Uncovering Social Changes of the Basque Speaking Twitter Community during COVID-19 Pandemic

Joseba Fernandez de Landa[1], **Iker García-Ferrero**[1], **Ander Salaberria**[1], **Jon Ander Campos**[2]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, [2]Cohere
{joseba.fernandezdelanda, iker.garciaf, ander.salaberria}@ehu.eus
jonander@cohere.com

Abstract

The aim of this work is to study the impact of the COVID-19 pandemic on the Basque speaking Twitter community by applying Natural Language Processing unsupervised techniques. In order to carry out this study, we collected and publicly released the biggest dataset of Basque tweets containing up to 8M tweets from September 2019 to February 2021. To analyze the impact of the pandemic, the variability of the content over time was studied through quantitative and qualitative analysis of words and emojis. For the quantitative analysis, the shift at the frequency of the terms was calculated using linear regression over frequencies. On the other hand, for the qualitative analysis, word embeddings were used to study the changes in the meaning of the most significant words and emojis at different periods of the pandemic. Through this multifaceted approach, we discovered noteworthy alterations in the political inclinations exhibited by Basque users throughout the course of the pandemic.

Keywords: Computational Social Science, Social Networks, Basque language

1. Introduction

In this constantly connected society (Castells, 2011), we are not exempt from the effects that remote communities generate in ours. Globalized problems such as climate change, nuclear accidents, pollution, war, refugees, and even pandemics, are becoming more frequent and widespread. These global challenges often transcend traditional boundaries of protection, leaving us in a state of uncertainty (Beck et al., 1992). Furthermore, there is an observable shift towards individualism as public institutions recede, thereby integrating us into a more globalized society (Bauman, 2013). The COVID-19 pandemic serves as an example of these trends. Therefore, we highlight the importance of conducting social research to understand the multifaceted impacts of such global incidents on specific communities.

Analysing the changes generated by the COVID-19 crisis has become a topic of main interest for many researchers as it can help in better understanding the new reality brought by the pandemic. Statistical analysis of virus infection levels has been one of the most used methods for modelling the trend of the disease. However, in this work we are focusing on the social change that COVID-19 has entailed. Understanding social changes is not an easy task and specially in a worldwide community where many different realities coexist. Moreover, the infection levels and restrictions taken by governments vary depending on the country, making global analysis misleading and dominated by greater communities. Thus, we focus on the Basque speaking Twitter community as all the users have shared similar restrictions and limitations during the different phases of the pandemic.

In recent years, social networks have become a mirror of society, and their use has greatly increased as a result of proposed health measures to combat the virus (Chakraborty et al., 2020). In addition, the ability to process massive data is greater than ever before due to current advances in hardware (Micikevicius et al., 2018). Along with this, neural network-based techniques have greatly developed the ability to obtain rich representations of words known as word embeddings (Mikolov et al., 2013; Devlin et al., 2019).

Therefore, monitoring public interactions in a social network such as Twitter provides an excellent opportunity to measure society's views on different events. In addition, the importance of social networks is even greater in times of change and they have shown their usefulness in analyzing the social effects of previous phenomena and actions (Buntain et al., 2016; Wang and Zhuang, 2017).

In this work, we want to analyze the response of the Basque speaking Twitter community to the pandemic of COVID-19 through the information provided by this social network, in order to better understand the impact of the pandemic on Basque society. To carry out this study, we have collected and analyzed the tweets posted by the Basque speaking Twitter community from September 2019 to February 2021 using different Natural Language Processing (NLP) techniques. Due to the different stages that the pandemic has experienced in the Basque Country, each one with its different restrictions and COVID-19 infection levels, we have distributed the collected tweets in different groups.

SIGUL2024 Workshop, pages 363–371
21-22 May, 2024.

This distribution enables us to analyze in much more detail the effect that the different events could have.

The main contributions of this work are the following ones: (1) We have collected and released the biggest dataset of Basque tweets ever, containing up to 8M anonymized tweets text from September 2019 to February 2021. The dataset is split over different pandemic stages enabling fine-grained and overall analysis of terms during period. [1] (2) We conducted an automatic exploration of the most representative terms during the different phases of the pandemic. Due to the combination of quantitative (frequency of use) and qualitative (meaning) analysis of those terms we are able to infer social phenomena from users' textual expressions. [2] (3) We spotted the change that the health crisis generated over people's main concerns. More specifically, we showed that general political issues have lost importance in favor of individual concerns.

2. Related Work

Since the beginning of the COVID-19 pandemic, many articles that monitor the activity of the Twitter social network have been published. Recent work has resulted in the creation of multiple datasets (Banda et al., 2021; López et al., 2020; Alqurashi et al., 2020; Chen et al., 2020a). These datasets typically contain tweets collected during the pandemic months of 2020 and 2021 and they tend to focus on the English language. Gathering English tweets enables us to collect huge datasets as the amount of English tweets is the biggest among all languages. However, as English is a worldwide spoken language, it brings difficulties when analizing social change due to all the different events that affect the English Twitter community. There are also some efforts that focus on smaller communities as the Arabic dataset presented by Alqurashi et al. (2020). All these datasets just extract tweets that contain COVID-19 related keywords as: "SaRS-CoV","COVID-19", "coronavirus"... and even if they are useful for many different tasks (Bullock et al., 2020) they do not offer information for analyzing social alterations caused by the pandemic.

In order to process unstructured text present on social networks, different NLP techniques (Chen et al., 2020b; Shahi et al., 2021) are used. To highlight the different themes treated around COVID-19, Chen et al. (2020b) use the Topic-Modeling technique by applying the LDA algorithm (Blei et al., 2003). The identified topics are visually represented through the UMAP dimension reduction technique (McInnes et al., 2018). In addition, general content analysis has also been performed on minority language scenarios such as Basque, applying Topic-Modeling (Fernandez de Landa et al., 2019) and interaction analysis (Fernandez de Landa and Agerri, 2021). Other studies use supervised techniques to analyze the content of social networks (Chen et al., 2020b; Shahi et al., 2021; Müller et al., 2020), also including Basque language (Agerri et al., 2021). However, in order to be able to train the supervised classification algorithms, previous manual work is needed, that is, an annotation expert must label different examples to be able to apply machine learning algorithms later on.

Analysis of changes in word semantics across time has been previously done by utilizing diachronic word embeddings. These embeddings have been applied for analyzing changes in culture (Hamilton et al., 2016), stereotypes (Garg et al., 2018) and political tendency (Azarbonyad et al., 2017). Similar methods were also used to model meaning change (Del Tredici et al., 2019) and to identify usage change of words across different corpora (Gonen et al., 2020). Closer to our case, Wolfe and Caliskan (2022) and Guo et al. (2021) use word embeddings in order to detect semantic changes in language on tweets related to COVID-19. Other approaches use contextual word representations (Devlin et al., 2019) to analyze the changes on the meaning of words inside specific sentences, instead of focusing on the word itself (Hu et al., 2019; Martinc et al., 2019). All those techniques are similar to ours, however, to the best of our knowledge, we are the first ones to apply this techniques into a controlled community over a specific phenomena such as the COVID-19 pandemic.

3. Data Collection

Twitter has been used as a great data source in order to analyze society and identify the latent dynamics that occur in it. This social network provides massive data for the analysis of small communities such as the Basque speaking one. Similar to any sample trying to represent social reality, ours also has a margin of error. Therefore, sample stratification problems such as age, socio-economic status or culture may occur if we extrapolate the results to the whole Basque society. Although we are able to extract information from the entire research population, our data collection is limited to Twitter users. Consequently, note that the references will center on Basque speaking Twitter community instead of Basque society. Data was gathered on February 2021 using the Twitter API.

[1] The collected data is publicly available here: https://github.com/joseba-fdl/basque_twitter_covid19_corpus

[2] Our code is publicly available here: https://github.com/ikergarcia1996/Ikergazte-Covid-Twitter-2021

As first step Basque speaking users were identified using *umap.eus* tool for Basque language monitoring in Twitter social network. This way, More than 10,000 Basque speaking users have been identified, obtaining 4M personal tweets and 4M retweets for a total of 57M tokens. Different from previous work, we consider all the tweets posted by the 10,000 Basque Twitter users and not just the COVID related ones. This decision is crucial for devising the impact of the pandemic on different aspects of society.

The collected data has been divided into five different periods or stages in order to enable a fine-grained temporal content analysis. As shown in Table 1, each division has been identified with striking moments of the pandemic that have heavily affected the Basque speaking Twitter community. In addition to that, start and end dates of each stage, as well as the distribution of tweets, retweets and word tokens are presented in the same table.

The different groups of the dataset are selected taking the following moments into account:

(0) First, a zero point has been set for the 2019 pre-pandemic era. This stage represents the moment when little or no information was known about the pandemic.

(1) This stage covers the period from the start of 2020 to the lockdown established by the Spanish government. In this period, people started getting infected with COVID-19 in the Basque Country and Spain, but no actions were taken by the authorities.

(2) The second stage consists of the duration of the lockdown order. Lockdown in Spain was defined as the obligation to stay at home, only being able to go out for essential things like buying food. After this moment, wearing a mask was compulsory.

(3) Stage 3 starts after the end of the lockdown era. This period was named as the *New Normality* and restrictions on mobility and social gathering gradually began to be lifted.

(4) Finally, the fourth stage starts when restrictive measures were again introduced due to a new increase of cases. This last stage finishes on February 2021, which was the data extraction date. In this phase important restrictions on social interactions (hospitality, gym, cultural acts...) and curfews were re-enabled in response to the increase of infections. Mobility between towns and cities was also reduced. We have named this stage as the *New Restrictions* period.

The collected data has been anonymized as the only available source is the textual one not keeping any metadata. This way, the authors of the tweets can not be tracked using our dataset, preserving the right to be forgotten. At the same time we keep user anonymity, we release a dataset based on pure text, permitting the reprodictuvity of the results as well as the use of this corpus as an informal Basque language data source.

3.1. Data Analysis

For data analysis purposes we have decided to take personal tweets and retweets into account, as these two elements are part of the content that each user makes public on their timeline. This way, this research is based on both texts of personal tweets and shared tweets (retweets). Apart from the words, that are the main component of the tweets, emojis have also been considered. These increasingly common emojis do not have an unambiguous dictionary definition, but they have their own meaning in certain contexts. Therefore, we study the frequency and meaning of different terms in order to analyze the effects of the pandemic on the Basque speaking Twitter community. We will also show how the use of terms has changed over time, while examining the impact of the pandemic on these changes.

We have carried out both quantitative and qualitative analysis using unsupervised NLP techniques grounded on the distributional hypothesis (Harris, 1954). On the one hand, we study how the frequencies of terms have changed over time, highlighting the terms that have become more and less mentioned as the pandemic has progressed. On the other hand, we have also studied the semantic changes that specific terms have undergone over time, showing the impact that the pandemic has had on the meanings of these terms.

3.1.1. Quantitative Analysis: Fluctuations in the Frequency of Terms over Time

The purpose of the quantitative study is to examine the terms with the greatest fluctuations of usage during the different pandemic stages. The quantitative study is based on the change of the frequency of the terms. We analyze the change of frequency using a linear regression over the frequency of the terms in the different dataset splits. We sort these values by the highest and lowest values to identify the terms with the biggest rise and biggest fall of usage.

First, we lemmatize all the terms using IXA pipes (Agerri et al., 2014) due to the great morphological richness of the Basque language. Suffixes and prefixes are very common and abundant in Basque and the same word can appear in very diverse forms. After lemmatization, as can be seen in Equation 1, we calculate the frequency of each

Stage	From	To	Tweets	Retweets	Word tokens
0. Before 2020	2019/09/01	2019/12/31	224,169	275,042	9M
1. Before lockdown	2020/01/01	2020/03/14	155,302	196,500	6M
2. Lockdown	2020/03/15	2020/06/21	296,627	349,368	12M
3. New normality	2020/06/22	2020/10/24	343,372	362,279	13M
4. New restrictions	2020/10/25	2021/01/31	415,388	347,533	14M

Table 1: Distribution of extracted tweets in Basque over different stages of the pandemic in the Basque Country.

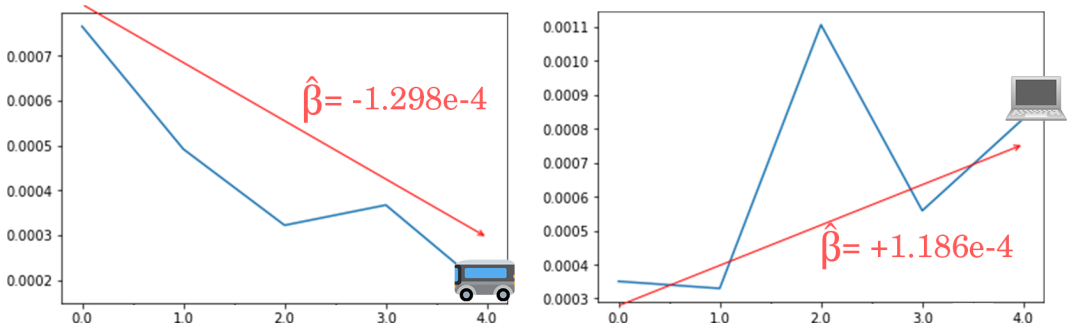


Figure 1: Laptop (💻) and bus (🚌) emoji trend. The Y-axis represents word frequency and the X-axis represents the different stages of the pandemic.

term for each dataset split that corresponds to a different moment of the pandemic. We calculate five different frequencies for each term, one for each dataset split. To calculate the trend of the term, we solve the Equation 2 linear regression system. The values $x_0..x_N$ represent the time splits and the values $y_0..y_N$ represent the frequency of each term in each time split. N is the total number of time splits. We use this linear regression to calculate the slope ($\widehat{\beta}$) of each term , which is an indicator of the trend of that term during the pandemic.

$$y = \frac{\textit{Number of tweets in which the term appears}}{\textit{Number of tweets}} \quad (1)$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad (2)$$

A positive slope or trend ($\widehat{\beta}$) means that the term has increased in use during the pandemic while a negative value means that the term usage has decreased. We rank all the terms described in the corpus according to their tendency. The 15 terms with the highest upward trend, and the 15 terms with the highest downward trend can be seen in Table 2. As an example, Figure 1 shows the trends of the laptop (💻) and the bus (🚌) emoji. The usage of the 💻 emoji has increased during the pandemic (especially during times when tougher restrictions were imposed) while the 🚌 emoji usage has decreased.

Terms that have increased in use can be seen in Table 2a, some of which are directly related to the pandemic like health-related terms (*covid, measure, health, pandemic, vaccine, positive, case, care, virus, #covid19*). In addition, we also have terms indirectly related to the pandemic (*online, confinement, hospitality, mask*) corresponding to some side effects such as: the increase in online communication, the reduction in hospitality and opening hours, the use of the mask in everyday life... Finally, the increase in the frequency of the word *crisis* can also be seen as a way to define the situation itself. Thus, most of the terms with the highest positive variability are directly related to pandemic issues, showing the impact of the pandemic on the Basque-speaking Twitter community.

On the other hand, Table 2b shows the terms with the most significant drop in usage. These terms are mainly related to political issues (*strike, feminist, Altsasua, pension, women, Catalonia, demonstration*) and collective initiatives (*presentation, conference, organize, lecture*). Thus, it can be confirmed that there has been a significant decline in the usage of political terms that were previously common on the social network. Feminism (*feminist, women*), economics (*strikes, pensions*) and other political issues (*Catalonia, Altsasua*) have lost their importance in the Basque community as the focus has changed to the pandemic. It also seems that terms related to political action or proclamations have lost their significance. This shows a significant loss of

Term		Trend
covid	*covid*	7.31
neurri	*restriction*	6.82
osasun	*health*	6.17
pandemia	*pandemic*	6.13
txerto	*vaccine*	5.02
positibo	*positive*	3.77
online	*online*	3.44
kasu	*case*	3.20
zaindu	*take care*	3.07
konfinamendu	*confinement*	2.80
birus	*virus*	2.79
krisi	*crisis*	2.78
ostalaritza	*hospitality*	2.75
#covid19	*#covid19*	2.70
maskara	*mask*	2.60

(a) The greatest positive variability.

Term		Trend
aurkezpen	*presentation*	-4.60
greba	*strike*	-4.43
feminista	*feminist*	-4.42
jardunaldi	*conference*	-4.23
Altsasu	*Altsasua*	-4.14
antolatu	*organize*	-3.83
pentsio	*pension*	-3.80
hitzaldi	*lecture*	-3.48
emakume	*women*	-3.22
elkartasun	*solidarity*	-3.20
Katalunia	*Catalonia*	-3.16
areto	*hall*	-3.11
aurkeztu	*presented*	-3.11
egitarau	*program*	-3.09
manifestazio	*demonstration*	-2.79

(b) The greatest negative variability.

Table 2: Variability in term usage over time.

importance of both political theory and practice, especially in Twitter, a social network with strong links to political demands and citizen protests.

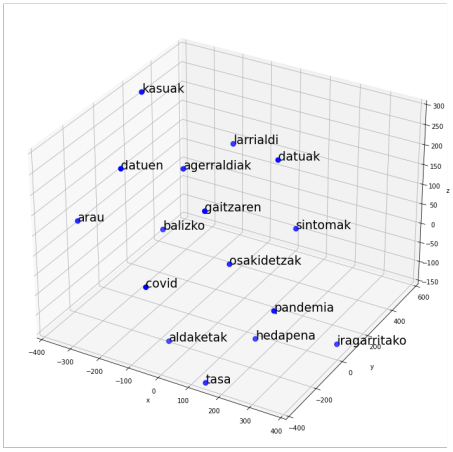
In summary, it is striking that the use of certain politically powerful concepts has decreased, while concepts such as health have gained a central place. Also, some words that have increased in frequency are related to practices that weren't common but have become everyday life, moving from abstraction to close reality. In addition, the frequency of various terms related to the restrictions or measures takes by the government has increased: the need to wear a mask, the permission to stay in bars or maximum number of people that can gather together, the way to communicate at a distance or the order to be locked up at home. It can be said that the focus has shifted to issues related to biopolitics (Foucault, 2009), that is, the regulation of human actions in everyday life. This concept alludes to measures imposed by governments or other power mechanisms that aim at regulating people's lives in their most personal and private facet. Following this reasoning, the presence of this kind of words manifests society's concerns about these restrictions, which seem to be understood as a form of control over their decision-making capacity as individuals. This way, the Basque speaking community in Twitter has shifted from focusing on general issues to focusing more on actions that affect everyday life.

3.1.2. Qualitative Analysis: Fluctuations on the Meaning of Terms over Time

The purpose of the qualitative study is to examine how the change in the meaning of terms has developed across time. Words change their meaning according to the needs of society, adapting their language to specific situations. In order to know which changes happened during the current pandemic we have used word embeddings. These word embeddings have the capability to represent semantics based on the distributional hypothesis (Harris, 1954). In this section, our intention is to generate different word embeddings for each stage and analyze whether the characteristics of terms have changed over time. We use emojis as words during the whole analysis as they are part of the usual vocabulary in social networks.

In order to represent the meaning of words and emojis, we use word2vec (Mikolov et al., 2013) and we obtain dense vector representation. Static word embeddings are used in order to capture the general meaning of the word across time. Thanks to vector representations we can get semantically similar terms, as similar terms have similar representations in the vector space. This way, vectors close to a given term can be used to identify words that are similar, that is, words that have a similar meaning. As words around each term define their meaning, we have computed word embeddings for each time period. For each stage we save the closest words of a given term and check whether there have been any changes between stages.

In order to find out how the meaning of words has changed over time, we have obtained a vector representation of words for 5 different stages. Each of these will combine the semantic features of a stage creating independent representations. To create each dense representation, we use the CBOW method with a 5-token window and 100 dimensions. Thus, we obtain 5 different instances of dense vector representations, placing terms in the corresponding vector space according to the stage and context in which the term was used. An example of the results obtained with this technique



Agerraldiak (*appearances*), aldkaetak (*changes*), arau (*rule*), balizko (*valid*), datuak (*data*), datuen (*of data*), gaitzaren (*of illness*), hedapena (*expansion*), iragarritako (*predicted*), kasuak (*cases*), larrialdi (*emergency*), Osakidetzak (*Osakidetza: Basque public healthcare system*), pandemia (*pandemic*), sintomak (*symptoms*), tasa (*rate*).

Figure 2: Closest words to the term *Covid* during the 3rd stage. Below, translations of the terms can be found.

can be seen in Figure 2, which shows the representation of the word *Covid* and the 16 semantically closest words. In this way, words related and similar to the chosen term are obtained, which will help to define the meaning of the word *Covid* in the 3rd stage.

To perform this qualitative analysis, we selected those terms that have experienced a significant increase in the frequency of use, and that experienced a clearer meaning change: positibo (*positive*), kasu (*case*) and segurtasun (*safety*). The emoji of the mask (😷) has also been chosen for the qualitative study, as it is among the emojis with the highest use frequency variation. Then, to understand each term's connotation, semantically similar words have been obtained using dense word vector representations. Similar words will define the meaning of the selected term. To illustrate how terms' connotations have changed through time, we have selected 5 similar words for each stage, as it can be seen in Table 3.

By analysing the term *positive*, it can be seen that at stages 0 and 1, it is related to many different words (*technique, difficulty, concept, h5n8, reason*..). At stages 2, 3 and 4, surrounding words have changed to terms such as *infect* and *coronavirus*, highlighting the effect of the pandemic in the meaning. During the pandemic era, this term has been used to define people who have been infected with the disease, being totally correspondent to the meaning of the term at stages 2, 3 and 4.

The term *case* at stages 0 and 1 is related to words like *affair* or *account* and also to words related to time (*moment, time, current*). On the contrary, similar words change at stages 2 to 4 showing again relations with the pandemic (*coronavirus, cases, infected*) are present 2, 3 and 4. In addition, it should be noted that the word *positive* is the closest, probably due to the appearance of the bigram *positive case*. Once again, we show that the term has now a direct relationship with the issues of the pandemic.

At stage 0 *safety* is related to words like *law*, *administration* or*system*, terms related to management. As it progresses, at stage 1 the meaning changes to words related to control (*control, to control, reduction*...) but always related to the pandemic (*coronavirus*). It should be said that from stages 2 to 4 the term has been related to words like *prevention* and *hygiene*, closely related to self-control, again showing a close relationship with the concept of biopolitics previously mentioned. In this case, the term has more relation to the regulation of daily life actions than to health status, showing a direct relationship with the impact of the pandemic on everyday life.

Regarding 😷 emoji, at stages 0 and 1, this emoji appears associated with terms related to environmental pollution (*#pollution, filter, chimney, spill, fog*...). As we move forward in time, the meaning changes again in stages 2 and 3, as they appear alongside words directly related to the pandemic (*capacity, hydroalcoholic*...) and with the need to wear the mask to avoid disease infection (*avoid, compulsory, #alwaysmask*...). Thus, the meaning of the emoji has also changed, from environmental pollution related topics to the pandemic, once again shifting to issues related to the regulation of everyday life.

Positive, *case*, *safety* and 😷 terms are excellent indicators of the situation, while they are terms directly related to pandemic issues, the changes in meaning are clearly visible. Although one might expect such changes based on common sense, we are able to demonstrate via a qualitative analysis that the previous meanings have been modified in a specific time period. Thus, this methodology is able to show the meaning of the selected term at each stage, giving the capacity to detect the moment and matter of the modification. The analysis has shown that the changes in meaning over time are closely linked to the pandemic. Those changes in the way Basque speaking Twitter users express themselves can be a sign of meaningful alterations. The modification of the written expressions is a way to show significant variations of the popular imagination of Basque users generated by the pandemic. Specifically in the terms *safety* and 😷, the changes in meaning are again closely linked to biopolitics, as they focus on concepts related to regulation of

Term		Related words on each stage
positibo (*positive*)	0.	teknika (*technique*), zailtasun (*difficulty*), kontzeptu (*concept*), ikusmen (*vision*), gertakizun (*event*)
	1.	h5n8, arrazoia (*reason*), egoiliarri (*resident*), aktiboko (*active*), ontzat (*okay*)
	2.	kutsatu (*infect*), koronabirus (*coronavirus*), kasu (*case*), PCR, infektatu (*infect*)
	3.	koronabirus (*coronavirus*), negatibo (*negative*), kutsatu (*infect*), positiboen (*positive*), PCR
	4.	kutsatu (*infected*), koronabirus (*coronavirus*), ospitaleratze (*hospitalization*), biztanleko (*per capita*), atzemandako (*detected*)
kasu (*case*)	0.	afera (*affair*), galdera (*question*), une (*moment*), kontu (*account*), zentzu (*sense*)
	1.	oraingo (*current*), garai (*time*), mota (*type*), legegintzaldi (*legislature*), afera (*affair*)
	2.	positibo (*positive*), koronabirus (*coronavirus*), kasuak (*cases*), PCR, kutsatu (*infect*)
	3.	positibo (*positive*), koronabirus (*coronavirus*), proba (*test*), kutsatu (*infected*), test
	4.	positibo (*positive*), kasuak (*cases*), test, hildako (*dead*), kutsatu (*infected*)
segurtasun (*safety*)	0.	sistemak (*systems*), hondakinen (*waste*), murrizteko (*reduction*), administrazio (*administration*), legearen (*law*)
	1.	prebentzio (*prevention*), kontrol (*control*), murrizteko (*reduction*), koronabirusak (*coronavirus*), kontrolatzeko (*to control*)
	2.	prebentzio (*prevention*), distantzia (*distance*), higiene (*hygiene*), errespetatu (*respect*), beharrezko (*necessary*)
	3.	prebentzio (*prevention*), higiene (*hygiene*), zorrotz (*strict*), neurriekin (*measures*), protokolo (*protocol*)
	4.	prebentzio (*prevention*), higiene (*hygiene*), malgutu (*adjust*), ezarritako (*established*), mugikortasun (*movility*)
😷	0.	#kutsadura (*#pollution*), albistegitan (*in the news*), #nipenanigloria (*#neitherpitynorglory*), #bizitzaerdigunera (*#lifeinthecenter*), Margaret
	1.	isurketa (*spill*), filtro (*filter*), argindar (*electricity*), tximinia (*chimney*), laino (*fog*)
	2.	saihesteko (*avoid*), besteekiko (*others*), musukoa (*mask*), maskara (*mask*), derrigorrezkoa (*compulsory*)
	3.	#maskarabeti (*#alwayswearmask*), aforo (*capacity*), #euskotrenmetrobilbao (*#train&underground*), edukiera (*capacity*), hidroalkoholikoa (*hydroalcoholic*)
	4.	bidalketa (*submission*), #htxonline, #getxo, #udalsarea2030, #amasavillabona

Table 3: Selected terms and related words over time.

everyday life (*control, to control, reduction, prevention, hygiene, avoid, compulsory, #alwaysmask*...).

4. Conclusions

This work examines the impacts of the COVID-19 pandemic on the Basque-speaking Twitter community, identifying significant changes in the ways of expression reflected in the textual data. The results generated may not fully represent the social reality, since the analyzed sample, despite being a large sample, is conditioned to the use of Twitter social network. While the results are not totally transferable from our selected sample into the entire Basque society, it can be said that they show some symptoms that affect many sectors of the general public.

With the intention of uncovering those variations, we carried out a massive collection of the available data from each of the Basque speaking community users that we identified. Our dataset generation strategy involved data collection and curation of tweets in the Basque language, resulting in the creation of the largest datasets in this minority language. This resource not only facilitates further research but also serves to amplify the visibility of the Basque language within the academic community.

Employing unsupervised Natural Language Processing (NLP) techniques allowed us to uncover significant transformations in language usage. Through a combination of quantitative analysis, tracking term frequency variations over time, and qualitative examination, utilizing dense word vectors to elucidate shifting word and emoji meanings, we are able to detect linguistic variations.

Fluctuations in word usage frequency and semantic meanings underscore the influence of the pandemic, showing how certain terms and symbols have significantly evolved. Moreover, the shift from discussions centered on general political matters to a focus on individual freedoms reflects a broader societal adaptation towards personal concerns, away from traditional political discourse. Nevertheless, these phenomena may be temporary, specific to the circumstances of the pandemic. Investigating

the long-term effects of these occurrences presents an interesting avenue for future research.

5. Limitations

Our research is constrained by its use of static word embeddings and frequency variations. While we acknowledge the existence of more sophisticated algorithms for learning unsupervised word representations, our technique demonstrates the capability to detect changes in word usage reflective of broader social shifts. The simplicity of our approach enables easy replication of experiments across various languages and contexts.

One limitation is our focus solely on a single small language and community. Although this choice facilitated analysis within a geographically confined community, our findings would hold greater significance if conducted across multiple small global communities.

In any case, the results that we have shown were reached due to our selected techniques, as evidenced by the linguistic shift observed among Basque users influenced by the pandemic.

6. Acknowledgments

This work has been partially supported by several MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) Disargue (TED2021-130810B-C21) and European Union NextGeneration EU/PRTR; (iii) AWARE (TED2021-131617B-I00) and European Union NextGeneration EU/PRTR. (iv) DeepR3 (TED2021-130295B-C31) and European Union NextGeneration EU/PRTR. This work has also been partially funded by the LUMINOUS project (HORIZON- CL4-2023-HUMAN-01-21-101135724).

7. Bibliographical References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, volume 2014, pages 3823–3828.

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Álvaro Rodrigo. 2021. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1509–1518, New York, NY, USA. Association for Computing Machinery.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Zygmunt Bauman. 2013. *Liquid modernity*. John Wiley & Sons.

Ulrich Beck, Scott Lash, and Brian Wynne. 1992. *Risk society: Towards a new modernity*, volume 17. sage.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Joseph Bullock, Alexandra Luccioni, Katherine Hoffman Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. 2020. Mapping the landscape of artificial intelligence applications against covid-19. *Journal of Artificial Intelligence Research*, 69:807–845.

Cody Buntain, Jennifer Golbeck, Brooke Liu, and Gary LaFree. 2016. Evaluating public response to the Boston Marathon bombing and other acts of terrorism through Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Manuel Castells. 2011. *The rise of the network society*, volume 12. John wiley & sons.

Tanusree Chakraborty, Anup Kumar, Parijat Upadhyay, and Yogesh K Dwivedi. 2020. Link between social distancing, cognitive dissonance, and social networking site usage intensity: a country-level study during the COVID-19 outbreak. *Internet Research*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020a. COVID-19: the first public coronavirus twitter dataset. *CoRR*, abs/2003.07372.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020b. Tracking social media discourse about

the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseba Fernandez de Landa and Rodrigo Agerri. 2021. Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

Joseba Fernandez de Landa, Rodrigo Agerri, and Iñaki Alegria. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6):212.

Michel Foucault. 2009. *Nacimiento de la biopolítica: curso del Collège de France (1978-1979)*, volume 283. Ediciones Akal.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Yanzhu Guo, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. How covid-19 is changing our language: Detecting semantic shift in twitter word embeddings. *arXiv preprint arXiv:2102.07836*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Christian E. López, Malolan Vasu, and Caleb Gallemore. 2020. Understanding the perception of COVID-19 policies by mining a multilanguage twitter dataset. *CoRR*, abs/2003.10359.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed Precision Training. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media*, page 100104.

Bairong Wang and Jun Zhuang. 2017. Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. *Natural hazards*, 89(1):161–181.

Robert Wolfe and Aylin Caliskan. 2022. Detecting emerging associations and behaviors with regional and diachronic word embeddings. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 91–98.