Tweets data Analysis and Usage Proposal

Seminar on Social Media Content Analysis

Julià Camps JULIA.CAMPS@ESTUDIANTS.URV.CAT

Ana Cocho ANACOCHO@COAG.ES

Abstract

The aim of this report is to present a discussion on possible applications of knowledge extracted from tweets[1].

In order to provide general understanding on the problem of learning knowledge from social media content. This paper presents three different datasets of tweets, extracted from Twitter social network. Furthermore, discussions on the knowledge extraction process, for each of the datasets proposed, commenting on the real possible applications of this knowledge, are included.

The knowledge extraction process, in this paper, is presented as the composition of the following subprocesses: data acquisition, data analysis and interpretation on the analysis results.

Keywords: Twitter tweets, Knowledge extraction, Social media content, Dataset, Data acquisition and analysis.

1. Problem statement and goals

The evolution of the Web 1.0 towards the "Social Web"[2], together with its widely popularity, has led towards the current situation, where social media generate data massively on real time.

Let be remarked that, currently, in average, around six thousand tweets are generated per second, which corresponds to over 350000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year (see ratio).

Therefore, it can be assumed that when working with these social media, as data sources, for solving a certain problem, usually, it will be necessary to deal with *Big data*[3].

From analysing this data may provide valuable insights on many different domains (e.g. relevance of current events, detection of hot topics, public opinions on different topics, spatial movement of users and disease tracking). This information can be extracted from the

1. Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets".
2. The social web is a set of social relations that link people through the World Wide Web.
3. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.

content, plus meta-data[4], features analysis.

The application of Artificial Intelligence techniques, such as NLP[5], to this area has recently gained the attention of several research groups.

The main aim of this paper is to present the process of attacking the described problem and a brief discussion on possible interesting analysis procedures.

1.1. Content

In this paper we present three application cases, on the described problem 1, and a discussion on the analysis of each of them. In section 2 all related work found, previous analysis and literature, regarding the ideas conceived for implemented the required tweets datasets, is reviewed. Providing an overview on some of the existing most similar work reviewed to the proposed approaches in this project. Section 3 presents the new dataset presented in this paper. In section 4 the analysis procedures designed, for representing the knowledge willing to be extracted, are described. Section 5 illustrates the discussion on the outcomes from the full process of performing the proposed strategies on the designed datasets. In section 6 the testing dataset proposed, the designs and evaluations of the procedures, are discussed, presenting a comparative view with the planned settings with the implemented configurations. In section 7 the possible extensions of this project, together with the strengths and weaknesses analysis of the execution are commented. Finally, in section 8 some concluding remarks, regarding the project, are discussed.

1.2. The target scenario

Before going further let us recall on the target scenario properties.

Let us assume the following situation:

> "It is possible to retrieve tweets, by means of TwitterAPI tools, using filtering criteria such as: the language of the tweet, the senders of the tweet, a particular "string" present in the tweet, the domain of the tweet the location from which the tweet was sent or the time in which the tweet was sent."

The available raw criteria to be considered is the following:

Based on tweets content

- **Specific text content:** The most basic idea when filtering by content, is to search for content matching.

  Some categories of content that might be interesting for selecting tweets are:

  - Words matching: specific words appearing in the message content.

4. Additional information contained on a tweet rather than the message plain text content (e.g. time, sender ID and location).
5. Natural Language Processing.

  - Hashtags: the presence of "hashtags", which are detected by the presence of the special character: '#'.
  - URLs: determining the filter on the presence/non-presence of URLs included in the message content.
    Although, it could also be used for searching for specific URLs, in this case, we would consider it as *words matching* filtering.

- **Language:** Languages of the tweet, representing the membership as fuzzy information.

  Since we know that we can determine the language of a tweet, it is sensible to assume that the language of a tweet fraction may be also determined in order to detect presence of a certain language, instead of a predominance harsh classification.

  Although, this feature could be presented when necessary as part of the algorithm, we think that it is more interesting to delegate this task to the filtering tool. However, we are aware, that the current version of it, might not incorporate the mentioned feature, forcing it to be implemented as an intermediate preprocessing step.

- **Domain:** Tweets may belong to an specific domain (e.g. football, politics, religion, health or travelling).

  Usually, this classification may be performed, also, as part of the analysis process. However, it is assumed that we can filter tweets in a domain-wise notion.

Based on tweets meta-data

- **Sender:** User ID of the sender of a certain tweet.
- **Location:** Geographical coordinates (or spatial region) from where a certain tweet is sent.
- **Time:** Sending time of a certain tweet.

2. Related work

Due to the huge amount of state of the art related literature reviewed, it was decided to divide this section in an dataset wise way. Thus, having three subsections where on each, some related work regarding the concrete dataset is reviewed.

2.1. Regarding the first dataset presented 3.1

In this section some work related with the strategies used for designing the first dataset is listed.

- Mocanu u. a. (2013): Surveys worldwide linguistic indicators and trends through the analysis of a large-scale dataset of micro-blogging posts.

- Bergsma u. a. (2012): Explore the use of Twitter to obtain authentic user-generated text in low-resource languages such as Nepali, Urdu, and Ukrainian.

2.2. Regarding the second dataset presented 3.2

In this section some work related with the strategies used for designing the second dataset is listed.

- Gabrielli u. a. (2013): This paper proposes and experiments new techniques to detect urban mobility patterns and anomalies by analyzing trajectories mined from publicly available geo-positioned social media traces left by the citizens (namely Twitter).
- Jabreel u. a. (2016): The framework presented in this work constitutes the first semantic methodology for a large-scale automatic analysis of the communication of emotional values by destinations through social media.

2.3. Regarding the third dataset presented 6.1

In this section some work related with the strategies used for designing the third dataset is listed.

- Sriram u. a. (2010): Propose to use a small set of domain-specific features extracted from the author's profile and text. The proposed approach effectively classifies the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages.

3. Designing the datasets

In this section we discuss two of the three tweets datasets designed for this paper.

Since, the last dataset is discussed individually in section 6, due the fact that is the only one which has been implemented in order to support drive stronger conclusions from it, rather than hypothetical analysis results.

3.1. Real time language fluctuation dataset

The first dataset proposed is defined as a dataset containing real time truism information.

3.1.1. Real time language fluctuation dataset design

It is obtained by applying the following filters:

1. Territorial range → defines the origin territorial space, which tweets are considered as candidates for being kept in the dataset.

2. Set of languages → defines the set of interesting languages, and presence degree, that must appear on a certain tweet, in order for it to be kept.

3. Temporal window → defines the time thresholds of which tweets should be kept/removed from the dataset.

Notice that this dataset is an real time dataset, and is changing continuously over time.

3.1.2. Hypothetical implementation of real time language fluctuation dataset

Let us present a possible configuration of the just mentioned filters. Notice that, this example may be of interest in order to be able to comment the hypothetical results of the analysis phase, proposed for this dataset in section 4.

The *Tarragona* language fluctuation dataset

Suppose that we want to obtain the set of tweets regarding real time language influence in Tarragona[6] city.
A possible interesting configuration would be the following:

1. **Territorial range:** about a hundred of kilometres around Tarragona city.
   For knowing the possible influence on Tarragona city, it is unnecessary to have a much larger range. However, narrowing the range to much, would lead, in most cases, to precluding the opportunity of anticipate to relevant interesting changes on the scenario.

2. **Set of languages:** any predominant language on the tweets collected within the mentioned range, excluding Catalan, Spanish and English.
   Usually the most interesting knowledge to extract from data is the unknown information. Moreover, it is well known than *Tarraco*[7] is an international tourist attraction. Therefore Catalan, Spanish and English language information could be discarded, since it will be redundant with the already known information.

3. **Temporal window:** the window size should be relatively small, we would think on one or two days at most.
   Since the interesting information in this case is the real time information. However, due to the very specific data gathered, it may be easy to extend the window to one week in order to be able to perform more complex predictions on the individuals tracked.

3.2. Residents information dataset

The second dataset proposed is defined as a dataset on evolution of residents tweets.

6. Tarragona is the capital of the Tarragona province, located in Catalonia, Spain.
7. Tarraco is the ancient name of the current city of Tarragona. It was the oldest Roman settlement on the Iberian Peninsula and became capital of the Roman province of *Hispania Citerior*, and of *Hispania Tarraconensis* during the Roman Empire.

3.2.1. Residents information dataset design

It is obtained by applying the following filters:

1. Evaluation period → period of time willing to be evaluated.
2. Time discretization unit → fractions of time to be grouped an atomic unit within the period.
3. Resident area → geographical space to be evaluated.
4. Resident threshold → defines minimum number of times that a certain user must have tweeted from a certain location, in order to be considered as resident of it.

Notice that this dataset is a fixed dataset. The dataset is built for a certain time period, using past tweets.

3.2.2. Hypothetical implementation of residents information dataset

Let us present a possible configuration of the just mentioned filters. Notice that, this example may be of interest in order to be able to comment the hypothetical results of the analysis phase, proposed for this dataset in section 4.

The *Vilanova i la Geltrú* residents dataset

Suppose that we want to obtain the set of tweets regarding people related to *Vilanova i la Geltrú* city.
A possible interesting configuration would be the following:

1. **Evaluation period:** a reasonable period of time to be evaluated could be five years. In order to be able to have a more representative smoothed general evolution on the information generated by the residents on a certain area.
2. **Time discretization unit:** one month could be a good choice.
   In order to have a avoid having to much detailed case dependent information.
3. **Resident area:** for this example we propose the to evaluate the information of *Vilanova i la Geltrú* city residents.
   Notice that although the city currently has about sixty thousand inhabitants (see Idescat), the ones that will use Twitter regularly will be fewer. On the other hand, the definition stated for someone being considered resident, dose just imply having Twitter usage records accomplishing with some properties, rather than living on the area evaluated.
4. **Resident threshold:** we decided to set this threshold to a periodicity representation, from which a user being resident of a certain area in the current month, is defined as

$$((\text{number of tweets}) \in \text{AREA}) \geq \frac{4}{\text{month}} \quad \text{for at least 2 consecutive months} \qquad (1)$$

   . Which indicates that at least the user must have performed 4 tweets per month, during the previous 2 months, in order to be considered resident of the area, in the current one.

Finally, the retrieved dataset would be composed by the batch of tweets performed by users within the defined geographic are and evaluated.

4. Analysis on the datasets

This section presents some analysis strategies, for knowledge extraction procedure, and discusses them for each of the *hypothetical* datasets presented in sections 3.1 and 3.2.

In order to provide a better understanding on the analysis proposed. Let us assume performing the analysis on the example datasets defined at the final parts of sections 3.1.23.2.2.

4.1. Analysing dataset on language fluctuation 3.1

Recall: The first dataset presents the tweets flow (streaming), within a two days temporal window, containing foreign languages, different to English, sent from an area of 100 km radius surrounding Tarragona city.

The tools and techniques proposed for this dataset are next discussed.

1. Build a graph representation among tweets of different users which have tweeted from near locations in the same language.
2. From this graph infer possible groups (clustering) of "travelling-together" people.
3. Collect the tweets one week period of the deduced members.
4. Performing an complete NLP feature extraction decomposition, using ANNIE information extraction system. From this analysis it would be possible to extract useful information regarding the size of the tourist group that travelled together to Catalonia.
5. When the previous information being too uncertain, or for supporting the results obtained in the previous step. Estimate the group size using case base reasoning techniques.
6. Determine the most likely size of the group from the previous information.
7. If existing among the previous information, any mention regarding Tarragona visiting intentions, keep it for supporting future hypothesis.
8. Supposing the travelling together relation to be correct, it is possible to track the group from the individual members. Thus, detect movement indicators towards Tarragona, of any member of the group. Using the support information of the previous tweets analysis, when possible.
9. By means of probabilistic temporal models, infer the probability of the group visiting Tarragona. On the current and following day, thus defining a two a 48 hours prediction model representation.

10. Add the learnt knowledge to the overall predictions model. With the groups sizes information, the probabilities of travelling to different destinations.

11. The information could be represented as the following composition of charts:

    - 2D chart where the x-axis represent the size of the group while the y-axis represents the provability of coming to the city, containing the flags of the countries of the groups placed on the space representation.
    - Pie chart representing the presence of each language.
    - 3D blocs towers representation, where the color of each block determines the language, all blocs have the same size and each one represents one member, and the towers are displaced on the a circular map representation of the city, where the distances correspond to estimated time arrival, therefore, if the displacement is on the opposite direction, the tower will get off of the representation space.

    And having an additional feature that when you click any of the items all related items of the other plots are highlighted.

Notice that some of the steps of the just proposed procedure could be performed using several different combinations of tools, however, they should be able to deal with a large set of languages (e.g. WordNet, WordBreaker and WS4J would be excluded from the possible candidates, since this tools, currently, only accept English input words).

4.2. Analysing dataset on residents tweets 3.2

Recall: The third dataset presents the tweets records of users that are considered to be residents of Vilanova i la Geltrú city area, over the last five years, when being on the mentioned area.

The tools and techniques proposed for this dataset are next discussed.

1. At this point the dataset is just composed by the overall tweets created in the area of *Vilanova i la Geltrú* city.

2. Next the time discretization has to be applied, leaving the data represented as a composition of batches, where each batch contains the tweets generated during one month.

3. For selecting the residents, the threshold criteria must be applied to the dataset:

    (a) Track all the senders appearances.

    (b) Decided for each month if they should be considered as residents in the next batch

    (c) Store for each month, which are the residents of the area.

4. Build a model for each month, where each instance is represented by the full collection of tweets send by a single sender considered resident.

   From this specification of the model, we are applying an implicit normalization on the amount of tweets generated by user. Thus, all residents are equally representative.

5. For each of the instances specified, perform sentiment analysis detection, discerning just between positive/neutral/negative (e.g. quantified as $1/0/-1$ respectively) categories.

6. Averaging the values for computing the sentiment value of each instance.

7. Finally we would the average of the values calculated, in order to obtain an simple sentiment temporal description of the residents on an area, per month.

5. Discussion on the Generation and Analysis strategies proposed

This section reports the discussion on the outcomes of the hypothetical experiments proposed in this paper.

The following sections are divided into three main parts:

1. The motivations for developing the proposed dataset, are presented.
2. The results on the analysis of problems that could be faced during the overall procedure, are exposed.
3. Finally some further extensions on the referred dataset are presented.

5.1. Discussion on the strategies proposed for the dataset on language fluctuation

This section reports the discussion on the hypothetical outcomes encountered during the execution of the procedures proposed for the dataset presented in 3.1.2.

5.1.1. Motivations for implementing the dataset on language fluctuation

From applying the proposed techniques, we would be able to build a short term tourism fluctuation predictable model.

Therefore, being able to anticipate *unusual*(i.e. international, but non-English writers) tourists arrival to Tarragona city.

Although this information might be interesting for several possible applications. The main purpose that made us conceive this dataset, was to provide information openness to the local establishments around the city, in order to create competitivity on adaptation to

language fluctuation.

Offering to all business interested the possibility to adapt to potential customers changes, using real time predictions. Therefore, this system would balance the information distribution, while forcing the services to use the information for surviving.

This set of features, by forcing the services to improve their products, basing the offer on real time demand. Finally, could lead the city to become a reference on adaptability to tourists necessities. Therefore, perhaps, gaining popularity and more costumers.

5.1.2. Problems that could be faced when during the execution of the procedures proposed for the dataset on language fluctuation

In this section, some possible problems that are likely to be faced are reported when implementing this dataset.

- Having to few predicted groups to be significant and highly disperse when languages.
- Local people taking lectures on foreign languages and immigrant people, could lead to false positives.
- Predicting false positives could imply having taken unnecessary adaptation measures, and perhaps, even decreasing the quality of services, instead of increasing them.
- Big changes may be difficult for most of local business to adapt. Leading finally to only benefiting big companies, instead of building an equilibrated competition model.
- Lack of enough data, since this dataset would be ideally built under the assumption that, at least enough tourists, will communicate actively using Twitter platform, when willing to visit *Tarragona* city.
- One of the main concerns of this systems is the problem of error accumulation. Since, the process will use the previous self output, for computing the next step, one small error on the early stages, could lead to a complete wrong result.

  Thus, quality of data is a *must* requirement when performing this procedure.
- Problems when building the groups, since tourists tent to visit monuments, or other places where are many other tourists. It could be very difficult to build the groups correctly, since usually, the amount of nearby people tweeting the same language will be larger than the actual group.

  Notice that this problem may be addressed using the previous information analysis, in order to discover if the tweets senders actually live and travelled together.

5.1.3. Extensions for the dataset on language fluctuation

In this section, some possible extensions aimed to deal with the problems commented in the previous section, are presented.

- Regarding the false positives problem, using information of the users could provide enough insight on the past of these ones, in order to better ensure them being tourists.
- Concerning the groups problem, performing a more exhaustive tracking of the individuals activity (e.g. analysing their travelling, sleeping, eating and moving together, and their relations on Twitter social network).

Although there are more issues commented in the above section, the rest are not under the control of the method itself, are external problems.

5.2. Discussion on the strategies proposed for the dataset on residents tweets

This section reports the discussion on the hypothetical outcomes encountered during the execution of the procedures proposed for the dataset presented in 3.2.2.

5.2.1. Motivations for implementing the dataset on residents tweets

From applying the proposed techniques, we would be able to build a model representing the changes experienced by the people usually tweeting on a certain area. Describing this information for the last five years, in blocks of one month.

Therefore, being able to evaluate the impact of infrastructural changes performed in the area among residents. Such as if the mayor of the city projects performed in the area, had some significant impact on the residents emotional indicators.

Although, not being enough expressive for extracting strong conclusions, could serve as an very interesting indicator, moreover for providing support to other information sources.

In order to retrieve the information to the user, it would be interesting to accompany the graph of the sentiment evolution, with the temporal model representation, with a sliding bar for selecting the month evaluated. And plotting the information as the individual tweets sentiment value on the city map.

5.2.2. Problems that could be faced when during the execution of the procedures proposed for the dataset on residents tweets

In this section, some possible problems that are likely to be faced are reported when implementing this dataset.

- Having a non homogeneous distribution of the information on the area evaluated. Thus, having a non appropriate description of the information.

- This procedure is trying to assume influence relations between the tweets sentiment and the location where they where created. And although it should influence, there will be a lot of noise, of location agnostic data (e.g. free Wifi areas, areas where people are waiting for something such as train/bus stations).
- If there existed areas where there is not Internet connection, there would be non represented areas in the dataset. Which could even introduce more noise in the surrounding areas.

5.2.3. Extensions for the dataset on residents tweets

In this section, some possible extensions aimed to deal with the problems commented in the previous section, are presented.

- It is straight forward that the best approach for dealing the non homogeneous distribution could be downscaling the area filter to neighbourhoods in order obtain a more specific representation or changing the time discretization.

Whereas, regarding the other two issues commented in the above section, they are not under the control of the method itself, are external problems.

6. Implemented tweets dataset

In order to provide a complete understanding view on the procedures reviewed, we decided to implement this dataset.
However, since only this dataset was physically generated. It was decided to present the design and analysis discerning two different stages: first the *hypothetical* case (as has been done with the previous), then the real design and analysis faced during the implementation procedure, and finally a discussion on the comparison of both.

6.1. Correlated users information dataset design

The dataset proposed for implementing is defined as a dataset containing information on interests of correlated users.
It is obtained by applying the following filters:

1. Profile selection criteria → any set of characteristics that might define the target group of users (e.g. territorial information, domain of the messages or reference to a certain topic).
2. Temporal window → defines the period of time to be evaluated.
3. Interested threshold → defines minimum number of times that a certain user must have demonstrated interest in the topic, in order to be considered interested on it.

Notice that this dataset is a fixed dataset. The dataset is built for a certain temporal window, using past tweets.

6.2. Hypothetical implementation of correlated users tweets dataset

Let us present a possible configuration of the just mentioned filters. Notice that, this example is just an hypothetical case, moreover, the final implementation performed is documented in section 6.5.

The *Castells* users information

Suppose that we want to obtain the set of tweets regarding people interested in *Castells*[8]. An possible interesting configuration would be the following:

1. **Profile selection criteria:** words, hashtags and domain recognition within the *Castells* topic.
   In order to select the tweets sent referring the *Castells* topic.

2. **Temporal window:** one year could be enough significant, in fact up to two years could be interesting.
   Since it might be the case that people low their Twitter usage, but not their interest within the topic selected.
   Notice that the size of the window may vary depending on the selection criteria (e.g. when selecting users interested in football topic, since the interest on this topic is too extended, perhaps one month, as window size, would be a better choice, rather than one or two years).

3. **Interested threshold:** in this paper for this example we propose at least a minimum of 2 tweets per month.
   Here a trade of must be reached, since it is interesting to detect as much of *true positive* users as possible, while avoiding a high presence of *false positives*.
   Thus, for the proposed case, during two years, having two tweets per month frequency, would imply roughly 48 tweets within the *Castells* domain.

Finally, the retrieved dataset would be composed by the batch of tweets performed by these users within the defined period of time evaluated.

6.3. Analysing dataset of correlated users tweets 6.1

Recall: The second dataset presents the tweets of users that "demonstrated interest on Casltells" (i.e. having performed least three tweets related with the Castells topic during the evaluated year).

The tools and techniques proposed for this dataset are next discussed.

1. Group the set of tweets by sender IDs, in order to normalize the influence of each user recorded in the dataset.

8. Castells, also known as human towers, are one of the most impressing Catalonia's cultural displays and, also, one of Europe's most genuine and unique cultural displays.

2. For each user set of tweets perform an complete NLP feature extraction decomposition, using ANNIE information extraction system. From this analysis step we could learn the interests of the users represented.
3. Clean the current data gathered. Notice that in the current dataset we will have many redundant information. Then it would be interesting to ignore the information related with the *Castells* domain. Since the *Castells* domain was the selection criteria for searching the users. Thus, this information is redundant with the knowledge previous to the analysis.
4. Performing fuzzy definitions on the senders remaining interests information.
5. Extracting the predominant interests from the modelled data.

6.4. Discussion on the strategies proposed for the dataset of correlated users tweets

This section reports the discussion on the hypothetical outcomes encountered during the execution of the procedures proposed for the dataset presented in 6.2.

6.4.1. Motivations for implementing the dataset of correlated users tweets

From applying the proposed techniques, we would be able to build a model representing the main interests of users having some common characteristics.

Thus, being able to evaluate the popularity of different proposals of combined events, from learning which are the main interests of people belonging to the target group.
Some interesting applications could be the following:

- Evaluating if organising a competition of the activity of predominant interest among *Castells* interested people (e.g. if it was football, the Intercasteller competition already exists, and is very popular).

  Perhaps another possible competition could be focused on climbing or other *adventure sports*.

- Filtering the main interests of people belonging to an area as grouping criteria, could provide interesting feedback in order to decide which combination of events could suit better the inhabitants, during the city festivities. Let us suppose that for a small city, the outcome is:

  1. Rock music
  2. LaPatum
  3. *Castells*

  Perhaps they realise that *La Patum* date overlaps with their festivities dates, but since they do not dispose for the budget for copying it, they just decide to organise two buses for helping people to displace to it, whereas they organise the rock concert and the *Castells* event with the remaining budget.

There may exists endless possible applications of these type of datasets, since they are very generic, and have a very wide coverage.

In this paper, the proposed case, as stated in 6.2, is the *Castells* interested people case. Which was implemented in order to bring up conclusions on the real data gathered from Twitter.

The knowledge extracted from this application could be retrieved in several formats, but the one considered as the most adequate was: a ranking on the learnt alternatives, additionally, indicating the popularity of each of them. Since this representation is very simple and would provide a view on the all the knowledge extracted.

6.4.2. Problems that could be faced when during the execution of the procedures proposed for the dataset of correlated users tweets

In this section, some possible problems that are likely to be faced are reported when implementing this dataset.

- Having most users to be mono-topic interested users (e.g. people interested in *Castells*, do not care about anything else, and the same happens with people interested in football). Thus, the results are not representative of the group defined.

  Notice, that from the popularity information on the alternatives retrieved (proposed on the previous section), this phenomena could be easily detected.

- Problems shared with the previous datasets would be:

  - Having enough Twitter contribution.
  - Noise introduced on the data. Such as the medias (e.g. TV, radio, etc.) Twitter accounts.
  - False positives when selecting the users interested in *Castells*.

- Having information of people who, although being correctly detected to be interested in *Castells*, would not be interesting for the final decision (e.g. people interested in *Castells*, but living in China (see XiquetsdeHangzhou), would not be usually interesting when organising events in Catalonia).

- When organising combined events based only on geographical characteristics, the raw results retrieved could lead to inappropriate decisions.

  For explaining better this point, let us suppose the following case:

  On a certain town, 1/2 of people are interested in rock music, 1/3 of people is interested in running and 1/4 is interested in triathlon and 1/5 people is interested in running with bulls. And the final client could decide to organise a triathlon with bulls. Although, this activity may add more participation rather than the rock concert, it would be interesting to really understand if this subgroups are one part of the other, and which interests are actually compatible.

6.4.3. Extensions for the dataset of correlated users tweets

In this section, some possible extensions aimed to deal with the problems commented in the previous section, are presented.

- The media Twitter accounts issue could be detected by applying filters to the users considered for generating the dataset.
- In order to exclude most part of non interesting information, it could be a initial valid approach to use a geographical filtering.For the *Castells* case, in order to obtain information of tweets generated in Catalonia territory, and nearby areas.
- For addressing the last issue, concerning the possible wrong interpretation of the information, we propose performing *Association Rule Mining* (see Tan u. a., 2005) on the interests extracted, in order to learn hidden relations between interests of people selected.

  Actually, this strategy could deal with building a complete model of the relations existing between different interests of Twitter users, in an unsupervised way. Thus, without being necessary to define the target group. This could be interesting in order to extract general knowledge from tweets. However, notice that further discussing on the results of such a general analysis is non sensible without implementing it, moreover, is beyond the scope of this paper.

Whereas, regarding the other issues commented in the above section, they are not under the control of the method itself, moreover are external problems.

6.5. Discussion on the implementation procedure for the dataset of correlated users tweets

In this section the process of implementing the dataset described in 6.2 and the analysis proposed 4 is discussed.

6.5.1. Design configurations for the dataset of correlated users tweets

In this section the parameters chosen for the generation of the dataset finally used are exposed.

CONFIGURATIONS	PARAMETERS				
	SELECTION CRITERIA	TEMPORAL WINDOW	INTERESTED THRESHOLD	SOCIAL MEDIA TAGS	AREA
PROPOSED	*Castells* domain	2 years	48	none	Catalonia
IMPLEMENTED	*Castells* domain	about 2 weeks	2	"tv" "radio" "info" "diari"	300km surrounding Barcelona

Table 1: Configurations the dataset on correlated users tweets.

Table 1 shows the comparison among the proposed configuration and the implemented one.

As can be observed in the table presented above, the final configuration used for the implementation phase of the project, differed with the initial proposal. This changes are discussed in the remaining part of this section.

- Selection criteria: it is the same than the proposed one.
- Temporal window: the tools used for implementing the dataset (TwitterAPI), had limitation regarding the period from which the data could be collected. The limitation was reviewed to be about two weeks.
- Interested threshold: although the proportional alternative for this parameters, would be to chose 1, we decided to maintain it to 2.

  Although that, due to this choice we might have discarded some users which would actually have been accepted using the proposed configuration, it was considered that using the threshold to 1, would introduce all false positives present in the tweets gathered. Thus, we decided to loose some data, as a trade of for improving quality.
- Social media tags: this parameters is introduces as a possible extension, and since it was noticed to be a significant part of the noise appearing in among the users retrieved (about 20% of users retrieved). Thus, it was considered to finally introduce some filtering tags, in order to discard as many non personal Twitter accounts as possible. The tags chosen for this experiment were: "tv", "radio", "info" and "diari". Since they showed allowing to remove most of this undesired senders.

- Area: although this parameter was also presented as an extension, for it there was a setting proposed, which was taking into account the area belonging to Catalonia. However during the while implementing, the way of defining the area was reviewed to be by providing a point coordinates and the radius distance, in order to having a circumference being drawn.

  Due to this format it was decided to amplify the area considered, in order to take into account possible tweets performed by temporal displacement to nearby areas of users, together with the tweets generated in the Balearic Islands.

  The distance selection was calculated by drawing a circumference on the world map, it was reviewed that the distance between Barcelona and Saragossa was enough for our purpose (i.e. $255.35km$ in straight line (distanceCalculator)), however, it was decided to round up the radius to $300km$, in order to be more flexible when deciding if an area is interesting for the final purposes of the dataset generated.

  Notice, that all information is kept, thus, additional filters may be done as part of the analysis procedures.

6.5.2. Dataset generated characteristics

The dataset of the tweets gathered from users interested in the *Castells* topic had the following characteristics:

- Size: 839 MB
- Number of tweets: 32822
- Users represented: 358 (out from the 652 retrieved in the previous stage, due to limitations on the Twitter API usage (see APIlimitations)).

6.5.3. Analysis results performed on the generated dataset

In this section the analysis procedure is exposed, together with some discussion on the changes compared to the original planning.

From the generated dataset (in *json* format), we created a simplified version containing just the sender names and the tweet message contents, in order to use it for the analysis proposed.

Notice that finally it was impossible to implement the full analysis procedure due to its complexity and time constraints. Thus, it was decided to propose an alternative simplified analysis. Which is presented in the following part of this section.

1. Extracting the message content, and relating them with the senders.
2. Instead of using ANNIE software, it was decided to use python well known NLP libraries, in order to perform the information extraction.

3. Search for relevance of other domains in the dataset, in order to infer the interest correlation among *Castells* and them.
4. The domains chosen for this experiment were: football, climbing and volleyball.
5. Extracting conclusions on the relevance of the presence of each of the proposed topics.

6.5.4. Discussion on the results obtained from the analysis on the generated dataset

In this section the results obtained from the analysis phase are discussed.

Although finally, the full analysis procedure was simplified. It was possible to extract some conclusions from it, regarding the interest of people who are interested in *Castells* to the following topics: football, climbing and volleyball.

RESULTS

FOOTBALL	CLIMBING	VOLLEYBALL
0.042%	0.018%	0.029%

Table 2: Results from the analysis.

Table 2 shows the results retrieved from the analysis phase of the project. The vocabulary used for implementing this filters, was obtained from football, climbing and volleyball.

From the presented results the following weak[9] deductions may be performed:

1. "Football" appears to be the most interesting topic of the three evaluated, for people who are interested in *Castells*. This fact, may serve for explaining the huge popularity of the Intercasteller.
2. Initially we supposed that "climbing" would be more popular than "volleyball". However the results obtained, show that both options are equally popular.

9. Since it would not be appropriate to drive strong conclusions on such a simple experiment, with the finally used fraction of the dataset proposed.

7. Extensions, strengths and weaknesses

This section presents the evaluation conclusions on the performance analysis, of this project.

7.1. Extensions

Due the time limitation resources it was not possible to implement all the presented datasets, neither the algorithms selected interesting candidates for the analysis (e.g. Association Rule Mining), or to explore further interesting methods, such as probabilistic models (see Çelikyilmaz u. a., 2010).

Some other possible extensions for this project would be:

- Finishing the experiments discussed on the dataset proposed.
- Implementing the other proposed datasets and perform a comparison on the expectations and the results retrieved.
- Exploring more complex approaches to the target problems, such as RNN[10] (see Dong u. a., 2014).

7.2. Strengths

Big data, generated by social media, data mining are very trendy concepts, especially, when concerning Twitter and real time knowledge extraction. The paper gives a clear and brief insight about the most important elements in the NLP knowledge extraction process, focusing on the specific scenario of extracting information from tweets datasets. While giving a general global view on the state of the art of data mining strategies and NLP analysis methodologies.

Other strengths of this project would be:

- The information provided during the AIS[11] conferences, and the additional material facilitated, was enough for performing this project (i.e. without researching much more in the literature).
- There exist endless sources of literature regarding Twitter data mining.
- All the topics discussed within this paper are state of the art, in the field of AI[12], research topics.
- The Twitter API documentation is very complete and there are many coding tutorials for creating datasets of tweets (see APIdocumentation).

10. Recursive Neural Networks
11. Artificial Intelligence Seminar
12. Artificial Intelligence

- All the information used and presented in this paper is real information, hypothetically designed from the existing tools and methodologies reviewed, extracted from analysing the crafted dataset or reviewed in other experiments presented in the literature.

7.3. Weaknesses

During the development of this project some problems appeared, forcing changing the initial planing, to the finally presented in this paper. The two most relevant issues during this work were:

- The impressive amount of definitions required for giving to the project the robustness and support necessary for presenting it as a serious report, on a real study case using real data.
- The lack of previous experience regarding the modelling of the social media data, at any level.
- The huge amount of previous work reviewed in the literature related with the initial designs of the three datasets, in fact, several times, exactly the same concept was already proposed in other papers.

  This made us restart the project several times, moreover, complicating the datasets in order to reduce the probability of matching any previous idea published by other research groups.
- Several problems where found while implementing the proposed datasets using Twitters free developers API (TwitterAPI).

  Next we list some of them:

  - Limitation of time availability of information (see forum).
  - Limitation of requests performed by application (see APIlimitations) (see supportPage) (see APIerrCodes).
- The previous two points, leaded to an unavoidable overload of work, which precluded the possibility of programing the knowledge extraction from the dataset full process presented.

  As a remarkable fact, notice that just building the raw dataset of tweets, it toke about three days due to the requests limitations. While without this limitation, would have been achieved within one hour.

8. Conclusions and Future work

After this project we built a strong understanding on the techniques and strategies discussed during the seminar.

This paper also presents a practical application case of the techniques discussed during the seminar.

We present a dataset for extracting information of people who are interested in the *Castells* topic. By filtering the users who tweeted within the past two weeks using words referring to the *Castells* domain, at least two times.

The selected methods for performing the analysis were finally not implemented due to issues faced when working with TwitterAPI, however, an alternative simplified analysis was performed, retrieving, overall, interesting results.

8.1. Future work

As has been commented in section 7.1, there exist many improvements which could be performed in order to continue this project. However, it would be perhaps more reasonable to test the current design before going further, in order to detect functional weaknesses on the model presented.

It has already been noticed that the *geographical* regulation, is actually perhaps to strict, although it has a sensible measurement strategy. A better approach to this problem, would be finding an alternative evaluation on the user profiles, in order to decide if they are interesting candidates for building the dataset.

Performing further exploration on the possible additional interesting information, that could be used for enriching the knowledge extracted, using different analysis goals (e.g. extracting *friendship* relations among users, detecting ideologies of users, participation in activities, rather than just showing interest on them, etc.).

8.2. Conclusions

Rather than just the implemented case, this paper is focused in providing a global overview on applications of the state of the art techniques used for knowledge extraction from social media sources.

Thus, it provides full description on several application cases, supported always by appropriate discussions, on the decisions taken for designing the required procedures on any situation.

Related literature has been reviewed showing that a lot of work is being done with this methods, but still there is a lot to explore.

The research performed regarding the knowledge extraction from tweets datasets leaded too some main conclusions:

1. For solving properly a real problems, a strong description and previous knowledge on the target domain is required, in order to avoid having useless information, or even performing wrong assumptions on the final results.

2. Although the data exists, not much work has been done within the local topics at low scale. In fact, the datasets do not exist by them selves, difficulting the access and usage to data. Which could held to mutual benefits, by creating applications and promoting the local culture, at the time that benefiting local users.

3. Big-data is leading to changes in the modelling paradigm, since there is too much data to be able to model it traditionally. Which, in any case, would lead to an overfitted model.

4. Several research is being done on the sentiment analysis area. Which seams that will lead to a dramatic change on the way that RS[13] will work. Since when able to find underlying relations between different topics, they will be able to perform very accurate predictions on the users preferences, just using partial information of their profiles.

References

[ANNIE] : *ANNIE Example information extraction system included in Gate.* http://services.gate.ac.uk/annie/ 7, 14, 18

[distanceCalculator] *Distance calculator.* http://es.distance.to/Barcelona/Zaragoza. – Accessed: 12-06-2016 18

[Idescat] *Idescat.* http://www.idescat.cat/emex/?id=083073&lang=es. – Accessed: 12-06-2016 6

[LaPatum] *La Patum.* http://www.lapatum.cat/. – Accessed: 12-06-2016 14

[TwitterAPI] *Twitter API.* https://dev.twitter.com/rest/public. – Accessed: 06-06-2016 2, 17, 21, 22

[APIdocumentation] *Twitter API Documentation.* http://docs.tweepy.org/en/v3.5.0/api.html. – Accessed: 12-06-2016 20

[APIerrCodes] *Twitter API Error codes.* https://dev.twitter.com/overview/api/response-codes. – Accessed: 12-06-2016 21

[APIlimitations] *Twitter API Limitations.* https://dev.twitter.com/rest/public/rate-limiting. – Accessed: 12-06-2016 18, 21

13. Recommender Systems

[forum] *Twitter developers forum*. https://twittercommunity.com/. – Accessed: 12-06-2016 21

[Twitter] *Twitter homepage*. https://twitter.com/. – Accessed: 06-06-2016 1

[supportPage] *Twitter support page*. https://support.twitter.com/articles/344781#. – Accessed: 12-06-2016 21

[ratio] *Twitter Usage Statistics*. http://www.internetlivestats.com/twitter-statistics/#trend. – Accessed: 07-06-2016 1

[climbing] *Vocabulary climbing*. https://latevaruta.wordpress.com/vocabulari-a-la-muntanya/. – Accessed: 13-06-2016 19

[football] *Vocabulary football*. http://www.termcat.cat/ca/Diccionaris_En_Linia/3/Fitxes/. – Accessed: 13-06-2016 19

[volleyball] *Vocabulary volleyball*. http://mundovoley.galeon.com/aficiones1520828.html. – Accessed: 13-06-2016 19

[WordBreaker] : *Word Breaker Demo Microsoft Research - Word Breaker Demo*. http://weblm.research.microsoft.com/WordBreakerDemo.aspx 8

[WordNet] : *WordNet WordNet is a large lexical database of English*. http://wordnetweb.princeton.edu/perl/webwn 8

[WS4J] : *WS4J WordNet Similarity for Java) measures semantic similarity/relatedness between words*. http://ws4jdemo.appspot.com/ 8

[XiquetsdeHangzhou] *Xiquets de Hangzhou*. https://ca.wikipedia.org/wiki/Xiquets_de_Hangzhou. – Accessed: 12-06-2016 15

[Intercasteller] *XV TCF CJXT*. http://collajovetarragona.cat/activitats/torneig-casteller-futbol/. – Accessed: 12-06-2016 14, 19

[Bergsma u. a. 2012] Bergsma, Shane ; McNamee, Paul ; Bagdouri, Mossaab ; Fink, Clayton ; Wilson, Theresa: Language Identification for Creating Language-specific Twitter Collections. In: *Proceedings of the Second Workshop on Language in Social Media*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2012 (LSM '12), S. 65–74. – URL http://dl.acm.org/citation.cfm?id=2390374.2390382 4

[Dong u. a. 2014] Dong, Li ; Wei, Furu ; Tan, Chuanqi ; Tang, Duyu ; Zhou, Ming ; Xu, Ke: Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland : Association for Computational Linguistics, June 2014, S. 49–54. – URL http://www.aclweb.org/anthology/P14-2009 20

[Gabrielli u. a. 2013] Gabrielli, Lorenzo ; Rinzivillo, Salvatore ; Ronzano, Francesco ; Villatoro, Daniel: From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. In: Nin, Jordi (Hrsg.) ; Villatoro, Daniel (Hrsg.): *CitiSens* Bd. 8313, Springer, 2013, S. 26–35. – URL http://dblp.uni-trier.de/db/conf/citisens/citisens2013.html#GabrielliRRV13. – ISBN 978-3-319-04177-3 4

[Jabreel u. a. 2016] Jabreel, Mohammed ; Moreno, Antonio ; Huertas, Assumpció: Semantic comparison of the emotional values communicated by destinations and tourists on social media. In: *Journal of Destination Marketing Management* (2016), S. –. – URL http://www.sciencedirect.com/science/article/pii/S2212571X16300117. – ISSN 2212-571X 4

[Mocanu u. a. 2013] Mocanu, Delia ; Baronchelli, Andrea ; Perra, Nicola ; Gonçalves, Bruno ; Zhang, Qian ; Vespignani, Alessandro: The Twitter of Babel: Mapping World Languages through Microblogging Platforms. In: *PLoS ONE* 8 (2013), 04, Nr. 4, S. 1–9. – URL http://dx.doi.org/10.1371%2Fjournal.pone.0061981 3

[Sriram u. a. 2010] Sriram, Bharath ; Fuhry, Dave ; Demir, Engin ; Ferhatosmanoglu, Hakan ; Demirbas, Murat: Short Text Classification in Twitter to Improve Information Filtering. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA : ACM, 2010 (SIGIR '10), S. 841–842. – URL http://doi.acm.org/10.1145/1835449.1835643. – ISBN 978-1-4503-0153-4 4

[Tan u. a. 2005] Tan, Pang-Ning ; Michael, Steinbach ; Kumar, Vipin: *Association Analysis: Basic Concepts and Algorithms*. Kap. 6, S. 327–404. In: *Introduction to Data Mining*, Addison-Wesley, 2005. – URL http://www-users.cs.umn.edu/~{}kumar/dmbook/ch6.pdf. – ISBN 0321321367 16

[Çelikyilmaz u. a. 2010] Çelikyilmaz, Asli ; Hakkani-Tür, Dilek ; Feng, Junlan: Probabilistic model-based sentiment analysis of twitter messages. In: Hakkani-Tür, Dilek (Hrsg.) ; Ostendorf, Mari (Hrsg.): *SLT*, IEEE, 2010, S. 79–84. – URL http://dblp.uni-trier.de/db/conf/slt/slt2010.html#CelikyilmazHF10. – ISBN 978-1-4244-7903-0 20