tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

EMS2021-245

www.tecnalia.com

Towards the operational use of tweets data in high impact weather scenarios: data mining and analytics in Basque Country

S. Gaztelumendi [1,2]

1- Basque Meteorology Agency (EUSKALMET). Parque tecnológico de Álava, Miñano, Araba, Basque Country.

2- TECNALIA BRTA (Basque Research and Tehcnology Alliance), Meteorology Area. Parque tecnológico de Álava. Miñano, Araba, Basque Country.

Abstract

In this contribution, we present different aspects related to the operational use of Twitter data in the context of high impact weather scenarios at local level. We present some results and experiences from a proof of concept project that demonstrate how different data mining and advanced analytics techniques, can be used in order to include social media data information for different operational tasks and particularly during severe weather events in a weather service context like the Basque Meteorology Agency.

Introduction

Although social media industry is nowadays a very congested Marketplace, Twitter continues to maintain its status as a popular social media platform with 330 million monthly active users and 145 million daily active users sending more than 6,000 tweets every second in the world (Businessofapps 2021). In Spain case, 85% population are social media users, with around 5 million Twitter profiles for a population around 47 million (Statista 2021). In the autonomous community of Basque country (CAE), with 2.17 million inhabitants in 7.234 km², we can estimate that around 70% of total population are social media users with a Twitter penetration of around 20%, after some inference from Spanish data, Eustat and a study of 'Basque Youth and Social Media' (EUSTAT 2020, Basque Youth Observatory 2018). It is noteworthy that Twitter is a social tool that enables users to post messages (tweets) of up to 280 characters supporting a wide variety of social communication practices including photo and video attachment. The Basque Meteorology Agency @Euskalmet with more than 115,3 K followers remain as one of the most popular accounts in Basque Country (Gaztelumendi et al 2013).

Twitter is not only an opportunity to instantaneously spread messages to people without intermediaries (Gaztelumendi et al 2013, 2015, Martija et al 2014, Palacio et al 2014), but also as a potential platform for valuable data acquisition using twitter API capabilities (Twitter 2021).

The most important activity in Euskalmet are actions in severe weather events (GV 2018, Gaztelumendi et al 2012, 2013). Before the event, mainly centered in forecasting and communication, during the event in nowcasting, surveillance and impact monitoring and after the event in post-event analysis (Gaztelumendi 2013, 2016). During all these complex processes, real time tweets posted by local users offer a huge amount of data that conveniently processed could be useful for different purposes. For operational staff, working at office during severe weather episodes, it is critical to understand in real time the local effects that a severe phenomenon is causing and the correct perception of the extent of impact and social alarm. For this purposes, among others, different information associated with posted tweets can be extracted and exploited conveniently.

Methodology and Data

In this paper we summarize our experience during a proof of concept project for automatic real time twitter mining and analysis and the key aspects to consider for the development of an operational tool for Twitter API data exploitation at local level. We present the main challenges and problems that we have had to face, including how to deal with the lack of geolocation information, since in the case of the Basque country, as in other parts of the world, tweets containing geotags are the exception, not the rule.

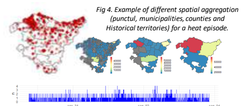
All data acquisition, data processing and analysis in the proof of concept project are carried out with different scripts prepared by the author in R environment (R Core Team 2020). Particularly useful in this context are the packages rtweet (Kearney 2019), tidytext (Silge 2016), tidyr (Wickham 2020), tm (Feineter 2020) and as usual dplyr (Wickham 2020) and ggplot (Wickham 2016).

Key findings and conclusions

- ✓ The increase in social interest around weather during severe weather episodes, is clearly reflected in social networks, despite the size of the territory (approx. 100x100km) and the inherent limitations of the Twitter API (theoretically only 1% of the live stream can be collected for free) (see fig 6)
- ✓ Twitter content examination and analysis in order to extract some derived quantities aggregated for different spatial and temporal scales could be the basis for an automatic surveillance and monitoring social system that could be useful on real or deferred time.
- ✓ It is important to note that different Knowledge aggregation levels are present from spatial, temporal and content characteristics and that information extraction is also possible at different levels of aggregation (see fig 4 and 5).
- ✓ Not only text content could be useful, available metadata (user characteristics, attached information, etc) could have an important role.
- ✓ Exact location information for a particular tweet (geo coordinates from where a tweet is send) is rarely available but in many cases must be inferred from location field at different level of spatial aggregations (municipalities, county, historical territories, etc..) (see fig2 and fig 3).
- ✓ As a general rule, each single tweet without any location information or outside the area of surveillance or impact interest (Basque country and surroundings) is not considered (see fig 4)
- ✓ All tweets are categorized for potential credibility based on users characteristics. Known users are categorized according to their general reliability considering different characteristics (official sources, media, weather regular contributors, etc.). Unknown users are considered at low level credibility until an analysis is made. As a general rule each single tweet content is considered inaccurate and suspicious.
- ✓ A plausible rule-based methodology could be implemented, for tweets text content and location mining, at a relatively low cost and serves as the basis for further and more complex developments.
- ✓ In spite of the general positioning of Euskalmet in the Basque Country and the high penetration of @Euskalmet in the Basque Twitter community, the officially used severe weather hagstaghs (defined in Basque language and theoretically the key for discussion and topic analysis) apparently have a limited impact outside the institutional users.
- ✓ Dictionaries / lexicons (topics, locations, etc.) must be implemented taking into account particular local idiomatic aspects (Basque and Spanish mixed words) and peculiarities of language usage in social networks (abbreviations, spelling mistakes, etc).
- ✓ Different metrics (e.g. increasing ratios of number of a particular topic tweets) could be real time monitored at different spatial and temporal scales as a sort of "social sensor" network. Population distribution and other socio-cultural aspects need to be included in order to extract conclusions in the side of impact
- ✓ In this PoC project, a preliminary real-time tweet collection and classification system is implemented, as well as some reduced monitoring and surveillance tools.
- ✓ Visual data analytics techniques are essential for rapid human interpretation and they might be actionable at real time.
- ✓ Both, fully automatic and supervised systems are needed for a full operational exploitation of available data.

Fig 3. Proportion of available data from different available fields (left) and new features after location module execution (right).

Fig 4. Example of different spatial aggregation (punctul, municipalities, counties and Historical territories) for a heat episode.



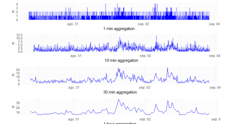


Fig 5. Example of different temporal aggregation for a storm episode.

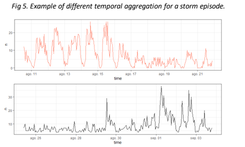


Fig 6. Timeline of number of tweets related with heat topic (top) and storm topic (down) during two severe weather periods in CAE

Acknowledgements

The authors would like to thank the Department of Security of the Basque Government and particularly to the Directorate of Emergencies and Meteorology for operational service financial support. We also would like to thank all our colleagues from DAEM, EUSKALMET, AZTI and TECNALIA for their daily effort in promoting valuable research and services for the Basque Society. We would also like to thank Twitter users, R community and all institutions and people that support open data and tools. This work has been partially funded by the LIFE-URBANKLIMA2050 project

EUSKO JAURLARITZA — GOBIERNO VASCO
HERRIZAINGO SAILA — DEPARTAMENTO DE INTERIOR
Herrizaingo Sailburuordetza — Viceconsejería de Interior
Larrialdiei Aurre Egiteko eta Meteorologiako Zuzendaritza — Dirección de Atención de Emergencias y Meteorología

Euskalmet — EUSKAL METEOROLOGIA AGENTZIA

URBANKLIMA2050

Results and Discussion

General Objective: To have better general and impact information before-during-after severe events that makes it possible to deal correctly with different Euskalmet's real time (RT) and delayed time (DT) operations in severe weather scenarios or other natural hazards events.

Final Goal: The implementation of an operational surveillance system of impact considering social media, news and emergencies data, fully integrated with actual monitoring and surveillance capabilities present in Euskalmet.

Context: Small country, highly concentrated population, bilingual with different penetration level of Basque Language. High usage of internet, connected mobile phones, and social network usage with an acceptable proportion of Twitter usage. Free Twitter API availability. Euskalmet presence in Twieter since 2010 with more than 117.000 followers and 64.000 posts. Euskalmet today performs some kind of social networks surveillance (particularly Twitter) based on human monitoring with tweetdeck (see more details in context section, on the right).

First Iteration: Lean development, successive iterations, minimum viable product, modularity, focusing on users, open source and low cost.

Proof of concept (PoC): in order to demonstrate the practical potential of Automatic real time twitter data mining and exploitation in real time operational tasks. Focusing on two main key aspects key aspects: data mining and analytics.

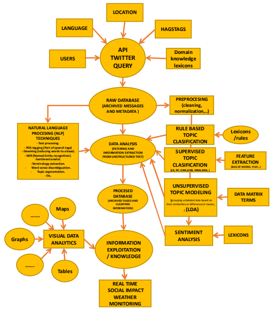


Fig 1. General conceptual framework.

Twitter mining and analytics.

Twitter, that originally was designed for effective and efficient two way communication, could be considered today as one of the simplest (and most redundant) public communications tool. Twitter provides high-volume, high-velocity and high-variety unstructured data (big data) that can be used to support decision-making (e.g. O'Leary, 2015), particularly in the Euskalmet case (Gaztelumendi et al 2015b) and around meteo-climatic business (Gaztelumendi et al 2016c).

There is substantial quantitative and qualitative information available for mining and analysis in Twitter, including number of tweets, number of retweets, number of followers and many others statistics and metrics (e.g. O`leary 2015). In addition, there is substantial non-numeric qualitative information in terms of text that we need to automatically convert in some kind of actionable quantitative data.

Twitter mining (e.g. O`leary 2015) and Twitter analytics (e.g. Kumar et al 2013) is concerned with providing structure to the unstructured data in order to extract and exploit information. Twitter text messages and metadata are our "data mine" and we mine that data for its potential usage in the field of local impact weather. For this purposes different modules and submodules are implemented for data acquisition, filtering, cleaning, geolocation extraction, topic classification, analysis of content and knowledge exploitation.

When dealing with content analysis, different approaches are possible (e.g. Kruspe et al 2021,Reuter et al 2017, Kuman et al 2013). In our case, different and specific lexicons has been prepared for different topics and in two different flavours: one for Twitter API querying and other for general message content mining. In this PoC, just fully human-based lexicons are prepared with the most frequents terms related with the so defined key subjects nearly the same as causes available in the warning/alert/alarm operational system (GV 2018, Gaztelumendi et al 2012). Lexicons are defined as bilingual, containing Basque as Spanish terms. Note that single queries to the Twitter API are limited to 500 characters so, if needed, multiples queries are performed.

Different typologies could be used to categorize social media messages, based on content and metadata, covering many dimensions as information provided, emotional content, source, credibility, time, location, etc. (see tab 4.1 in chapter 4 Castillo 2016 for more details). In our case, we include all metadata directly available for each tweet from the API (more than 90 fields), including time, geolocation (if present), and user id. We also include some relevant derived information from the different classification modules. In this PoC, mainly based on location and severe weather typology according to warning/alert/alarm operational criteria.

Such modules are executed sequentially is an automatic process that organizes and understands the large collections of available twitter data, by assigning "tags" or categories according to each individual tweet characteristics (mainly from field "location" and "text"). As a first iteration we implement different rule-based classification modules by directly programming a set of hand-made rules based on the content of textual fields. Defined rules are able to extract location information from the "location" and "text" fields and to discern between tweets of different topics by looking directly at semantically relevant elements of a particular "text" field content. Rules are defined using different lexicons (lists of words) and using different metrics for similarities as the highest frequency specific topics words.

Data provision module.
Set of Twitter APIs queries for data acquisition based on different R scripts and rtweet package executed recursively each 5 minutes. Queries are pre-filtered (at API level) by key-words and location. Final dataframes (DF) are conveniently merged and labelled.

User classification module.
Extract relevant user information for categorization purposes. Users are classified and labelled according to different categories. In the PoC just 4 users typologies are included: governamental/public, news/media, collaborators, others. In next steps further classification must be included with the final idea of characterizing potential credibility to a given content using derived information available from the user side (e.g. influence metrics). User related labels are included for each tweet in the tweets DF.

Location classification module.
Location data extraction at different aggregation level, directly from geotogag available data (exact punctual lon lat coordenates, box lon lat, etc) or indirectly extracting information from the text location field (municipality, county, province, etc..). Those tasks oriented to identifying the presence of key terms from different geographical lexicons prepared for this purpose. All the information available is included in the tweets DF.

Message content classification module
Categorization of Tweets messages with rule- based methodology (e.g. Berka 2020). Different keyword-based rules are used (as in a expert system) to separate messages into categories defined by the presence of different key-words in the message. Although it is generally considered an ineffective approach, is the simplest and most straightforward methodology, and this may work for certain information categories that have a small, well-defined, unambiguous set of terms that are highly discriminative (Castillo, 2016), as is the case. This module includes various Natural Language Processing techniques for text pre-processing (e.g. tokenization, stop words extraction, etc.)

Representation module.
At the end of the automatic process, messages are converted to a format suitable for automatic representation. Final DF contains original data and new features from text and metadata content analysis. Aply different visual data analytics techniques in order to achieve a better as possible inclusion of relevant information in the vigilance process. In the PoC just some basic graphics are generated.

Fig 2. Squemaric representation of PoC modules.

CONTEXT

Twitter and its structure.

Twitter general usage.

A "**tweet**" is a social media message posted on Twitter.com. It is restricted to 280 characters. Though most tweets contain mostly text, it is possible to embed links (URLs), pictures, GIFs GIFs, videos or emojis. Once a tweet is sent by a user it becomes immediately visible to all followers. If it is found worthy enough, it can be retweeted on to secondary followers. A **retweet** ('RT' for short) is a tweet by a user X that has been shared by user Y to all of Y's followers. Thus, retweets are a way of measuring how popular the tweet is.

This occurs either as a "**pure retweet**", where the entire tweet is re-messaged forward by one of the secondary followers, or as a "**modified retweet**," where a part of the tweet is forwarded or what is known as a "**user mention**" tweet where only the original Twitter username or "**handle**" (with the '@' symbol) is transmitted. The original Twitter handle may be mentioned in any subsequent tweets, so it is easily traceable back to the original message. Usernames or handles are recognized by the '@' symbol. A user can direct a message to another user by adding the handle, with the '@' symbol. An @, or a mention, is when you include somebody's @twittername in the tweet. The person will be alerted that you mentioned them. Use it to send a public "hey, over here," or to add somebody on a conversation that's currently happening.

Tweets usually contain components called "**hashtags**", which are words that capture the subject of the tweet. They are prefixed by the '#' character. Hashtags are clickable, too, so you can tap on a hashtag to see all the tweets related to that topic.

A user can '**favorite**' a tweet (analogous to a 'Like' on Facebook or instagram). A Reply on twitter means responding to a message or tweet from a person while to retweet is to broadcast (like forwarding an email) a tweet or message posted by a person to others.

There are two ways to reply to tweets. There is a @ reply where you use "@username" in your message. Such replies are public i.e. visible on your own Twitter page. If you want to reply privately, you can send a DM (direct message) which is sent only to the recipient like a private e-mail.

Once a message is retweeted either in an exact manner or user mentioned it becomes visible to their multiple new followers. This visibility allows for an exponential sharing of information that in turn can lead to a 'viral' dissemination of the original message. This viral propagation results in messages being widely popular, referred to in the Twitter jargon as a "**trend**" spread, becoming visible even by non-followers and by those at large on the Internet.

Twitter data.

In addition to the content of the tweets themselves, a few other relevant information such as username, time stamps, location, etc, can play an important role to understand the origination of tweets, profile inclination, and timeline analysis. Twitter company allows for the storing of tweets and will further publicly stream every tweet and its associated metadata of information. This metadata includes 20 fields of data that are sorted into three primary categories: location (including the time zone), biography (occupation-personal interest), and the tweet messages themselves.

Twitter APIs.

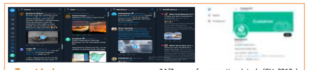
The Twitter APIs provides access to a variety of different resources, including information about conversation happening on Twitter. Twitter offers three API types: API Streaming (used to access public tweets in real time), Search API (allows access to the last 1500 tweets of the last 7 days) and API Rest (allows access to the central twitter data to the last 3200 tweets). Return APIs queries provide data encoded using JavaScript Object Notation (JSON) including Tweets and Users. This APIs provides access to a variety of different resources including the following: Tweets, Users, Direct Messages, Lists, Trends, Media, Places, Timelines, etc. It is important to note that using the free version of Twitter API it is impossible to retrieve historical data from the past as you can only go back to 7-days. The Twitter Search API in standard (free) version searches against a sampling of recent Tweets published in the past 7 days. The standard search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If completeness is important the premium or enterprise non-free search APIs should be considered. Note that recently, new version Twitter API v2 is going to be launched, and some changes may be introduced.

Although Twitter APIs can be used directly to programmatically retrieve and analyze a collection of relevant Tweets matching a specified query, usually some R (e.g. rtweet) or Python (e.g. Tweepy) packages are used for this purpose. Such packages make interacting with Twitter's APIs more approachable and data more accessible to researchers.

Two main approaches to data retrieval are possible:

Get historical: Requests are for a period of interest, with no focus on the real-time nature of the data. A single request is made, and all matching data is delivered using pagination as needed. This is the default mode for Search Tweets.

Polling or listening: Requests are made in a "any new Tweets since my last request?" mode. Requests are made on a continual basis, and typically there is a use case focused on near real-time 'listening' for Tweets of interest.



Tweetdeck

For management of @Euskalmet account we use Tweetdeck (GV 2018a) a free customizable web based Twitter tool for real-time tracking, organizing, and engagement (tweetdeck.com). Usually with the TweetDeck's dashboard we display separate columns of activity from @Euskalmet accounts. Operationally a window display is 24/7 open for operational task (GV 2018a) including real time surveillance of selected items (see fig b1). Usually, four timelines columns are 24/7 monitored, the Home, the User, the Mentions and the Notifications column, in order to have a correct perspective of what is happening around @Euskalmet. Note that in the Home column timeline we monitor tweets from 112, Traffic, Fireman, and other accounts relevant for impact.

Tab 1. Euskalmet promoted hashtags for severe weather

Basque	Spanish	Topic (Dealing with plausible or observed impact due to:)
[illegible]		yellow warning level
[illegible]		orange alert level
[illegible]		red alarm level
[illegible]	[illegible]	snow
[illegible]		ice
[illegible]		storms
[illegible]		lightning
[illegible]		hail
[illegible]		wind
[illegible]		sea waves
[illegible]		splashing waves
[illegible]		overtopping waves and coastal floods
[illegible]	[illegible]	coastal trapped disturbances local phenomena (known as "enbata" or "galerna")
[illegible]	[illegible]	floods
[illegible]		high temperatures
[illegible]		low temperatures
[illegible]		forestfires
[illegible]		fog or low visibility

Hashtags.

A hashtag is a metadata tag that is prefaced by "#", it is used on different social networks ([illegible], Instagram, etc.) as a form of user-generated tagging that enables cross-[illegible] of content sharing a subject or theme. Hashtags permit anyone on the Internet ([illegible] and non-followers) to access the tweet and to disseminate messages across [illegible] communities without restrictions. It serves as one of the most powerful [illegible] enhancing exponential messaging, in other words it enhances the opportunity for [illegible] go viral. Hashtags are clickable, so users can tap on a hashtag to see all the [illegible]

Meteorology and Emergencies – Department of Security – Basque Government) has defined specific hashtags for high impact weather [illegible] monitor, track, and consolidate information around [illegible] (Gaztelumendi et al 2012,2013, GV 2018b). [illegible] promote and use the predefined set of hashtags in [illegible] episodes in social media.

In our case, the so call "warning hashtags set" [illegible] events (GV 2018a), and are defined with the [illegible]

Social and cultural aspects in Basque Autonomous Community (CAE) .

Population Spatial Distribution.

The total population of the of the Basque Country Autonomous Community (CAE) (see fig a1) is around 2,200,000 with a mean density of 304.4 inhabitants/km2. 80% of the population of the CAE resides in urban areas, particularly in municipalities with more than 10,000 inhabitants (see fig a2). The distribution by territory is unequal (see fig a2 and a3), with 52% in the territory of Bizkaia, 33% in Gipuzkoa and 15% in Araba. Likewise, 35.2% of the population resides in one of the three capital cities. The population of Araba is highly concentrated, with almost 75% living in Vitoria-Gasteiz (see fig a2). In the Bizkaia case almost 70% lives around great Bilbao area (see fig xx). In Gipuzkoa, the population is more spread out than in the other territories, with nearly 50% living in small or medium municipalities with less than 20.000 inhabitants (see fig a3). See GV 2020 and EUSTAT (https://www.eustat.eus and) for more details.

Basque Language.

The Basque Language is one of the two official languages in the territory of CAE spoken by less than half of inhabitants, along with Spanish language spoken by nearly all population. Basque is a minority European original language that has resisted different uniformization and standardization historical periods (e.g. Urrutia et al 2008), nowadays it benefits from international minority language status. Basque Government adopt different measures leading to increasing the presence of Basque in the population and the media, tending towards the progressive and theoretically equality in the use of both official languages.

Today in CAE 41% speak Basque and another 15% are able to understand it (GV 2020), in Gipuzkoa 56%/13% , in Bizkaia 35%/16% and in Araba 27%/16% (see figure xx). In recent years, the number of bilingual people has risen significantly and an even sharper rise is expected in the future. In figure a4 we can see the actual distribution of active and passive speakers. The majority of bilinguals over five years old (66.3%) live in urban areas, mainly in the metropolitan areas of the three capital cities (Bilbao, San Sebastian and Vitoria) and in towns with more than 10,000 inhabitants. This is due to the demographic concentration falling mainly in urban nuclei (GV 2020)

The Basque language is extended far away the CAE, in fact covering a collection of territories inhabited by the Basque people, known as Euskal Herria (in Basque language "the land of Basque language"), that includes two Spanish autonomous communities (CAE and Navarre) and the French Basque Country with a total of seven traditional territories (see figure a1) with different degree of Basque language presence (always in minority).

The Basque language is a predominantly agglutinative language from a morphological viewpoint. It is an ergative or active language with a basic word order in which Subject precedes Object and Verb (SOV) (see more details in Igartua et al 2021). Spanish (Indo-European or Romance language based on its origins) is mosthly SVO, classified as somewhat inflectional because of the extensive use of word endings used to indicate attributes such as gender, number, and tense. Those differences are important for Natural Language Processing (NLP) needed for automatic tweets content analysis tools implementation.

Information habits and connectivity.

48.4% of Basque people from CAE are mainly informed by newspapers or digital media. Local information is preferred by 84.2% of Basques. 54.8% watch regional or local television. Although 26% do not trust generalist media, there is a high level of trust in radio and press (6.1 / 10) especially with regard to local information. Around 40% of CAE citizens use social networks for regular information (Periodismo.2030.com 2021).

91% of Basque household with a member aged between 16 and 74 have an internet connection. Computers are present in 70.9% of households, mobile phone is found in 95.2% of Basque households. From the point of view of the population, the percentages are very similar: 87.4% population have access to the Internet at home, 78.3% have a computer and 97.2% a mobile phone. Practically the entire Basque population that uses Information and Communication Technologies (ICT) has communicated via WhatsApp or other instant messaging applications (96.1%), has searched for information on goods or services (83.8%), used email (79.4%) or read news stories, newspapers or magazines (74.3%). See more details in Information society overview from Eustat (Eustat 2020)

Social networks in CAE.

In 2020, 62% of population (aged 13+) of Western Europe are active social media users, in Spain case users are 71% population (aged 13+) with 98% of them accessing via mobile phones (Hootsuite 2020). In CAE a large part of society interacts regularly via social networks, especially the 99% of younger age group (290.000 with 15-29 years old) with 51.4% of them using an average of 3 or 4 social networks and 97.5% connecting daily to 3 networks (BYO 2019). The most used social networks in CAE (see fig a5) are WhatsApp, Facebook, Instagram and Youtube, in second place are Twitter, Snapchat and Spotify (BYO 2019). Instant messaging applications are owned/used daily by 99%/98%, Youtube 99% /73%, Instagram 81%/74% and Facebook 58%/33% (BYO 2019). In the case of Twitter 37.7% of young people have it and 17.7% use it daily, while 60% never use it (BYO 2019). The social networks preferred by young people in the CAE are Instagram (52%), Facebook 17%, Youtube 16% and Twitter 10%. Young people who "cyber-act" use Twitter and Facebook to a greater extent than the average (BYO 2019).

81.5% of 15-29 year-olds social networks users from CAE mainly use Spanish, 15.0% mainly use Basque and 3.3% use English (BYO 2019). In Gipuzkoa, 30.0% of young people mainly use Basque when chatting or posting content on social networks; in the case of Alava and Bizkaia, the percentages are considerably lower (see fig a5) (BYO 2019). More than 80% of Basque speakers do not use Basque on social networks, as it is not only their own knowledge of the language that influences them, but also that of the other people they interact with on social networks (BYO 2019).

Twitter in CAE.

In Spain 18% of population use Twitter (Hootsuite 2020).Since 2012, Twitter has had a version in which the entire interface is offered in Basque. Although it is quite complicated to know the real presence of Basque on social networks (Elorriaga et al 2019), according to http://indigenoustweets.com/ (Bhroin 2015, Kegan et al 2015) with data from 2017, Basque language is the minority language with the largest presence on Twitter, with 17052 users and more than 110,000,000 Tweets sent since 2012 (nearly half in Basque language). UMAP.eus uses a tool that collects and filters the flow of those who use Basque on Twitter (umap.eus). According to UMAP data in March 2021, from 21,192 users analysed, 10,542 were active, posting 582,957 tweets during a month, with 40.6% exclusively in Basque (UMAP 2021). In Figure xx we can see the evolution of the number of annual tweets issued in recent years by identified Basque language users from umap.eus.

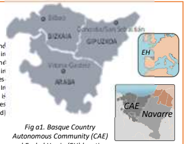


Fig a1. Basque Country Autonomous Community (CAE) and Euskal Herria (EH) location.

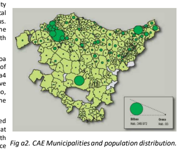


Fig a2. CAE Municipalities and population distribution.

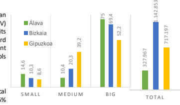


Fig a3. Total and small / medium / big municipalities population distribution among territories in CAE

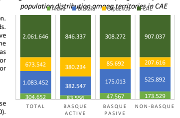


Fig a4. Basque language speakers distribution among CAE territories

Fig a5. Social network usage among young people in CAE

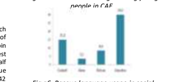


Fig a6. Basque language usage in social networks for young people in CAE.

References

- Agerri, R. Josu Bermudez and German Rigau (2014): "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland. http://ixa2.si.ehu.eus/ixa-pipes/papers/ixa-pipes-lrec2014.pdf
- Alam, F., Ofli, F., Imran, M. and Aupetit, M. (2018) A Twitter tale of three hurricanes: Harvey, Irma, and Maria. Available at: http://arxiv.org/abs/1805.05144 [Accessed 28th February 2019].
- [illegible]

Contact info:
santiago.gaztelumendi@tecnalia.com

TECNALIA
Parque Tecnológico de Bizkaia
C/ Geldo Edificio 700
E-48160 DERIO (Bizkaia) Spain
www.tecnalia.com

tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

BRTA BASQUE RESEARCH & TECHNOLOGY ALLIANCE

EMS Annual Meeting 2021
3–10 September
Online

EMS European Meteorological Society