ISSN 1870-4069

Temporal Language Analysis in News Media and Social Networks

Fernando S. Peregrino, David Tomás, Fernando Llopis

University of Alicante, Department of Software and Computing Systems, Spain

{fsperegrino,dtomas,llopis}@dlsi.ua.es

Abstract. The amount of text we can find on the Internet is constantly growing, what makes not feasible to manually analyse such as quantity of information. The Natural Language Processing (NLP) research field has provided a set of tools and techniques that allow human beings to extract relevant data from unstructured pieces of text that come from electronic sources such as digital newspapers or online social networks. The aim of this article is to make a temporal analysis of both formal (newspaper articles) and informal (Twitter messages) texts sources. In this article, we will analyse how some terms evolve in time and the correlation between the formal and informal corpora.

Keywords. Geographical focus detection, geographical information retrieval, natural language processing.

1 Introduction

Nowadays information society is characterised by having a huge amount of information and for the need to allow both a rapid access and disclosure of this information. To this end, storing such information in a digital format has become crucial. Access to this digital information has to be as fast and accurate as possible. With this motivation arise Information Retrieval (IR) systems, also known as search engines.

IR is the science of searching for information in digital documents, producing as a result a subset of the initial documents sorted according to the relevance of a given query. IR deals with the representation, storage, organization of and access to information items such as documents and web pages. The representation and organization of the information items should provide the user with easy access to the information she is interested in [1].

The core tasks of these systems is generally split in two stages: indexing and searching. In the former, all the terms in the documents are indexed. To this purpose, the system records in which documents and how often appears each of these terms, as well as different statistics related to the number terms in every single documents. In the latter, a query is sent to the system to retrieve a set of document sorted by the relevance to this query, trying to match the terms which are in the query with the ones that have been indexed in the previous phase.

Among all the data that IR systems have to deal with, geographical information is a specially relevant type. In accordance with the study conducted by [6], 12.7% over four

million sample queries have a toponym[1] at least. This has been corroborated by the work carried out in [4], where 36 million queries submitted to an IR system were analysed. It was found that between 18% and 22% of these queries where geographically bounded.

Geographical Information Retrieval (GIR) is a specialisation of IR systems, where documents have associated geographical metadata. GIR systems require semantic information, i.e. geographical features associated with the documents. Because of this, in GIR systems document processing and indexing is usually separated from geographical indexing. In other words, each document is indexed according to the place or places in which the document is focused on.

In order to discover weaknesses and opportunities when this geographical index is made, this work conducts a study on different text sources, comparing them to analyse aspects such as the evolution of the use of terms over time and the correlation between formal and informal sources taking into account its geographical scope.

The structure of this article is as follows: Section 2 describes the corpora used in this work; Section 3, provides a description of the analysis and results obtained; finally, Section 4 presents the main conclusions that can be drawn from the work carried out.

2 Corpora

This section describes the corpora analysed in this study. One of the main problems that GIR systems have to deal with is the different nature of texts. If we classify the texts according to their level of formality, they could be divided in two groups:

1. **Formal**. A great amount of formal texts can be found in digital format: newspaper articles, encyclopedia articles, reports, reviews, etc. In this paper we will focus on newspaper articles. More precisely, in the articles belonging to the *20Minutos*[2] newspaper.
2. **Informal**. Online Social Networks have facilitated the existence of huge amounts of texts written in informal language. In this paper we will worked on *Twitter*[3] as a source of this type of texts.

In the two following subsections, we will describe in depth the aforementioned corpora.

2.1 20Minutos

20Minutos is a free Spanish newspaper, with local editions in several Spanish cities. News articles are geographically classified according to the region that they belong to. In this work, the locations that we have worked with have been the 50 Spanish province capitals plus the two Spanish autonomous cities of Ceuta and Melilla. We chose the city capitals since they usually are the most populated cities en their province as well as the administrative location, which makes them a source of a large number of articles

[1] Toponymy is the study of place names (toponyms), their origins, meanings, use, and typology.
[2] `http://www.20minutos.es/`
[3] `https://twitter.com/`

that represent not only the given city but its province. The period of time comprised was from January 1st 2008 to December 31st 2011, inclusive. For all these four years and geographical regions, we obtained a total number of 519,563 geographically tagged newspaper articles.

In order to obtain the articles, we built a crawler which iterated over the section where the local news where published. Each piece of news found from any of aforementioned cities and time period were stored.

A pre-process was carried out on the corpus, removing all the *URLs*, punctuation symbols and special characters such as underscores, slashes, etc. In addition, as part of this pre-process, all the terms were lower-cased, except for the first character of each term, which was kept with the purpose of identifying proper nouns such as locations, people or organisations.

Because of the comparison between *20Minutos* and Twitter corpus in the experiment conducted in Section 3.1, a supplementary set of newspaper articles belonging to the year 2013 were crawled. The dates encompassed in this set was from March 1st 2013 to July 31st 2013. This set of newspaper articles was pre-processed as described in the previous paragraph.

2.2 Twitter

Twitter is an online social networking service that enables users to send and read short 140-character messages called *tweets*. Twitter was created in March 2006 and launched in July 2006. The service rapidly gained worldwide popularity. As of August 2017, Twitter has more than 325 million monthly active users[4].

To obtain our corpus of tweets, we used the Twitter *SEARCH API*, currently included with some restrictions in the Twitter *REST API v1.1*[5]. Through this API, a set of geo-referenced tweets was acquired from the 50 Spanish capital cities plus its two autonomous ones, as in the *20Minutos* corpus (see Section 2.1). The dates of these tweets span from April 20th 2013 to June 10th 2013, 52 days.

Given that Twitter users can send tweets from more than one location in the collected corpus, tweets were grouped by user and location. That is, were we to have a set of tweets from a user which has tweeted from both the location *A* and *B*, those set of tweets are treated as two different sets instead one. The reason for doing this is that when users tweet from a location, in many cases they refer to a place or entity in the place where they are writing, what means it could be useful when it comes to extract features from the given place.

The most relevant data extracted from the obtained tweets is:

- The text of the tweet. Up to 140 characters.
- The tweet location. It is one of the 52 location aforementioned.
- The user. The user who has sent the given tweet.

Notice that if a user has tweeted from n different locations, it will count as n users. What underlies this is that we are trying to figured out where a user is analysing a set

[4] `https://about.twitter.com/company`
[5] `https://dev.twitter.com/docs/using-search.`

of tweets that he or she has sent, and since these set of tweets could differ considerably depending on the location where they were sent, we are splitting these tweet according to the origin, in order to avoid extracting noisy features from the different places.

As in the *20Minutos* corpus, a pre-process was carried out in the corpus of tweets. All the punctuation signs, special characters (underscores, slashes, etc.) and *URLs* were removed. The terms which started with # or @ were kept, since these characters have a special meaning in Twitter, representing *hashtags* and user names respectively, what could be useful to determine the users location.

All the repeated tweets as well as the re-tweeted ones were also removed, as they do not give any additional information. No additional information (e.g. *followers* or location field in the profile) was taken into account.

3 Analysis

This section describes an analysis over the two corpora mentioned in the previous section. This analysis exposes the temporal texts terms evolution according to their geographical and temporal arrangement.

The aim of this analysis is to give a deeper insight into the existing relationship among the used terminology in order to improve the geographical focus detection in GIR systems.

To this end, the following aspect will be analysed in both corpora:

1. Correlation between terms from both corpora.
2. Terms evolution over time.

3.1 Corpora Correlation

In this section, we will show the correlation between the terms of the *20Minutos* and Twitter corpus. To accomplish this, we will measure how similar the texts from both corpora are. This similarity will be carried out comparing the Twitter corpus with a corpus of newspaper articles that encompass three different periods in the year 2013, as described in section 2.1:

1. **Before**. The 52 previous days to the Twitter corpus.
2. **During**. The 52 days which coincide with the Twitter corpus.
3. **After**. The 52 days after the Twitter corpus.

The reason under this corpus split is to corroborate the following hypothesis: the messages expressed in Twitter usually reflect what is happening in day-to-day, what means that a strongest correlation with contemporary newspaper articles must exist. Thus, the terms from each of these three newspaper corpora were compared with the Twitter corpus. This comparison was carried out at city level, i.e., for a given location and period of time (*before*, *during* or *after*) its terms were compared with the terms of the same city in the Twitter corpus.

In order to check the correlation between each of these texts Kullback-Leibler (*KL*) divergence was used [5]. In probability theory and information theory, the KL

divergence is a measure of the difference between two probability distributions *P* and *Q*. In practice, *P* represents the real distribution of data, observations, or a precisely calculated theoretical distribution, while *Q* represents a theory, model, description, or approximation of *P*. In other words, it is the amount of information lost when *Q* is used to approximate P.

For discrete probability distributions *P* and *Q*, whose vocabulary is in the set of finite terms χ, the KL divergence from *Q* to *P* is defined by equation 1:

$$D(P||Q) = \sum_{x \epsilon \chi} P(x) log \frac{P(x)}{Q(x)}. \tag{1}$$

Given that the KL divergence is not symmetric, we have calculate the KL symmetric divergence, i.e. the KL distance, based on the approach exposed in [3], as the equation 2 shows:

$$D(P||Q) = \sum_{x \epsilon \chi} (P(x) - Q(x)) log \frac{P(x)}{Q(x)}. \tag{2}$$

In this way, all the terms in the corpus of tweets for each location is obtained along with their frequency. On the other hand, the same operation is carried out with the set of newspaper articles which are included in the period that we want to measure the KL distance with.

The union of all the term from both corpora, that is the vocabulary from both corpora, is expressed by the symbol χ in the equation 2. *P* and *Q* represent the Twitter and *20Minutos* terms frequency normalised vectors respectively.

So as to accomplish this normalisation, the total number of terms and their frequency that appear in each of the given locations is considered. Equation 5 shows the smoothed applied to the terms that are not in the Twitter corpus but exist in χ.

Once all the terms from the Twitter corpus of a given location are normalised, a vector with the weight of all the terms that are in χ is created. For all those terms which are in the vocabulary (χ) but not in the corpus of tweet of the analysed location, a residual value is assigned. This value is calculate dividing the smoothing value ϵ, obtained in the equation 5, by the total number of terms that are in χ but not in the vocabulary of the location under analysis. The frequency of these terms is also considered. Consequently, this normalisation is carry out as is shown in equation 3:

$$P(t_i, c_j) = \begin{cases} \frac{freq(t_i)}{d}, t_i \; is \; in \; the \; corpus \\ \frac{\epsilon}{d}, t_i \; is \; not \; in \; the \; corpus. \end{cases} \tag{3}$$

In this equation, $P(t_i, c_j)$ represents the term *i* probability from the location *j*. $freq(t_i)$ is the stands for the term *i* frequency in the location *j*. d is the divisor that distributes the ϵ weight over the terms that are not in the corpus. Finally, ϵ is the smoothing factor applied to the terms which are not in the corpus.

The resultant vector is denoted in the equation 2 by the symbol *P*, whilst *Q* represents the vectors of terms gathered from the location analysed in the *20Minutos* corpus

in the given time span. This vector has been built following the procedure employed for vector *P*.

The calculation of the divisor *d* shown in the definition 3 is achieved in accordance with what is exposed in equation 4:

$$d = np * \epsilon + fp, \tag{4}$$

where *np* represents the number of terms in χ that are not in the analysed corpus, ϵ is the smoothing value, and *fp* is the total number of terms and their frequency which appear in the analysed corpus.

The calculation of the smoothing value ϵ is effectuated as shown in equation 5:

$$\epsilon = \frac{\frac{1}{1+(|vp-vq|)}}{1+p+q}. \tag{5}$$

This equation is based on the work by [3], where $|vp - vq|$ represents the total number of terms which are not shared by the two analysed corpora. *p* is the total number of terms for the Twitter analysed location, where each term is multiply by its frequency in the given corpus. *q* stands for the *20Minutos* total number of terms from the analysed location and period multiply by their frequency.

In such manner, KL distance has been calculated at city-level between the Twitter corpus of these locations and these same locations from the *20Minutos* corpus for each of its three periods.

Notice that the smaller the KL distance, the bigger the correlation between corpora. If we look at the 10 largest Spanish locations, in 9 out of 10 (90% precision) the distance is lower in the *During* period, showing a largest correlation between the terms in both corpora when day occur in the same time span.

It is noteworthy the different nature of these text sources, where the *20Minutos* texts are related to current issues in a concrete geographical area in a specific date. On the other hand, in Twitter, we can find a huge range of topics. Frequently, this topics are not related to the location where they were sent and, therefore are dealing with a completely different subject. Actually, in many occasions, these tweets talk about subjects specific from other locations.

We should also keep in mind that there is little geographical information in the vast majority of tweets, which makes more difficult to match these messages with the given local news.

Another aspect to consider is that the three (*before*, *during* and *after*) *20Minutos* corpora used in this analysis where published in consecutive dates, that is these newspaper articles where very close in time, so that the treated issues in these periods overlap each other, specially when it comes to the period in the middle, the one which is supposed to get the minimum KL distance with respect the Twitter corpus.

All in all, having into consideration all the difficulties previously described, we can conclude that our results get a significant accuracy, mainly when it comes to the largest cities.

3.2 Temporal Language Evolution

In this section, we will show the temporal evolution of a set of terms included in both corpora at location granularity. In this way, we will be able to see how the relevance of the selected terms is changing in each of the presented locations, allowing to detect or predict the main citizens' concerns for determined periods of time.

So as to analyse the evolution of this set of terms, corpora were split in periods, as it will be explained in sections 3.2 and 3.2. The terms and their frequency were obtained for each of this periods in each of the corpus locations. With these terms and their frequency, the standard score (*z-score*) [2] of each term was calculate. To calculate this measure, each term in a period was compared with the rest of periods for each given location.

The standard score is the signed number of standard deviations an observation is above the mean. A positive standard score indicates an observation above the mean. The calculation of the standard score is shown in equation 6:

$$z = \frac{x - \bar{a}}{\sigma}, \tag{6}$$

where x is the analysed term normalised frequency for a given period of time. $\bar{a}$ is the mean of population, i.e. this same term normalised mean for a given location in all the corpus periods of time. Finally, σ represents the standard deviation of those values.

The set of terms to be analysed was chosen from the list of the main concerns for the Spanish citizens. This list was acquired from the Spanish *Centre for Sociological Research* (CIS: Centro de Investigaciones Sociológicas).[6]

The *CIS* surveys the Spaniards citizens monthly so as to detect their main concerns. For this survey, the *CIS* asks to the survey respondents for choosing their three main concerns from a set of options. The options list can be seen in the *CIS* web page[7]. The *CIS* will finally show the percentage of people that have selected each of these options.

For this analysis, the three chosen subjects are: *paro* (unemployment), *corrupción* (corruption) and *educación* (education). These subjects were among the most important Spaniards concerns according to *CIS* surveys.

In order to capture the three aforementioned subjects in *20Minutos* and Twitter texts, the presence of these terms without taken into account capitalisation or accents was observed. The temporal evolution of these concerns for a set of sample locations (Alicante, Madrid and Sevilla) and the country average (España - Spain) is shown in the sections 3.2 and 3.2. The results of the *CIS* survey is also displayed.

So as to check whether a concern gain or lose importance in each of the selected periods of time for a given location, all the terms that were in the texts of each of the corpora locations in each period were sorted according to their standard score in descending order, i.e., the greater their standard score, the higher in the returned list. Once obtained this sorted list of terms, the position in which the terms appears in the

[6] The *CIS* is a Spanish public research institute for sociological issues `http://www.cis.es/cis/opencms/EN/index.html`.

[7] `http://www.cis.es/opencms/-Archivos/Indicadores/documentos_html/TresProblemas.html`

list and the total number of terms that was in the period of time and location analysed were considered in order to obtain the final score as shown in equation 7:

$$ConcernIndex = 100 - 100 \times \frac{TermPosition}{TotalNumberofTerms}. \quad (7)$$

This equation provides values between 0 and 100, allowing to directly compare them with the values obtained in the *CIS* surveys. Notice that the closer these values are to 100, the more significant concern was shown in that period of time. Once the concern index for a given term in a period of time and location was calculated, it was compared with the rest of periods of time for this same term and location.

20Minutos The *20Minutos* corpus was divided by years, given as a result four different corpus for each location: 2008, 2009, 2010 and 2011.

The obtained results for the main citizens concerns in each of the aforementioned locations are displayed together with the country average and what the *CIS* survey obtained.

***Corruption* temporal evolution**

The evolution of the term *corrupción* in both the *20Minutos* articles and the *CIS* surveys is shown in the figure 1.

In the figure 1, it can be observed how the results between cities and *CIS* clearly differ.

Cities such as Alicante or Seville obtain a similar results than the country average, whilst Madrid has an opposed trend. At a city granularity, the citizens corruption perception can suffer great variations depending on the cases in the given places and when these have happened.

Thus, if a city such as Alicante is analysed, a case like Gürtel[8] that erupted in the Valencia Community at the end of 2007 causes that in the following years the corruption perception was rising, reaching its peak when the former president of this community, Francisco Camps, resigned in the mid of 2011.

If we reduce the geographical focus to Alicante, cases such as Brugal[9] or Caja de Ahorros del Mediterráneo (*CAM: Mediterranean Savings Bank*)[10] increased the corruption perception in 2010 and 2011.

Seville had a similar situation (ERE [11]), what makes a similar perception for this subject in this city.

But for Madrid, the rest of cities displayed and the country average coincide with the trend shows in the *CIS* surveys. This mean that the media is reflecting what was happening at that moment.

Another thing to take into consideration is that, according to *CIS*, corruption did not seem to be among the main concerns for the Spanish people in the analysed years. In the subsequent years, this perception suffered an exponential increase in the *CIS* surveys,

[8] https://en.wikipedia.org/wiki/G\%C3\%BCrtel_case
[9] https://es.wikipedia.org/wiki/Caso_Brugal
[10] https://en.wikipedia.org/wiki/Caja_de_Ahorros_del_Mediterr\%C3\%A1neo
[11] https://es.wikipedia.org/wiki/Caso_ERE_en_Andaluc\%C3\%ADa

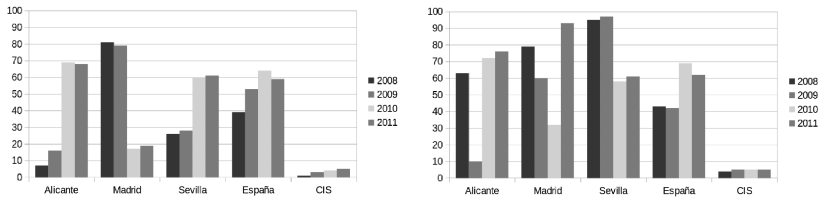

Fig. 1. Left. Corruption (left) and Right Education (right) temporal evolution according to the index of concern of the term *corrupción* in the *20Minutos* articles in the aforementioned locations.

reaching values over 60% in 2014, and around 50% in 2016. We should analyse whether this increase is due to corruption increase or to its pervasive nature in news media, which clearly affects citizens opinions. If we accept the latter reasoning, this would mean that we are able to predict the future citizens concerns through news media avoiding traditional surveys.

***Education* temporal evolution**

The evolution of the term *educación* in both the *20Minutos* articles and the *CIS* surveys is shown in the figure 1.

One more time, belonging each of the location analysed to different communities, and having different policies on education each of these communities because they have derived this competency, it can be observed large variations between locations. This variations depend on the measures and funding cuts adopted in each of their respective communities. Still, it can be seen as the concern for education is growing nationally in the journal texts as the economic crisis progresses, and therefore cuts. Finally, among the cities shown, a similar degree of concern is displayed, which is far above from the one shown in the *CIS* surveys.

According to the *CIS*, the degree of concern by the citizenship for education in the years studied was around 5%. This degree has grown steadily, doubling the initial figure in 2016. Again, if you look at the rate at the national level, it appears that what is shown in the press anticipates what would later collect *CIS* surveys. This may also be because the measures taken by the government, generally cuts in the middle of the financial crisis, which echoes the press, have not a final impact until later. After this time, it is when the public begins to notice the effect of the measures taken, and therefore to express concern.

***Unemployment* temporal evolution**

The evolution of the term *paro* in both the *20Minutos* articles and the *CIS* surveys is shown in the figure 2.

For the 4 years analysed, it should be noted that unemployment was the biggest concern among Spanish citizens based on *CIS* surveys. This time, the concern shown by the *CIS* matches that shown in *20Minutos* in the given cities and the country's average.

According to the *CIS*, unemployment concern was growing during these 4 years analysed. Something like the example of Alicante, which is the most similar to those listed in the chart to the national average, with the exception of the last year (2011) that

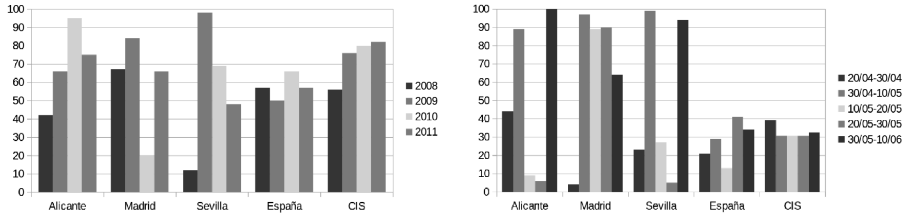

Fig. 2. Left.Unemployment temporal evolution according to the index of concern of the term *paro* in the *20Minutos* articles in the aforementioned locations.Right Corruption temporal evolution according to the index of concern of the term *corrupción* in the *Twitter* messages for the aforementioned locations.

began to decline. According to *CIS*, although unemployment remains the main concern of the Spanish people in 2016, it has decreased to levels that are between those shown for the years 2009 and 2010, which already could be observed on the results of the newspaper *20Minutos* nationwide.

Beyond that, the results shown by the media may affect the opinion of citizens, according to our study, these media seem to be a barometer of what would citizens opine in the future.

Twitter The *Twitter* corpus was composed by tweets collected in 52 different days. In order to be able to observe a temporal evolution of terms we decided to divide the corpus into five parts each of these divisions comprising periods of 10 days (the first and last period have 10 days plus the initial or final day respectively): 20/04/2013-30/04/2013, 01/05/2013-10/05/2013, 11/05/2013-20/05/2013, 21/05/2013-30/05/2013 y 31/05/2013-10/06/2013.

In the following sections, the evolution of the concerns, which were previously mentioned in the sample locations indicated in the figures, will be shown.

This time, it should be noted that since the *CIS* shows the results of their survey monthly, and because of each analysed period covers only 10 days, some of the *CIS* surveys values shown in the figures (the ones that comprise April) do not vary in several periods.

It also should be noted that because the time periods cover only 10 days, the results obtained from the standard score may considerably fluctuate.

***Corruption* temporal evolution**

The evolution in the presence of the term *corrupción* in *Twitter* between the indicated dates is displayed in the figure 2. This concern index is compared to that shown in surveys conducted by the *CIS* for this same time period.

In social networks, the average rate of concern for the term *corrupción* for the period of time indicated has been lower than the one shown in the newspaper *20Minutos* for the exposed years. Nevertheless, Spanish people concern about this subject has considerably increased (around 30%) as it can be observed in the figures 1 and 2.

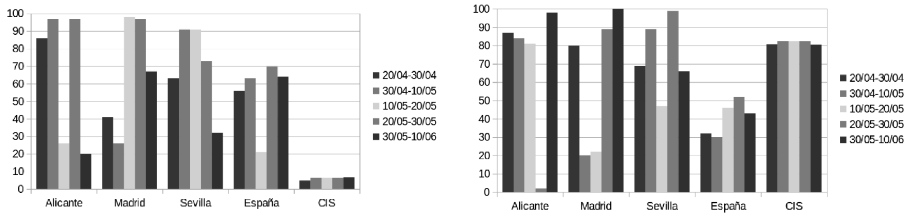

Fig. 3. Left. Education temporal evolution according to the index of concern of the term *educación* in the *Twitter* messages for the aforementioned locations. Right Unemployment temporal evolution according to the index of concern of the term *paro* in the *Twitter* messages for the aforementioned locations.

Among the displayed cities, it is visible how Alicante and Seville have a virtually identical trend for the term *corrupción*. On the other hand, this concern did not decrease in Madrid in the third and fourth period as it did for the other two locations.

We must highlight how the average score for all the cities is very similar to the achieved scores in the *CIS* surveys, especially when we calculate the average for the month of April, such as is made in the *CIS* survey. Thus, *Twitter* seems to accurately reflect what the *CIS* surveys show.

***Education* temporal evolution**

The evolution in the presence of the term *educación* in *Twitter* between the indicated dates is displayed in the figure 3. This concern index is compared to that shown in surveys conducted by the *CIS* for this same period of time.

This time, the term *educación* has obtained more differentiated values than with the previous term. The fluctuation experimented in the concern rate of this subject in *Twitter* has been significantly high, varying between 20 and 100 percentage points. This is due to the different meanings of the term '*educación*' in Spanish. The main senses of this term are education and good manners, more than academic education.

According to the graph shown in the figure 3, Alicante seems to be a meaningful sample that let us know how *education* has been perceived in the rest of the country in *Twitter*.

On the other hand, in accordance with the data shown in *CIS* and *Twitter*, concern for education is much higher in the social network than in the published survey. In *Twitter* do not seem to be one of the main concerns of citizens, as happened with that shown in the *CIS* results in Figure 3, where the results of the *CIS* surveys were compare with the results in the *20Minutos* newspaper. Despite this, comparing the *CIS* results from both graphs, it is indeed perceived to have a considerably increased from one date to another.

***Unemployment* temporal evolution**

The evolution in the presence of the term *paro* in *Twitter* between the indicated dates is displayed in the figure 3. This concern index is compared to that shown in surveys conducted by the *CIS* for this same time period.

Regarding the term *paro*, the data shows an unequal concern rate between the different cities. This time, unemployment is more relevant in the *CIS* surveys (it is clearly the main concern) than the average value obtained from the social network.

On the other hand, as previously mentioned, when such a short periods of time are analysed, great fluctuation can be found in the standard score values. If we omit the fourth time period, we can observe how this concern follows a very similar evolution between *CIS* and *Twitter* for the city of Alicante, as well as Seville if we instead omit the second period of time.

4 Conclusions

In this article, two different studies have been carried out which were intended to provide information on both formal and informal texts, the corpus of *20Minutes* and *Twitter* respectively.

In the experiment carried out to show the correlation between these corpora, the results show that there is a temporal correlation that reaches 90% in the top 10 largest cities of the country. This cities are in turn the ones from which a great number of tweets have been collected in the corpora.

This is a great result considering the difficulties of the task, such as: the different nature of the corpora, messages sent in *Twitter* usually refer to other places, little geographic information found in the tweets, or that the dates of the articles in the newspaper corpus were contiguous.

As for the experiments that showed the temporal evolution of the language, as it has been observed in the results obtained, it seems that what the press publishes, after a while, is reflected in the polls of the *CIS*. This can be useful to predict trends in the concerns that citizens will face in the future, either because the press anticipates these concerns or because it influences citizens.

Apart from that the results shown by news media can affect the opinion of the citizens, according to this study, these media seem to be a thermometer of what citizens will be concerned in the future.

With regard to *Twitter* and the temporal evolution of the terms, it would have to be followed in a longer period of time, since if these periods are divided into 10 days large shifts in the evolution of terms take place, especially when considered at city level.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
2. Benzécri, J.P., Bellier, L.: L'analyse des données, vol. 1. Dunod Paris (1976)
3. Bigi, B.: Using Kullback-Leibler distance for text categorization. Springer (2003)
4. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: Proceedings of the first international workshop on Location and the web. pp. 49–56. ACM (2008)
5. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics pp. 79–86 (1951)
6. Zhang, W.V., Rey, B., Stipp, E., Jones, R.: Geomodification in query rewriting. In: GIR (2006)