Extended Analysis of the Semantic Shift in Diachronic Word Embeddings for Spanish Before and After COVID-19

Esteban Rodríguez-Betancourt
Universidad de Costa Rica, Posgrado en Computación e Informática
San José, Costa Rica, 10501
esteban.rodriguezbetancourt@ucr.ac.cr

and

Edgar Casasola-Murillo
Universidad de Costa Rica, Escuela de Ciencias de la Computación
San José, Costa Rica, 10501
edgar.casasola@ucr.ac.cr

Abstract

Words can shift their meaning across time. This article shows the results obtained by the exploratory analysis of the semantic shifting on Spanish vocabulary using Diachronic Word Embeddings. Diachronic data consists of a 2018 Spanish corpus, before the COVID-19 outbreak, and a second corpus with documents from 2021. This paper addresses the construction of the diachronic Spanish word embeddings model, as well as the results obtained by the analysis using a non-supervised distance vector technique. In addition to the results shown in [1], this extended article discusses additional topics with the most semantic shift between those periods, like video games, cryptocurrencies, COVID-19 testing, COVID-19 vaccines, sanitizers, sports (soccer, basketball, cycling, tennis), pornography and K-Pop, in addition to COVID-19, masks, and vaccines treated in the first article.

Keywords: Linguistics, natural language processing, natural languages, pragmatics

1 Introduction

Semantic shift refers to the phenomenon of word meanings changing over time [2]. While the traditional approach to studying semantic shift involves analyzing long periods, recent resources like Google NGrams and Twitter have made it possible to analyze shorter time spans [2]. In this study, we extend the analysis initiated by our previous work [1] and explore the semantic shifts that occurred in Spanish vocabulary due to the COVID-19 pandemic.

With access to vast amounts of data, automatic analysis of semantic shifts has become possible [3]. Distributional methods can be used to detect this shift, by measuring the change in context where the words are used [2]. In this work, we analyzed millions of Spanish web pages collected by the CommonCrawl project[1].

To measure semantic shift, we used diachronic word embeddings trained on two corpora: one from before the COVID-19 outbreak, and another from 2021, after the pandemic emerged. Word embeddings are a prominent technique in natural language processing for capturing the meaning of a word. Several algorithms can be used to generate these embeddings. In general, they involve building a dense vector representation of a word from neighboring words [4]. These representations can be used in a variety of natural language processing tasks, such as information retrieval, classification, text generation, and analogy resolution [5].

[1] Available at `https://commoncrawl.org/`

Given that word meanings change over time, it follows that the relationships between words as defined by word embeddings will also change if a model is trained on corpora collected at different times [6]. Embeddings generated using diachronic corpora are referred to as diachronic embeddings.

Once words are represented as vectors, semantic shift can be measured using techniques such as cosine distance between the word embeddings of the same word at two different times [6][7].

The idea behind the previous calculation is that, by considering the distributional hypothesis, word embeddings can represent the meaning of a word, or more precisely, a concept. If we train two sets of word embeddings using different corpora, each model would map the same word to two different embeddings. The cosine distance allows us to measure the angle between two vectors. A cosine distance of 0 means that the vectors are the same, while a cosine distance of -1 or 1 would indicate that they are completely opposite. In this article, we are interested in identifying words that exhibit the largest angle between the embeddings trained with text from before and after the COVID-19 outbreak. For simplicity, the previous description omits the alignment step, which can be addressed using Orthogonal Procrustes [8], as explained by [7].

Because of these changes in word meaning, models of word embeddings that take temporal information into account have been proposed. Zijun et al. [9] proposed a method of word embeddings that is conditioned on temporal location. Hongyu et al. [10] proposed a model of embeddings that is conditioned on time and place, as well as the relationships between words. They intended to identify cultural trends and specific situations in different locations.

This research analyzes changes in the distances between words during the period surrounding the COVID-19 outbreak, comparing two word embeddings models trained on corpora collected in different years (2018 and 2021). One of the main contributions of our research is the construction of a diachronic Spanish corpus collected from the open internet, specifically for studying semantic shift during the COVID-19 pandemic. It should be noted that studies of this kind are relevant and recent areas of research [11][12][13]. For example, [11] studied COVID-19 neologisms in English using manual observations. In [13], frequency analysis was used to study semantic shift caused by COVID-19 in English. In [12], semantic shifts in English tweets related to COVID-19 were investigated. Although some works, such as [14], generated COVID-19 specialized word embeddings in Spanish, no studies were found on semantic shift in Spanish during this period. For this reason, exploratory studies in Spanish are necessary. Due to the increasing use of word embeddings in machine learning models that are used in our daily lives (search engines, content filtering, voice assistants, chatbots, etc.), it is important to understand how changes in language or our customs could render these models obsolete, requiring them to be retrained or losing effectiveness.

The following sections will explain the concept of word embeddings. Then, the process of obtaining the diachronic corpus (2018 and 2021), training of the word embedding model and its alignment will be explained.

Subsequently, the topics with the greatest semantic shift will be detailed, including various topics such as COVID-19, sports, electronics, singers, cryptocurrency and others. For each topic, the nearest neighbors in 2018 and 2021, the words that moved closer and farther away, and the shift of those topics relative to emotions will be explored. Additionally, a similar analysis will be conducted for specific words, such as “coronavirus”.

Finally, recommendations will be presented for future work related to word embeddings and possible application domains.

2 Background

To develop this work, different information retrieval and natural language processing techniques were used. The following sections will describe how these techniques were applied.

Word embeddings are a dense vector representation of the “meaning” of a word in latent space [15]. While there have been earlier vector representations, such as one-hot encoding, in this case we consider word embeddings to be dense representations that reduce the meaning of each word to a fixed number of dimensions (e.g., 300 or 100 dimensions).

These vector representations have interesting properties that are useful for natural language analysis. For example, Bengio et al. [16] describe how words with similar meanings tend to be close together in the vector space. In the work of Mikolov et al. [5], examples of analogy resolution using word embeddings are shown. This type of analogy resolution is illustrated in 1.

$$V[king] - V[man] + V[woman] \sim V[queen] \tag{1}$$

There are many algorithms to associate a word to a word embedding. One of the most recognized is Word2Vec [4], which uses the weights in the hidden layer in a neural network to represent each word. That neural network is trained to predict the surrounding context words given a central word (Skipgram) or a central word given its surrounding context words (CBOW). There is also GloVe [17], which leverages the statistics information of words. Another model is fastText [18], that uses subword information, allowing it to generate embeddings for unknown words. There are also improvements in the computational performance. For example, pWord2Vec [19] parallelized Word2Vec for architectures with many cores and fast RAM access between cores. BlazingText [20] adapted similar techniques to GPU and distributed execution.

Previous models associate a word to a single vector. But words can have different meaning given a different context. There are models like Universal Sentence Encoding [21] and Bidirectional Encoding Representations from Transformers (BERT) [22] that are capable of returning different embeddings based on the context the word is used.

Words can change their meaning through time, reflecting both language and society shifts [2]. A way to analyze those changes is using diachronic word embeddings [2], that are word embeddings built from a diachronic corpus. A diachronic corpus is a corpus that contains documents from several moments. In this research, the embeddings were built using BlazingText [20] with documents collected in week 25 of 2021, after the COVID-19 outbreak, and week 51 of 2018 (before any known mention of COVID-19). The data was collected from the project *CommonCrawl* (`https://commoncrawl.org/`), that periodically crawls documents from the open internet and makes them publicly available.

3 Methodology

The process of collection, embedding construction, and analysis was as follows:

1. Download a corpus of documents available on the internet in week 51 of 2018 and week 25 of 2021. The documents were downloaded from CommonCrawl, discarding the documents with less than 40% of its content in Spanish.

2. Train a word embeddings model using the corpus generated in step 1.

3. Align the word embeddings using the Orthogonal Procrustes algorithm, as proposed by [7].

4. Identify the words with the biggest semantic shift. For doing that, the cosine distance between the 2018 word embeddings and the 2021 word embeddings were calculated for each word. The words with a cosine similarity of less than 0.7 were chosen for the next step.

5. Determine which topics or areas had the greatest semantic shift using automatic clustering. A manual selection was done to discard irrelevant or incoherent topics (e.g., source code or gibberish from binary files).

6. For each selected topic, analyze the neighbors from the word embeddings in 2018 and 2021, and identify the greatest shifts that moved closer or farther. A similar analysis was done for some specific words.

7. Analyze the emotional shifts of the selected topics according to the emotional hierarchy of Shaver et al. [23] using the Word Mover's Distance (WMD), following a methodology similar to that proposed by [24].

3.1 Corpus Download

The documents from week 51 of 2018 and week 25 of 2021 were downloaded from CommonCrawl. Specifically, plain text files (without HTML) and metadata were downloaded, including information on encoding and language detected using CLD2[2]. The size of the resulting corpus after each stage is shown in Table 1.

Table 1: Characteristics of each stage of the collected corpus

	2018-51	2021-25
Total downloaded documents	3,086 million	2,394 million
Documents in Spanish	131 million	106 million
Documents with minimum length	131 million	106 million
Total paragraphs/sentences	2,424 million	2,157 million
Total paragraphs/sentences (deduplicated)	872 million	756 million
Total words	35,480 million	29,533 million
Compressed size (base 62)	102.1GB	86.6GB

[2] https://github.com/CLD2Owners/cld2

The documents with less than 40% of Spanish words according to the metadata were discarded. The reason we did not pick a higher number is because CLD2 sometimes tends to misclassify Spanish text as Portuguese or Galician. As CommonCrawl contains web documents, the resulting text contains text from all around the world. There is no easy way to determine the proportion of regional variations in the corpus, but it can be approximated using the TLDs, as summarized in Table 2.

Table 2: Proportion of the 25 most common Top-Level Domains in the Corpus

Top-Level Domain	2018-51	2021-25	Top-Level Domain	2018-51	2021-25
com	54.07%	49.61%	ec (Ecuador)	0.28%	0.39%
es (Spain)	16.33%	16.71%	edu	0.16%	0.48%
org	4.72%	5.29%	ve (Venezuela)	0.39%	0.19%
ar (Argentina)	4.20%	4.88%	tv	0.27%	0.28%
mx (Mexico)	3.83%	4.21%	de (Germany)	0.20%	0.26%
net	3.98%	3.44%	fr (France)	0.14%	0.30%
cl (Chile)	2.47%	2.98%	ru (Russia)	0.22%	0.17%
co (Colombia)	1.94%	2.49%	cu (Cuba)	0.22%	0.15%
pe (Perú)	0.94%	1.06%	do (Dominican Republic)	0.15%	0.23%
info	1.02%	0.66%	it (Italy)	0.16%	0.20%
eu (European Union)	0.62%	0.57%	cr (Costa Rica)	0.11%	0.24%
uy (Uruguay)	0.40%	0.58%	me (Montenegro)	0.14%	0.14%
cat (Catalan)	0.28%	0.47%			

After filtering the documents by language, the documents were transformed to UTF-8. To normalize the text, all the letters were converted to lower case and the non-Latin characters were removed using a regular expression.

Once the documents were normalized, the text was split in lines (corresponding to paragraphs or phrases) and duplicated lines were filtered out, using a bloom filter [25].

Finally, the corpus was compressed. The compression was done by replacing with word with its position or ranking in the Corpus of the Royal Spanish Academy [26]. For out-of-dictionary words, the word was copied verbatim, prefixed with the character `!`. The positions were encoded using base 62. As the Corpus of the Royal Spanish Academy is sorted by frequency, the most common words were represented with fewer digits, achieving compression to approximately half its original size. This scheme was chosen to allow building the word embeddings over the compressed text, without required decompressing or modifying the word2vec implementation.

3.2 Word Embeddings Training

BlazingText [20] was used for training the word embeddings from the collected corpus. It was executed in 8 instances `ml.c4.8xlarge` in AWS SageMaker Studio. Each of those instances has 36 cores, 60 GB of RAM and costs $1.909USD/hour. The corpus from 2021 took 15852 seconds to executing, while the corpus from 2018 took 18639 seconds. The total cost of training both models was $150.07 USD. The configuration used is shown in Table 3.

Table 3: BlazingText Configuration

batch size	12	buckets	10 000 000
early stopping	no	epochs	5
learning rate	0.5	max char	35
min char	1	min count	10
min epochs	2	mode	batch skipgram
negative samples	5	patience	4
sampling threshold	0.0001	subwords	false
vector dimension	300	window size	5
word ngram	2		

3.3 Alignment of word embeddings

Given that the generation of word embeddings has random aspects, it is not possible to guarantee that the same word will have vectors that can be compared across different models [7]. Therefore, it is necessary to

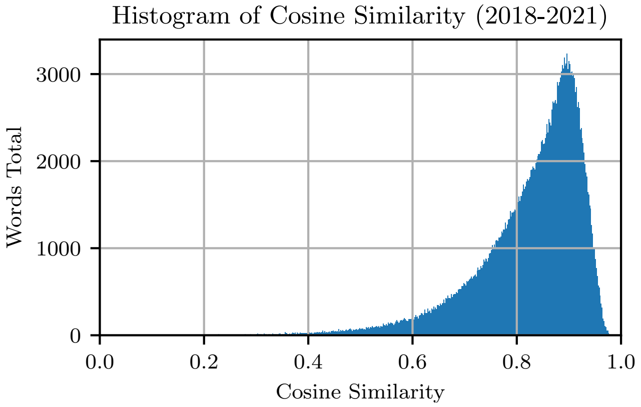


Figure 1: Histogram of the cosine similarity between words in the 2018 and 2021 word embeddings, after the alignment process, with 500 bins.

align the vectors.

There are various algorithms for aligning vectors, and the best one may depend on the task at hand. For example, [18] is used for aligning embeddings between different languages in an unsupervised manner. In this work, the same method as [7] was chosen, which involves treating the vectors as in the orthogonal Procrustes problem. The solution to this problem is to find the rotation matrix that minimizes the distance between each pair of vectors (i.e., the same word in both models for both years), which can be efficiently obtained using singular value decomposition [8].

In the Figure 1, a histogram is shown with the cosine similarities between the word embeddings of the words in the models trained with data from 2018 and 2021, after the alignment process. Here, the horizontal axis corresponds to the cosine similarity, where 1 means that the vector is identical and -1 means that it is completely opposite. The vertical axis corresponds to the number of words with such cosine similarity. The plot has a shape similar to a half bell with the maximum near 1, which coincides with the expectation that most of the words have a small semantic shift.

3.4 Determining topics with the greatest semantic shift

Automatic clustering was applied over the words with most semantic shift to determine the topics with most changes. With that algorithm, 575 clusters were obtained, that were filtered by hand, finding a total of 35 different topics.

First, from the 265505 words in common between both models, the words with a cosine similarity lower or equal to 0.7 and at least 1000 occurrences in the corpus were chosen. With that filtering, 24943 words were obtained.

Subsequently, these words were grouped using a variation of the DBSCAN algorithm [27]. Due to differences in density, the process began with an EPS[3] of 0.65, and the clustering algorithm was repeated for resulting clusters with more than 30 elements, decreasing the EPS by 10% each time. The process stopped when EPS < 0.1, there were fewer than 5 words remaining, or the process had been repeated 25 times. The resulting clusters were manually filtered, and a total of 35 topics were identified.

3.5 Analysis of selected topics

For each selected topic, an analysis was conducted to determine which words were closest to the cluster in both 2018 and 2021. Additionally, the words that moved closer or farther from the cluster were identified. A formula for relative displacement, shown in (2), was used to measure the degree of shifting. This formula utilizes a_t and b_t, which represent the word embeddings of the words being compared, and ϵ, a small value to prevent division by zero.

$$1 - \frac{\text{CosineSimilarity}\left(a_{t+\Delta}, b_{t+\Delta}\right) + \epsilon}{\text{CosineSimilarity}\left(a_t, b_t\right) + \epsilon} \tag{2}$$

[3] EPS, or epsilon, is the local radius used by the DBSCAN algorithm to look for possible neighbors for a cluster in expansion. If the distance between a point and a cluster is greater than EPS, then that point cannot be considered as part of that cluster. If the distance is lower, then the point can be added to the cluster, in which case the frontier of the cluster expands.

3.6 Analysis of the emotional shift

To measure emotional shift, the Shaver et al [23] hierarchy of emotions was translated into Spanish. For each tertiary emotion, groups of words were created that corresponded to the translated emotion and variations of the word (infinitive, gender change, gerund, participles, reflexive form, etc.). Then, Word Mover Distance was used to measure the proximity between each studied word and the words that represent each emotion, similar to the method proposed by [24]. One difference from [24] is that instead of determining the part of speech, a softmax function was applied to the distances, creating a probability distribution.

4 Semantic Shift Analysis

In this section, the semantic shift of the found topics is analysed. It is important to note that some words shown in the figures contains orthographic or encoding errors. Those errors, like "pandemía", "sosprechado" and "protecciãfæ" come from the corpus used to generate the word embeddings, and were not introduced while redacting this document.

The clusters analyzed in this article are summarized in the Table 4. The table includes the name of the cluster, the word that had the biggest semantic shift and its cosine similarity between 2018 and 2021.

Table 4: Clusters analyzed in this article, the word with the biggest semantic shift and its cosine similarity between 2018 and 2021

Cluster Name	Biggest Shift	Cosine Similarity
Video games	stadia	0.1689
Cryptocurrencies	aave	0.2127
COVID	covid	0.2261
COVID vaccines	covax	0.2379
Cycling	arkea	0.2980
Soccer players	sergiño	0.3032
Pornography	blizz	0.3136
Space Missions	tianwen	0.3261
Basketball	herro	0.3376
COVID testing	hisopar	0.3795
Tennis players	musetti	0.4116
Masks	nasobuco	0.4409
Korean pop	itzy	0.4466
Sanitizers	hidroalcohólico	0.4989
Vaccination	inoculará	0.6125

Additionally, the Table 5 contains groups that were identified but not studied in this work. Other groupings identified include various clusters of cities from around the world and collections of products in online stores, among others.

Table 5: Clusters identified but not analyzed in this article, the word with the highest semantic shift and its cosine similarity between 2018 and 2021

Cluster	Biggest Shift	Similarity	Cluster	Biggest Shift	Similarity
Painters	vrel	0.2203	Pirated Films	chilie	0.3873
Pills	veona	0.2219	Dating Apps	eharmony	0.4223
Electronics	narzo	0.2313	Freelancers	workana	0.4596
Gossip News	måneskin	0.2316	Women's names	lania	0.4734
Star Wars & Marvel	grogu	0.2933	Anime & Manga	nezuko	0.4994
Korean Actors	jihu	0.3514	Tiktok	tiktokers	0.5056
US Officials	walensky	0.3574	NSFW Content	onlyfans	0.5146
Online Bets	betwinner	0.3799	Shoes	ballerinas	0.5299

4.1 Video games

The cluster "video games", that is shown in Figure 2, includes words related with video games and consoles. The cluster centroid is the word "*godfall*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.4995.

“Video Games” Cluster

valheim, csgo, kakarot, squadrons, outriders, mediatonic, runeterra, teamfight, valorant, maneater, intergrade, carmageddon, dualsense, fenyx, genshin, replicant, warzone, godfall, resurrected, shadowlands, returnal, yuffie, wonderworld, onlygames and stadia

Neighbors 2018

confinament, digitalització, copyable, pandèmia, restriccions, sentilecto, unaccepted, elevades, youyour, delegacions, obteniu, mantindre, recalculating, descreet, queixes, depèn, collumn, recommenders, realted, sizereduction, swers, mutability, disjunctive, hytale and cubetype

Neighbors 2021

xenoverse, xcloud, xbox, housemarque, fighterz, nioh, playstation, titanfall, nier, battlefront, cataclysm, guacamelee, psyonix, biomutant, hellblade, torchlight, gamepass, darksiders, spelunky, fortnite, dualshock, crysis, pubg, juegosdb and battlefield

Top Shifted Farther

cubetype, sizereduction, abjuration, restricciónsubject, licitant, collumn, realted, toyour, depèn, queixes, recalculating, swers, mantindre, delegacions, descreet, unaccepted, elevades, obteniu, youyour, restriccions, pandèmia, digitalització, copyable, sentilecto and confinament

Top Shifted Closer

xcloud, xbox, playstation, overwatch, pubg, psvr, xenoverse, titanfall, legends, driveclub, nioh, gamepass, fortnite, soulcalibur, ouya, wiiu, smite, crossplay, dualshock, hellblade, hearthstone, retrocompatibilidad, housemarque, guacamelee and xbla

Figure 2: Analysis of neighbors, shifts closer and farther of the cluster about “Video Games”

In this cluster, the semantic shift was mainly due to the release of new video games or consoles after the collection of the 2018 corpus, such as Valorant (announced in October 2019 and released in June 2020) and Balan Wonderworld (launched in March 2021), among others. Some video games and consoles stand out in this cluster that were not yet on the market in week 51 of 2018 when the data was collected to calculate the first word embeddings used as a base, such as Valorant (a game announced in October 2019 and released in June 2020), Balan WonderWorld (a game released in March 2021), Maneater (a game released in May 2020), DualSense (a controller for PlayStation 5 announced in April 2020), Immortals Fenyx Rising (a game released in December 2020), Stadia (a streaming game platform launched in November 2019), Valheim (a game released in February 2021), Teamfight Tactics (a game released in June 2019), Star Wars: Squadrons (a game announced in June 2020), World of Warcraft: Shadowlands (a game announced in November 2019 and released in November 2020), and others. However, there are also some older video game terms in this cluster, such as Carmageddon (released in 1997) and CS:GO (an acronym for Counter-Strike: Global Offensive, released in 2012).

Since this cluster is mostly composed of new video games, it is interesting to see that some of the terms that came closest are related to video games, such as consoles (Xbox, PlayStation, WiiU, PS-Vita, Ouya) or companies in the sector (Nintendo, Ubisoft). Here we can see that more “creative” terms are more easily shifted by the release of a product. For example, products with a more generic name such as Nintendo Switch or Amazon Luna did not move those terms closer to this cluster.

4.2 Cryptocurrencies

The cluster “cryptocurrencies”, that is shown in Figure 3, includes words related with cryptocurrencies and decentralized finance (DEFI). The cluster centroid is the word “*wbtc*”. In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.4956.

“Cryptocurrencies” Cluster

yfi, chz, blockfi, doge, polkadot, robinhood, etherium, chainlink, wbtc, chiliz, ltc, shib, staking, algorand, busd, beeple, snx, nft, aave, taproot, defi, enjin, yearn, paxful, etherum and nfts

Neighbors 2018

oqi, uzl, vjj, owj, mqj, yny, gxo, mwj, lwy, nqi, jrq, jgq, yzh, oqr, uqn, iqoo, pxoa, uew, uuw, jqv, ethereum, jwv, litecoin, jtw and changenow

Neighbors 2021

usdt, qtum, usdc, tezos, ethereum, dogecoin, localbitcoins, cardano, makerdao, mainnet, bitgo, filecoin, vechain, dapps, kucoin, tokens, stablecoins, stablecoin, altcoin, litecoin, coingecko, changelly, zcash, blockchains and decentraland

Top Shifted Farther

dxu, jwv, jgq, zny, ljq, rqe, lwu, rzw, uuw, pxoa, lsz, jrq, uqn, iyu, nqi, oqr, mwj, lwy, gxo, yny, mqj, owj, uzl, vjj and oqi

Top Shifted Closer

qtum, dapps, ethereum, tezos, usdt, filecoin, cardano, altcoins, kucoin, altcoin, stablecoins, okex, blockchains, binance, litecoin, cripto, usdc, dogecoin, stablecoin, bitcoin, zcash, xrp, poloniex, bittrex and makerdao

Figure 3: Analysis of neighbors, shifts closer and farther of the cluster about “Cryptocurrencies”

Similar to the video game cluster, many of the terms in the cryptocurrency cluster correspond to projects launched after the collection of data for the first word embedding used, such as Yearn Finance (July 2020), shib (the symbol for Shiba Inu, launched in August 2020), busd (launched in September 2019), and Chainlink (launched in May 2019).

Additionally, some pre-existing cryptocurrencies, such as Doge and Ethereum (misspelled as etherium and etherum), as well as the application Robin Hood for buying stocks and cryptocurrencies, all launched in 2015, are also present in this cluster. Notably, many of these cryptocurrencies correspond to decentralized finance or DeFi systems. Since Ethereum is a platform that supports programmable contracts, many of these DeFi applications were developed on Ethereum.

Other cryptocurrencies and platforms for trading them are among the words that came closest to the cluster.

4.3 COVID-19

The cluster "covid-19", that is shown in Figure 4, includes words related with COVID-19 and related topics like quarantine. The cluster centroid is the word "*pandemia*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.4802.

"COVID-19" Cluster

covid, pandémico, pandemiaâ, ncov, cov, covi, contagiados, pandémicos, covit, sars, autoaislamiento, confinamientos, pandémicas, desconfinamiento, confinamiento, cuarentena, cuarentenas, pandémica, covd, coronavirus, pandemia and pandemía

Neighbors 2018

epidemia, contagiadas, sosprechado, iclr, pospandémico, infodemia, optm, epidémicas, itors, gripecita, vacunaría, autoconfinamiento, wpac, grupoesoc, swers, protecciãfæ, sanibrun, thinkbook, puntosfuertes, mubs, epidémicos, adenovirus, ftui, stalkerware and epidémico

Neighbors 2021

contagiadas, desescalada, contagios, infectados, crisisâ, fallecidos, virus, contagiado, influenza, fallecimientos, epidemia, confirmados, hospitalizados, actualâ, patógeno, crisis, norovirus, recesiã, rebrotes, contagiaron, pandã, micaâ, contagio, hospitalizadas and asintomáticos

Top Shifted Farther

ufsa, sanibrun, generalmills, maxdisplay, amdpress, nnum, mubs, dezombies, tnkr, simblicamente, fundamentall, bpage, gozazaragoza, websties, jsst, wpac, thinkbook, agroconcept, binor, puntosfuertes, swers, itors, optm, sosprechado and iclr

Top Shifted Closer

desescalada, virus, contagios, aspo, influenza, fallecidos, crisisâ, contagio, rebrotes, wuhan, contagiadas, patógeno, fallecimientos, convid, antiviral, ébola, epidemia, hantavirus, decesos, brote, distanciamiento, dengue, sanitaria, norovirus and contagiarse

Figure 4: Analysis of neighbors, shifts closer and farther of the cluster about "COVID-19"

Despite COVID-19 being a completely unknown disease until late 2019, many terms in this cluster were already in common use. As a result, some common terms shifted in meaning to become more closely associated with the COVID-19 pandemic. One exception is the term "COVID", which was not widely used before the outbreak and was likely mostly encountered as a typographical error.

4.3.1 Term covid

In the case of "covid", this term was previously associated with terms that appeared to be orthographic or encoding errors, such as *amdpress* or *gozazaragoza*. However, in 2021, it is associated with terms used to refer to COVID-19 (including typos), such as *covit*, *convid*, and *covd*. It is also associated with *contagio* (contagion), *contagios* (contagions), *virus* (virus), *pandemia* (pandemic), *coronavirus* (coronavirus), *influenza* (influenza), and *contagiados* (infected individuals). These changes are shown in Figure 5.

Term "*covid*"

Neighbors 2018

grupoesoc, sanibrun, websties, gozazaragoza, amdpress, maxdisplay, bcty, tádel, gascaribe, veritrade, idnivel, myftpupload, kitempleo, datarush, grupogodo, facturacionweb, waterpolosevilla, cookielawinfo, galussothemes, quibi, hceu, educasites, sercaman, letradox and metasiteid

Neighbors 2021

coronavirus, covd, cov, sars, pandemia, covit, contagios, covi, ncov, contagiados, virus, influenza, contagio, pandemía, convid, contagiadas, pandémica, epidemia, dengue, cuarentena, sanitaria, brote, ébola, pcr and pandémico

Top Shifted Farther

galussothemes, domainid, weburl, carbosin, gascaribe, gallerytype, cookielawinfo, educasites, agroconcept, grupoesoc, uclasificados, myftpupload, metasiteid, facturacionweb, bcty, grupogodo, kitempleo, veritrade, tádel, idnivel, sanibrun, maxdisplay, amdpress, gozazaragoza and websties

Top Shifted Closer

coronavirus, sars, pandemia, cov, contagios, contagiados, covd, contagio, covi, contagiadas, influenza, pandemía, pandémica, ébola, virus, epidemia, brote, cuarentena, dengue, contagiarse, rebrote, confinamiento, covit, rebrotes and pandémico

Figure 5: Analysis of neighbors, approaches and distances of the term "covid"

4.3.2 Term coronavirus

As seen in Figure 6, the term "coronavirus" was previously associated with other types of viruses and bacteria, such as adenovirus, enterovirus, bordetella, and bugdorferi. However, these associations shifted abruptly over time. In fact, some of the words that were previously close neighbors in 2018 moved the farthest in 2021.

Term "*coronavirus*"

Neighbors 2018

adenovirus, parainfluenza, herpesvirus, parvovirus, bordetella, enterovirus, sincitial, borrelia, rinovirus, flavivirus, rotavirus, micoplasma, mycoplasma, arbovirus, burgdorferi, babesia, leptospira, typhi, distemper, yersinia, trachomatis, rickettsia, aviaria, papilomavirus and sincicial

Top Shifted Farther

bartonella, giardiasis, babesiosis, strongyloides, chlamydia, enterovirus, ancylostoma, sincitial, leptospira, typhi, trachomatis, rinovirus, parvovirus, pasteurella, burgdorferi, flavivirus, babesia, mycoplasma, micoplasma, papilomavirus, bordetella, adenovirus, borrelia, herpesvirus and parainfluenza

Neighbors 2021

covid, sars, covd, cov, pandemia, contagios, covit, virus, contagiados, contagio, ncov, covi, brote, epidemia, contagiadas, pandemía, cuarentena, convid, influenza, hantavirus, ébola, rebrote, pandémica, dengue and confinamiento

Top Shifted Closer

covid, covd, covit, cov, covi, pandemia, convid, confinamiento, confinamientos, covis, sars, cuarentena, reaperturas, rebrotes, confinarnos, wuhan, sanitaria, rebrote, desconfinamiento, manaos, aforos, crisis, entornointeligente, dgratis and flexibilizaciones

Figure 6: Analysis of neighbors, approaches and distances of the term "coronavirus"

4.3.3 Term cuarentena *(quarantine)*

The term "cuarentena" (quarantine) stopped being close to terms like *zoosanitario* (zoosanitary) or *salmonela* (salmonella), and got closer to terms related to the COVID-19 outbreak, like *coronavirus*, *pandemia* (pandemic) or *covid*. It also shifted closer to terms related to the new lifestyle, like *teletrabajo* (teleworking), *presencialidad* (in-person attendance) or *autoaislamiento* (self-isolation). Such changes can be seen in Figure 7.

Term "*cuarentena*"

Neighbors 2018

cuarentenas, quarantine, infectados, cuarentenarias, infestados, infectado, infectada, salmonela, aphis, contagiados, hibernación, confinamiento, infestadas, zoosanitario, zoosanitaria, infestación, alerta, plaga, aviar, infestado, sospechosos, contagios, fumigado, aviaria and ébola

Top Shifted Farther

dramapaís, dqsa, apiario, ebtools, mamferos, recycler, permacultivo, hcsp, huevecillos, fenzoni, kyth, eslegal, rutinaria, zddr, ejegir, infestados, inspeccionar, peligro, restricciόntema, accecer, productero, turkishexporter, deequipaje, infestado and estriles

Neighbors 2021

confinamiento, cuarentenas, autoaislamiento, pandemia, aspo, desconfinamiento, coronavirus, confinamientos, pandemía, distanciamiento, covid, asilamiento, desescalada, covd, encierro, covit, contagios, confinarse, aislamiento, contagiados, contagio, confinarnos, contigencia, flexibilizaciones and rebrote

Top Shifted Closer

confinamiento, aspo, autoaislamiento, covid, distanciamiento, pandemia, desconfinamiento, teletrabajo, covd, covit, confinamientos, presencialidad, coronavirus, covi, asilamiento, desescalada, flexibilizaciones, teletrabajar, bimodalidad, cuidándonos, flexibilización, reaperturas, aislamiento, semipresencialidad and presenciales

Figure 7: Analysis of neighbors, approaches and distances of the term "cuarentena"

4.3.4 Term pandemia *(pandemic)*

Previously, the term *pandemia* (pandemic) was close to others like *epidemia* (epidemic), *aviaria* (avian), *influenza* (influenza) or *gripe* (flu). Now, such term is closer to terms like *coronavirus* (coronavirus), *covid* (covid), *confinamiento* (confinement) and *cuarentena* (quarantine). Within the terms that shifted farther from *pandemia* there is *bioterrorismo* (bioterrorism), *antivacuna* (anti-vaccine), *gripe* (flu), *aviar* (avian), *aviaria* (avian), *ébola* (ebola), *carbunco* (anthrax), *tiroidea* (thyroid), *poliomielitis* (poliomyelitis), and *multirresistente* (multidrug-resistant). The changes are shown in Figure 8.

4.4 COVID Testing

The cluster "COVID-19 testing", that is shown in Figure 9, includes words related with COVID-19 detection tests. The cluster centroid is the word "*hisopados*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.5438.

Term "*pandemia*"

Neighbors 2018

epidemia, pandemias, pandémico, pandémica, gripe, aviaria, epidémico, epidémica, pandémicas, endemia, epidemias, epidémicos, aviar, influenza, pandemía, epidémicas, ébola, infodemia, hambruna, pandémicos, sars, peste, plaga, poliomielitis and bubónica

Neighbors 2021

pandemía, covid, coronavirus, covd, epidemia, pandémica, pandémico, crisis, confinamiento, pandemiaâ, cuarentena, confinamientos, contingencia, covit, pandémicos, contigencia, sanitaria, cuarentenas, pandémicas, contagios, crísis, sars, coyuntura, pandemias and covi

Top Shifted Farther

sida, esquistosomiasis, epidémicos, leishmaniasis, meningitis, parotiditis, viruela, ántrax, prevenible, papilomavirus, multirresistente, endemia, tifoidea, epidémicas, ebola, pandemias, plaga, antivacuna, bioterrorismo, poliomielitis, carbunco, gripe, aviar, aviaria and epidemia

Top Shifted Closer

covid, covd, confinamiento, confinamientos, covit, coronavirus, covi, pandemía, cuarentena, sanitaria, aspo, convid, cov, cuarentenas, reaperturas, pandemiaâ, confinarnos, crisis, covidien, contingencia, pande, coyuntura, reapertura, teletrabajo and virtualidad

Figure 8: Analysis of neighbors, approaches and distances of the term "pandemia"

"COVID-19 Testing" Cluster

testearse, hisopado, pcr, pcrs, antígenos, hisopar, hisopados, pdia and testeos

Neighbors 2018

antígeno, anticuerpos, antigénicos, macrófagos, inmunoglobulinas, linfocitos, autoanticuerpos, testeo, antigénica, anticuerpo, antigénicas, mastocitos, monoclonales, hemocultivo, glicoproteínas, urocultivo, nasobuco, macrófago, qpcr, laboratoriales, monocitos, unicamanera, citocinas, glicoproteína and linfocito

Neighbors 2021

antígeno, testeo, serológicos, antigénicos, serológicas, antigénicas, anticuerpos, serológica, serológico, antigénica, serología, qpcr, malbrán, nasofaríngeo, vacunatorios, rastrillajes, igg, inocularse, antigeno, agendarse, vacunarse, inmunizarse, nasofaríngea, cov and sars

Top Shifted Farther

buscadon, thinkbook, nnni, pulsimetro, stalkerware, protecciãfæ, nasobuco, zddr, faciltarte, howtopronounce, jicotera, ejegir, infaltante, hidroalcohol, alcontacto, dezombies, ebtools, eslegal, deequipaje, simblicamente, perduraci, emuaidmax, enalimentos, sosprechado and unicamanera

Top Shifted Closer

serológica, vacunatorios, martorano, contagiados, serológicas, puratich, covid, serolã, contagiadas, serológico, vacunarse, testeo, aspo, asintomáticos, serológicos, chahla, confirmados, hbsag, contagios, seroconversión, serología, malbrán, hemocultivos, cribados and nasofaríngea

Figure 9: Analysis of neighbors, shifts closer and farther of the cluster about "COVID-19 Testing"

This cluster is characterized by terms such as *nasobuco*, *desconfinamiento* and *hidroalcohol*, which have also clustered together in other groups. Among the new neighboring words that have moved closer to this cluster are the names of politicians and institutions in Argentina, including Puratich (Minister of Health of Chubut, Argentina), SIPROSA (Provincial Health System of Tucumán, Argentina), Malbrán (National Administration of Laboratories and Health Institutes "Dr. Carlos Malbrán", an entity dependent on the Ministry of Health of Argentina), Chahla (Rossana Chahla, a medical doctor and deputy for Tucumán, Argentina), Santojanni (a hospital in Buenos Aires, Argentina), Masvernat (Hospital Delicia Concepción Masvernat, located in Entre Ríos, Argentina), Zgaib (Minister of Health of Río Negro, Argentina), Martorano (Minister of Health of Santa Fe, Argentina), and Perrando (a hospital in Chaco, Argentina). The strong influence of texts related to Argentina is also evident with terms such as ASPO (social, preventive, and mandatory isolation) and AMBA (metropolitan area of Buenos Aires).

4.5 COVID-19 Vaccines

The cluster "COVID-19 vaccines", that is shown in Figure 10, includes words related with COVID-19 specific vaccines and entities that developed them. The cluster centroid is the word "*astrazeneca*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.4953.

In the cluster related to COVID-19 vaccines, we find the names of vaccines such as Sputnik or Curevac, as well as the names of companies and institutions that manufacture them, including Sinovac, Pfizer, BioNTech, AstraZeneca, Janssen, Gamaleya, and Sinopharm. The COVAX vaccine distribution mechanism is also present in this cluster.

Some people or organizations that worked in the vaccination effort moved closer to this cluster. For example, Vizzotti (Health minister of Argentina) and Tedros (World Health Organization director), gavi (*the Vaccine Alliance*, co-leader of the COVAX effort), RDIF (Russian Direct Investment Fund, participated in selecting and financing COVID-19 testing systems, medications, and vaccines, including Sputnik V) and BIRMEX (a Mexican state-owned company that produces and sells vaccines and antivenoms).

“COVID-19 Vaccines” Cluster

sinovac, covax, inmunizante, inmunizantes, biontech, aztrazeneca, zeneca, janssen, novavax, biologics, astrazeneca, curevac, pzifer, sinopharm, pfizer, astrazéneca, gamaleya and sputnik

Neighbors 2018

novartis, glaxosmithkline, sanofi, dqsa, merck, ovhcloud, agroconcept, recommenders, marketresearch, booknet, swers, oekb, aventis, inmoinfo, remdesivir, datarush, jsst, exchnage, promotheus, genzyme, onespan, generalmills, sentilecto, gsk and themeinwp

Neighbors 2021

vacuna, vacunas, rdif, inoculantes, regeneron, inoculaciones, inoculante, inmunización, sanofi, inmunizar, arnm, bexsero, dosis, butantan, cilag, dmitriev, biofarmacéutica, cansino, aplidin, inmunizadas, fármaco, vacunadas, liomont, remdesivir and birmex

Top Shifted Farther

futboldelujo, hostadvice, dvam, fundamentall, sentilecto, wpac, customify, themeinwp, trangender, inmoinfo, feedrss, oekb, carbosin, copid, datarush, jsst, booknet, generalmills, booktitle, marketresearch, recommenders, swers, promotheus, dqsa and agroconcept

Top Shifted Closer

vacuna, vacunas, rdif, cansino, arnm, inoculaciones, inmunizar, dosis, vacunadas, inmunizaciones, inmunización, bourla, mrna, inmunizados, inoculación, vacunados, inmunizadas, vacunar, inoculados, remdesivir, dmitriev, bancel, inoculante, inmunizarse and vacunación

Figure 10: Analysis of neighbors, shifts closer and farther of the cluster about “COVID-19 Vaccines”

4.6 Masks

The cluster “masks”, that is shown in Figure 11, includes words related with masks and face protection. The cluster centroid is the word “*cubrebocas*”. In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.5654.

“Masks” Cluster

nasobucos, mascarilla, cubreboca, cubrebocas, ffp, tapaboca, barbijo, máscarillas, mascarillas, distanciamiento and nasobuco

Neighbors 2018

máscarilla, hidratante, tapabocas, peeloff, pantalonespantalones, zddr, protecciãfæ, sosprechado, sérum, exfoliante, daytox, cuticate, zddd, deequipaje, enalimentos, purificante, infomodule, exfoliantes, hisopar, dqsa, buscadon, emuaidmax, barbijos, loción and facialderm

Neighbors 2021

tapabocas, barbijos, máscarilla, autofiltrantes, autofiltrante, máscaras, máscara, antibacterial, distanciamientos, fpp, camisolines, autoaislamiento, hidroalcohólicos, earloop, iir, sanitizantes, sanitizante, higiénicas, alcoholado, contagio, hidroalcohólicas, aglomeraciones, respiradores, batas and cuarentena

Top Shifted Farther

pajaadictos, yardwe, puntosfuertes, superlarge, stalkerware, multianuncios, ebtools, plkd, reluzcas, chicklets, howtopronounce, dezombies, thinkbook, pantalonespantalones, dqsa, enalimentos, meojores, faciltarte, zddd, buscadon, sosprechado, cuticate, protecciãfæ, zddr and infomodule

Top Shifted Closer

tapabocas, bioseguridad, aforos, barbijos, fpp, iir, cuarentena, aglomeraciones, aforo, cuarentenas, alcoholado, hidroalcohólicos, máscaras, sanitización, sanitizantes, aspo, antibacterial, sanitarias, hidroalcohólico, quirúrgicas, cuidándonos, higienizarse, sanitizados, gatell and higiénicas

Figure 11: Analysis of neighbors, shifts closer and farther of the cluster about “Masks”

In the masks cluster, there are several synonyms for this tool for protecting both the mouth and nose. It is interesting to note the presence of the term *“nasobuco”*, which is a term for a mask that is primarily used in Cuba[4]. This demonstrates how word embeddings can capture words with analogous uses in different regions of the world. An additional interesting aspect is that the cluster moved farther away from facial masks or cosmetic treatments. Terms like *hidratante*, *peeloff*, *sérum*, *exfoliante*, *daytox*, *purificante*, *loción*, and *facialderm*, which were among the closest neighbors in 2018, shifted farther away in 2021.

4.7 Sanitizers

The cluster “sanitizers”, that is shown in Figure 12, includes words related with cleaning products. The cluster centroid is the word “*hidroalcohólico*”. In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.5903.

In the case of sanitizers, we can see that the focus has been on *hidroalcohol* (alcohol gel) and *vericidas* (virucides), that is, disinfectants designed against viruses. Among the words that came closest, we can see other words related to COVID-19 prevention, such as *distanciamiento* (distancing), *dispensadores* (dispensers), *tapabocas* (face masks), or *aforos* (capacity limits).

4.8 Vaccination

The cluster “vaccination”, that is shown in Figure 13, includes words related with vaccination. The cluster centroid is the word “*inmunizados*”. In average, the cosine similarity of the words in the cluster, relative to

[4] According to `https://www.rae.es/observatorio-de-palabras/nasobuco`

"Sanitizers" Cluster

hidroalcohólicos, virucidas, hidroalcohol, sanitizador, sanitizantes, alcoholado, hidroalcohólico, sanitizante, hidroalcoholico, hidroalcohã and hidroalcohólicas

Neighbors 2018

viricida, glicerinado, hidroalcohólica, hidroxietilcelulosa, desinfectante, desinfectantes, higienizante, sanitiza, antimicã, etanólico, germicidas, antianã, bacteriostáticos, antiasmã, virucida, mucolã, desincrustantes, bactericidas, carbómero, bakuchiol, dermoprotector, higienizantes, emuaidmax, desodorizante and antirreumã

Neighbors 2021

higienizante, hidroalcohólica, antibacterial, desinfectante, viricida, virucida, higienizantes, desinfectantes, sanitización, glicerinado, antiseptico, sanitizar, biocidas, bactericidas, sanitizacion, antibacteriales, germicidas, biocida, desinfecciã, antisépticos, geles, hipoclorito, jabã, lavandina and higienizador

Top Shifted Farther

infomodule, youghourt, coptis, enalimentos, vevor, disodio, alcontacto, dietilo, oogaa, ebtools, etilhexilglicerina, phytorelax, mucolã, hisopar, grnulo, bakuchiol, emuaidmax, dwmx, beacuir, antiasmã, dioctilo, antianã, hidroxietilcelulosa, etanólico and antimicã

Top Shifted Closer

antibacterial, distanciamiento, epis, glicerinado, sanitización, tapabocas, barbijos, higienización, hidroalcohólica, cubrebocas, desinfección, higienizante, anticontagio, barbijo, dispensadores, antibactérien, desfibriladores, sanitizar, pediluvios, mascarillas, higiã, higienizarse, aforo, desinfectarán and cubreboca

Figure 12: Analysis of neighbors, shifts closer and farther of the cluster about "Sanitizers"

itself in 2018 and 2021 is 0.6649.

"Vaccination" Cluster

inoculación, inmunizó, inmunizados, inoculados, inoculaciones, inocularse, vacunados, inmunizando, inmunizará, inoculará, vacunarán, vacunará, inmunizado, inmunizada, vacunó, vacunada and inoculadas

Neighbors 2018

vacunadas, vacunado, vacunaron, desparasitada, inmunizaron, vacunándose, desparasitados, vacunan, inocularon, inoculada, inmunizadas, inoculado, desparasitado, inmunizarse, vacunando, inoculaciã, vacunar, vacunaran, vacunen, vacunaría, esterilizada, vacunación, vacunemos, pandemía and inocular

Neighbors 2021

inmunizarse, vacunarse, vacunando, vacunaron, vacunadas, inmunizadas, vacunado, inmunizaron, inmunización, vacunación, inoculado, vacunan, inoculada, inocular, inmunizar, inoculando, inocularon, vacunar, vacunas, vacunaciones, inoculó, inoculaciã, vacunara, vacunaría and inmunizaciones

Top Shifted Farther

infodemia, cuticate, dismuyen, youghourt, zoonóticas, pandémico, dqsa, variedadesde, sosprechado, tuberculosas, sonicación, lobosnews, nezuko, esterilizada, sindemia, pandémicas, vacunándose, hbtes, lisados, pospandémico, pandémicos, desparasitado, desparasitados, pandemía and desparasitada

Top Shifted Closer

vacunarse, inmunizarse, vacunado, sinopharm, vacunación, vacunando, inmunización, astrazeneca, biontech, inmunizadas, astrazéneca, covax, pfizer, cansino, sputnik, aztrazeneca, zeneca, inoculada, vacunadas, inmunizar, vacunaron, inoculado, covid, vacunar and alabí

Figure 13: Analysis of neighbors, shifts closer and farther of the cluster about "Vaccination"

4.9 Sports

In the sports-related clusters, we found groups associated with cycling, soccer, basketball, and tennis. Generally, the changes in these clusters were due to the inclusion of the names of young players. When mentioning the age of a player in this section, it refers to their age as of the second corpus collection, which was conducted in the 25th week of 2021.

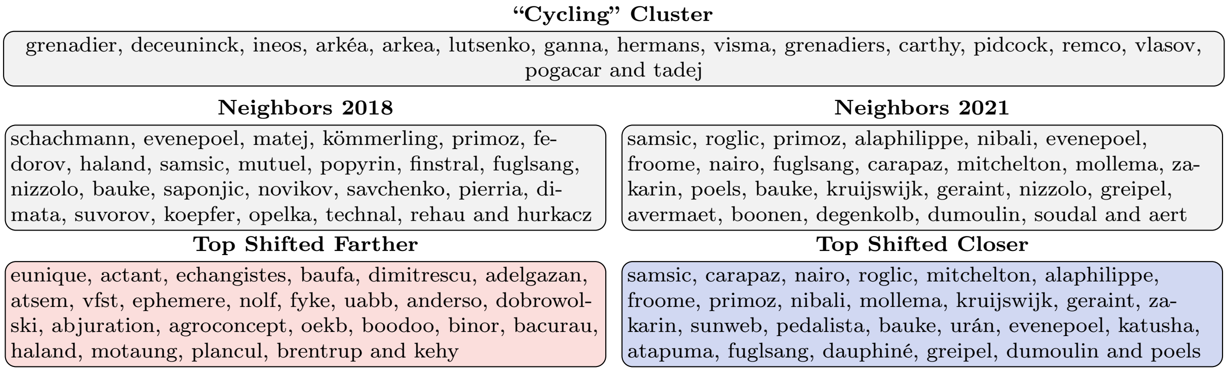
4.9.1 Cycling

The cluster "cycling", that is shown in Figure 14, includes words related with cycling teams and cyclists. The cluster centroid is the word "*grenadiers*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.5799.

In this cluster, we have cycling teams such as Arkéa, Visma (Jumbo Visma), INEOS Grenadiers, Deceuninck (Deceuninck-Quick Step) as well as cyclists like Remco (21 years), Carthy (27 years), Pidcock (22 years), Pogacar (23 years), Hermans (35 years), Lutsenko (29 years), Ganna (25 years) and Vlasov (25 years). Age appears to play a role in the semantic displacement of the cluster, similar to other domains like video games or cryptocurrencies. However, surnames may not necessarily refer to the intended cyclist; for example, "Lutsenko" may also refer to a Ukrainian politician.

4.9.2 Soccer

The cluster "soccer", that is shown in Figure 15, includes words related with soccer players. The cluster centroid is the word "*pedri*". In average, the cosine similarity of the words in the cluster, relative to itself in



“Cycling” Cluster

grenadier, deceuninck, ineos, arkéa, arkea, lutsenko, ganna, hermans, visma, grenadiers, carthy, pidcock, remco, vlasov, pogacar and tadej

Neighbors 2018

schachmann, evenepoel, matej, kömmerling, primoz, fedorov, haland, samsic, mutuel, popyrin, finstral, fuglsang, nizzolo, bauke, saponjic, novikov, savchenko, pierria, dimata, suvorov, koepfer, opelka, technal, rehau and hurkacz

Neighbors 2021

samsic, roglic, primoz, alaphilippe, nibali, evenepoel, froome, nairo, fuglsang, carapaz, mitchelton, mollema, zakarin, poels, bauke, kruijswijk, geraint, nizzolo, greipel, avermaet, boonen, degenkolb, dumoulin, soudal and aert

Top Shifted Farther

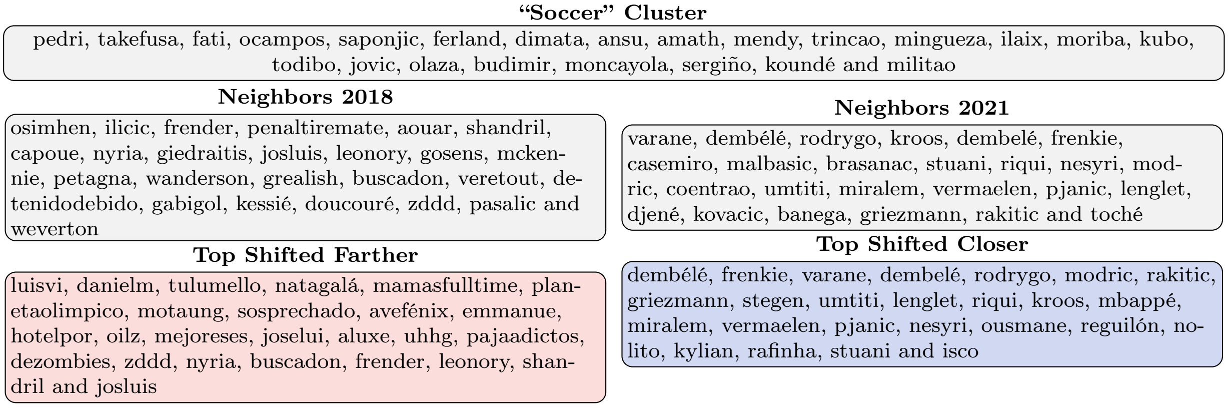
eunique, actant, echangistes, baufa, dimitrescu, adelgazan, atsem, vfst, ephemere, nolf, fyke, uabb, anderso, dobrowolski, abjuration, agroconcept, oekb, boodoo, binor, bacurau, haland, motaung, plancul, brentrup and kehy

Top Shifted Closer

samsic, carapaz, nairo, roglic, mitchelton, alaphilippe, froome, primoz, nibali, mollema, kruijswijk, geraint, zakarin, sunweb, pedalista, bauke, urán, evenepoel, katusha, atapuma, fuglsang, dauphiné, greipel, dumoulin and poels

Figure 14: Analysis of neighbors, shifts closer and farther of the cluster about “Cycling”

2018 and 2021 is 0.5517.



“Soccer” Cluster

pedri, takefusa, fati, ocampos, saponjic, ferland, dimata, ansu, amath, mendy, trincao, mingueza, ilaix, moriba, kubo, todibo, jovic, olaza, budimir, moncayola, sergiño, koundé and militao

Neighbors 2018

osimhen, ilicic, frender, penaltiremate, aouar, shandril, capoue, nyria, giedraitis, josluis, leonory, gosens, mckennie, petagna, wanderson, grealish, buscadon, veretout, detenidodebido, gabigol, kessié, doucouré, zddd, pasalic and weverton

Neighbors 2021

varane, dembélé, rodrygo, kroos, dembelé, frenkie, casemiro, malbasic, brasanac, stuani, riqui, nesyri, modric, coentrao, umtiti, miralem, vermaelen, pjanic, lenglet, djené, kovacic, banega, griezmann, rakitic and toché

Top Shifted Farther

luisvi, danielm, tulumello, natagalá, mamasfulltime, planetaolimpico, motaung, sosprechado, avefénix, emmanue, hotelpor, oilz, mejoreses, joselui, aluxe, uhhg, pajaadictos, dezombies, zddd, nyria, buscadon, frender, leonory, shandril and josluis

Top Shifted Closer

dembélé, frenkie, varane, dembelé, rodrygo, modric, rakitic, griezmann, stegen, umtiti, lenglet, riqui, kroos, mbappé, miralem, vermaelen, pjanic, nesyri, ousmane, reguilón, nolito, kylian, rafinha, stuani and isco

Figure 15: Analysis of neighbors, shifts closer and farther of the cluster about “Soccer”

In this cluster, there are soccer players such as Ocampos (27 years), Ilaix Moriba (18 years), Koundé (23 years), Trincão (22 years), Takefusa Kubo (20 years), Saponjic (24 years), Dimata (24 years), Pedri (19 years), Ansu Fati (19 years), Jović (24 years), Militao (23 years), Ferland Mendy (26 years), Olaza (27 years), Amath (Ndiaye, 25 years), Todibo (22 years), Budimir (30 years), Sergiño (21 years), Moncayola (23 years), and Mingueza (22 years). Similar to the case with cyclists, this group is mostly composed of young players. From a semantic shift perspective, it is reasonable to see further shifts in the names of young players since they are starting their career, and their names are being associated with new events that may trigger semantic shifts.

4.9.3 Basketball

The cluster “basketball”, that is shown in Figure 16, includes words related with basketball players. The cluster centroid is the word “*herro*”. In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.6204.

In the basketball cluster, one can find players such as Avramović (27 years), Adebayo (24 years), Caris (27 years), Pippen (56 years), Bolmaro (21 years), LaMelo (20 years), Polonara (30 years), Hardaway (by Tim Hardaway Jr. 29 years), Herro (21 years), LeVert (27 years), DeMar (32 years), Kobe (1978-2020), and LeBron (37 years). It is worth noting that the NBA has a minimum age requirement of 19 years for players. As observed, the majority of players in this cluster are young players. Additionally, the cluster includes the team Basket Zaragoza (known as Casademont Zaragoza) and a basketball blog.

4.9.4 Tennis

The cluster “tennis”, that is shown in Figure 17, includes words related with tennis players. The cluster centroid is the word “*popyrin*”. In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.6198.

In the tennis player cluster, we find players like Karatsev (28 years old), Janko Tipsarević (37 years old, retired in 2019), Zidanšek (24 years old), Cerúndolo (23 years old), Jannik Sinner (20 years old), Świątek

"Basketball" Cluster

levert, adebayo, kobe, polonara, lamelo, bolmaro, herro, blogdebasket, avramovic, hardaway, pippen, demar, casademont, caris and lebron

Neighbors 2018

vaulet, jordan, jordans, gosens, saponjic, hispanosnba, pellistri, vapormax, pierria, cosachov, pertossi, iguodala, nike, sassoli, derozan, olynyk, vivafutbol, dwyane, uptempo, rondoblaugrana, mainoldi, mcgrady, monarriz, fitipaldo and hield

Top Shifted Farther

cristof, tvandshow, vfst, gosens, martucci, arandahoy, kyxy, adelmar, sujatovich, vivafutbol, joselui, culturemap, etchebest, palaciodelamusica, elplacerdeleer, jordans, operti, vapormax, pertossi, varesi, italiahttps, dimitrescu, cosachov, franciahttps and sassoli

Neighbors 2021

giedraitis, pierria, kawhi, lakers, campazzo, shengelia, doncic, westbrook, antetokounmpo, lillard, dwyane, derozan, olynyk, embiid, dragic, nurkic, calathes, korver, cavs, prepelic, vildoza, montrezl, iguodala, millsap and janning

Top Shifted Closer

olynyk, derozan, iguodala, korver, dragic, millsap, valanciunas, javale, vanvleet, montrezl, lamarcus, ayton, kemba, domantas, shengelia, oubre, brogdon, draymond, vucevic, redick, lillard, nurkic, jaylen, giedraitis and jrue

Figure 16: Analysis of neighbors, shifts closer and farther of the cluster about "Basketball"

"Tennis" Cluster

karatsev, janko, zidansek, cerúndolo, jannik, swiatek, opelka, koepfer, barbora, bublik, badosa, andreescu, kenin, sonego, popyrin, musetti, barty, podoroska, krejcikova, davidovich, gauff and sinner

Neighbors 2018

oggero, onestini, pellistri, polonara, betular, quaglia, bernabei, ballarini, lillini, mangiaterra, contentti, tunzi, rotondi, amorosi, operti, sonego, sassoli, muzzio, weigandt, guzzini, maddaloni, solimo, corvetto, ferrario and gargiulo

Top Shifted Farther

gorof, pellistri, kroser, quaglia, onestini, isamit, corvetto, guzzini, santucci, tunzi, anez, varesi, solimo, ballarini, arcidiácono, adelmar, maddaloni, ezequie, oggero, lamota, amorosi, contentti, operti, casamassima and mangiaterra

Neighbors 2021

sonego, berrettini, cecchinato, krajinovic, lajovic, youzhny, pouille, bedene, tipsarevic, monfils, hurkacz, berlocq, struff, delbonis, bublik, mannarino, khachanov, bagnis, thiem, miñaur, baghdatis, popyrin, wawrinka, troicki and aliassime

Top Shifted Closer

krajinovic, lajovic, khachanov, youzhny, monfils, pouille, cecchinato, thiem, raonic, wawrinka, bedene, nishikori, tipsarevic, karlovic, rublev, zverev, soderling, tsitsipas, shapovalov, azarenka, kohlschreiber, baghdatis, troicki, svitolina and berrettini

Figure 17: Analysis of neighbors, shifts closer and farther of the cluster about "Tennis"

(20 years old), Opelka (24 years old), Koepfer (27 years old), Barbora Krejčíková (26 years old), Búblik (24 years old), Badosa (24 years old), Andreescu (21 years old), Kenin (23 years old), Sonego (26 years old), Popyrin (22 years old), Musetti (19 years old), Barty (25 years old), Podoroska (24 years old), Davidocich (22 years old) and Gauff (17 years old). Again, it is a group of words mostly composed of young player names.

We can also see how some shifts were caused by other areas, such as video games. For example, "Dimitrescu" stopped being strongly related to the tennis players Alexandru Dimitrescu and Claudia-Gianina Dimitrescu, to be associated with the character Alcina Dimitrescu from the video game *Resident Evil: Village*, released in 2021.

4.10 Pornography

The clusters related to pornography were the most abundant among the identified groups, comprising at least 4.17% (24 out of 575) of the automatically generated clusters. Instead of examining all 24 clusters, we focused on the ones with the most significant semantic shifts. The cluster that experienced the most semantic change contained the word *colochas*, with an average cosine similarity of 0.5654 between its words in 2018 and 2021. Notably, in 2018, the nearest neighbors did not include terms associated with pornographic content, but in 2021, they did. This shift can cause issues with automatic systems designed to filter out such content, as the distribution shift may render them obsolete if they are not updated.

4.11 Space Missions

The cluster "space missions", that is shown in Figure 18, includes words related with recent space exploration missions. The cluster centroid is the word "*zhurong*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.4898.

In this cluster, we find the following terms related to recent space exploration missions: Tianwen (CNSA's orbiter, lander, and rover. Launched in July 2020, it landed on Mars in May 2021), CNSA (China National Space Administration, China's space agency. Established in 1993), Yutu (launched and landed in December

"Space Missions" Cluster

tianwen, cnsa, yutu, ingenuity, starship, perseverance y zhurong

Neighbors 2018

wenbin, wenliang, inosuke, ziyun, yue, zhurong, qingyue, genshin, zhen, lijian, yun, jian, zhan, jinwoo, guan, zhou, zuo, kalec, huo, shan, jin, tian, jeongguk, zong y qing

Neighbors 2021

zhurong, róver, orbitador, perseverance, tiangong, isro, cnsa, exomars, shenzhou, alunizar, tripulada, jaxa, ryugu, orbiter, churyumov, ingenuity, gerasimenko, planitia, philae, tianhe, nasa, soyuz, curiosity, cosmódromo y roscosmos

Top Shifted Farther

shan, zong, futakuchi, guan, duan, shandril, kathos, ravelia, yeho, gawan, chu, rathma, aluxe, panth, jian, yun, jeongguk, huo, ngot, qingyue, ziyun, yue, inosuke, wenbin y wenliang

Top Shifted Closer

róver, orbitador, perseverance, exomars, tiangong, ingenuity, orbiter, cnsa, isro, tripulada, churyumov, philae, jaxa, zhurong, alunizar, gerasimenko, sonda, planitia, nanosatélite, soyuz, curiosity, baikonur, nasa, hirise y suborbital

Figure 18: Analysis of neighbors, shifts closer and farther of the cluster about "Space Missions"

2013. A lunar rover from CNSA's Chang'e 3 mission), Ingenuity (first flight on Mars in April 2021. Part of NASA's Mars 2020 mission), Starship (name of SpaceX's high-capacity rockets, announced in September 2018), Perseverance (Mars rover, part of NASA's Mars 2020 mission), and Zhurong (rover from CNSA's Tianwen-1 mission. Deployed in May 2021).

4.12 K-Pop

The cluster "K-Pop", that is shown in Figure 19, includes words related with K-Pop groups and singers. The cluster centroid is the word "*ateez*". In average, the cosine similarity of the words in the cluster, relative to itself in 2018 and 2021 is 0.6362.

"K-Pop" Cluster

boyz, lightstick, ateez, blackpink, monsta and itzy

Neighbors 2018

infite, jisoo, gfriend, nswf, newu, mamamoo, gmaq, kyxy, cyou, cuepoint, arrowed, sublabel, fastlove, vyral, bts, runeterra, bamboozle, cubetype, tiktokers, shinee, bizarrap, geeh, bangtan, dreame and hundle

Neighbors 2021

mamamoo, gfriend, bangtan, shinee, jisoo, taeyeon, taemin, kpop, hyuna, bts, soompi, seventeen, jyp, mnet, kdrama, joong, boyband, babymetal, normani, stray, btob, idols, nct, hyuk and seokjin

Top Shifted Farther

psih, generalmills, xxyyxx, peolpe, cyou, ebbl, ffrk, xlju, sublabel, culturemap, fastlove, interratial, newshttps, geeh, wadoo, hundle, bamboozle, arrowed, cubetype, kyxy, newu, nswf, infite, gmaq and cuepoint

Top Shifted Closer

mamamoo, gfriend, hyuna, taeyeon, bts, boyband, shinee, kpop, bangtan, mnet, soompi, jyp, jisoo, taemin, hallyu, nct, seventeen, idols, chainsmokers, jeon, normani, joong, quavo, cnco and aew

Figure 19: Analysis of neighbors, shifts closer and farther of the cluster about "K-Pop"

In the K-Pop cluster, we found groups like Ateez (debuted in 2018), Monsta (2015), Blackpink (2016), Boyz (2017), and Itzy (2019). These groups were created relatively recently, and the word embeddings associated with their names have moved closer to the word embeddings associated with more established groups or singers like Hyuna (2007), Taeyeon (2007), BTS (2013), or record label companies like JYP (1997).

5 Emotional shift analysis

To illustrate how semantic shifts can impact natural language processing models, we implemented an unsupervised emotion classification model based on the algorithm proposed by [24]. Instead of determining the emotion for the words in the cluster, the *softmax* function was applied over the distances between each word and the words that belong to each emotion. Those values result in a probability distribution, where the sum of all the "intensities" is one.

5.1 Video games

In Figure 20 the emotion shift between 2018 and 2021 for this cluster is shown. The closest emotions to the video games cluster in 2021 were joy/cheerfulness, joy/zest, anger/rage, and fear/horror. Although the changes were small, there was an increase in joy/optimism. In the primary emotions, joy and sadness got closer, while anger and love moved farther apart. However, note that love/longing moved closer. Fear and surprise remained constant, although for some games, the closeness to fear increased, presumably because of the game's nature.

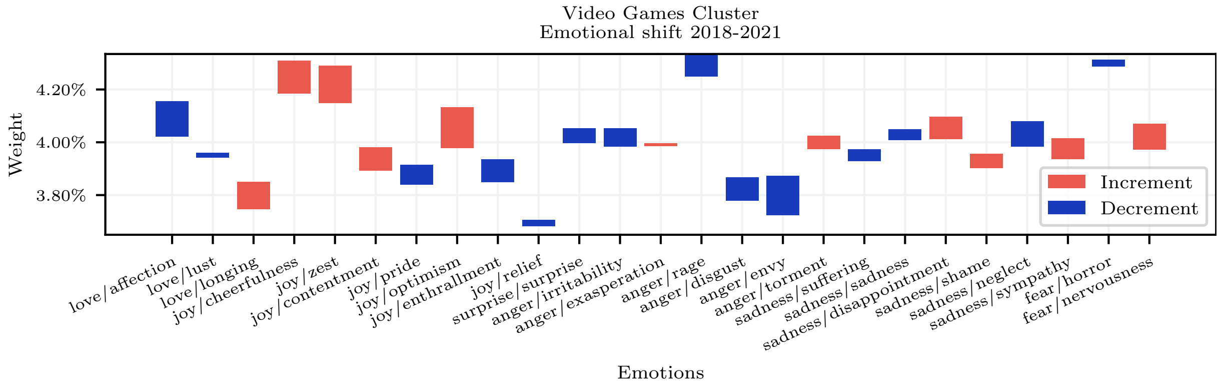


Figure 20: Relative emotional shift of words in the “Video games” cluster to secondary emotions

5.2 Cryptocurrencies

In Figure 21 the emotion shift between 2018 and 2021 for this cluster is shown. In the context of cryptocoins, the emotion with the highest increase was sadness/sympathy. In fact, that is the dominant secondary emotion, followed by joy/zest and joy/optimism. An interesting change happened with the term *taaprot*, which is a Bitcoin improvement to implement smart contracts and improve privacy. The increase in the weight of joy/optimism is one of the biggest in this cluster, which suggests that this proposal gained trust in the community.

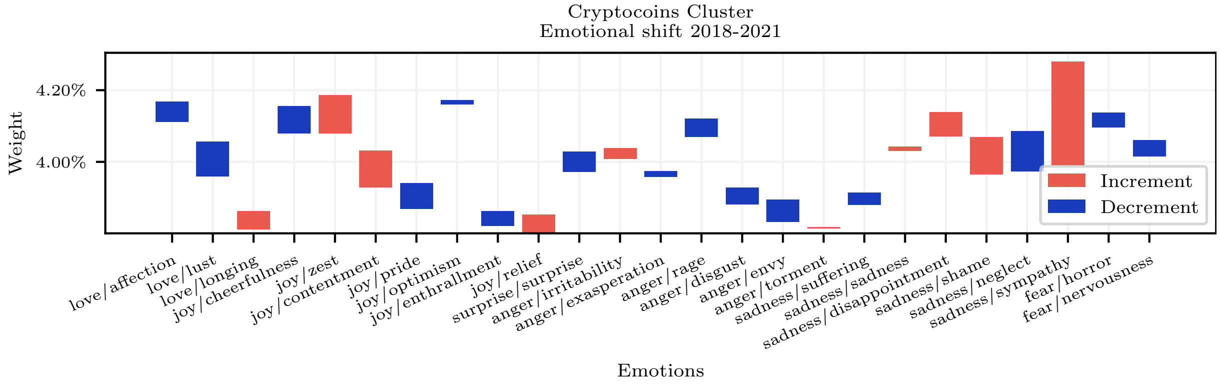


Figure 21: Relative emotional shift of words in the “Cryptocoins” cluster to secondary emotions

5.3 COVID-19

In the case of the COVID-19 cluster, the dominant emotions correspond to fear/horror, fear/nervousness, and sadness/neglect, as shown in Figure 22. These three emotions also had the greatest proximity. Although “COVID” as a disease emerged in 2019, the other words in the cluster existed before, so it is possible to measure a difference in the proximity to emotions.

If we review the tertiary emotions within fear, we can observe that the greatest increases were in horror/alarm and nervousness/suspense. These changes are shown in Figure 23.

5.4 COVID-19 Vaccines

In the COVID-19 vaccines cluster, the biggest emotional shifts occurred in love/longing (closer) and joy/cheerfulness (farther), as shown in Figure 24. The dominant emotions in this cluster were joy/optimism, anger/irritability, and fear/horror.

5.5 Vaccination

In the vaccination cluster, as shown in Figure 25, the dominant secondary emotion is fear/horror, followed by anger/rage and sadness/neglect. However, these were not the emotions that underwent the most significant changes. The graph suggests that the fear of vaccines already existed, as fear/horror was the dominant emotion in both periods. The most notable shifts occurred towards love/longing and joy/optimism.

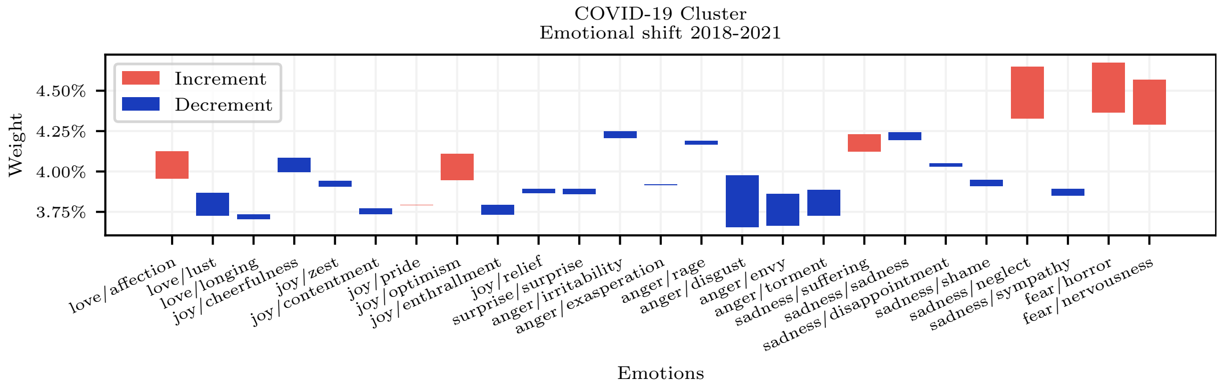


Figure 22: Relative emotional shift of words in the “COVID-19” cluster to secondary emotions

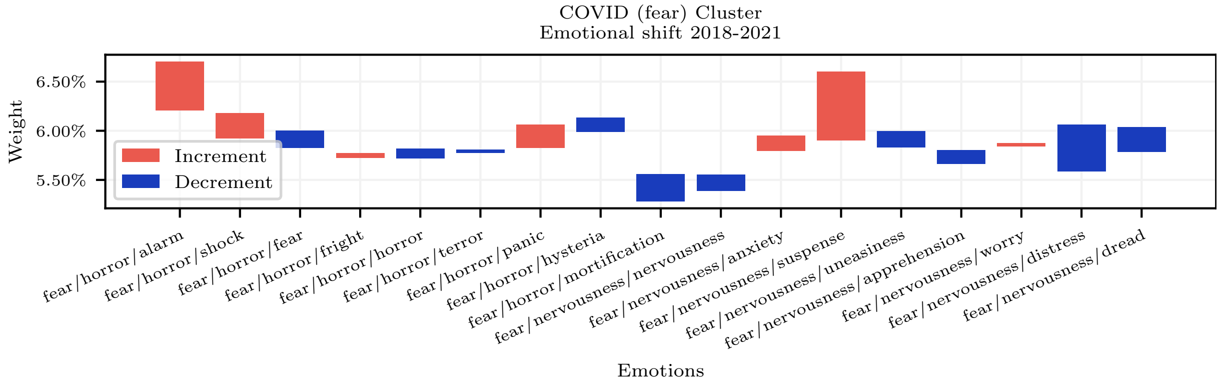


Figure 23: Relative emotional shift of words in the “COVID-19” cluster to tertiary emotions of fear

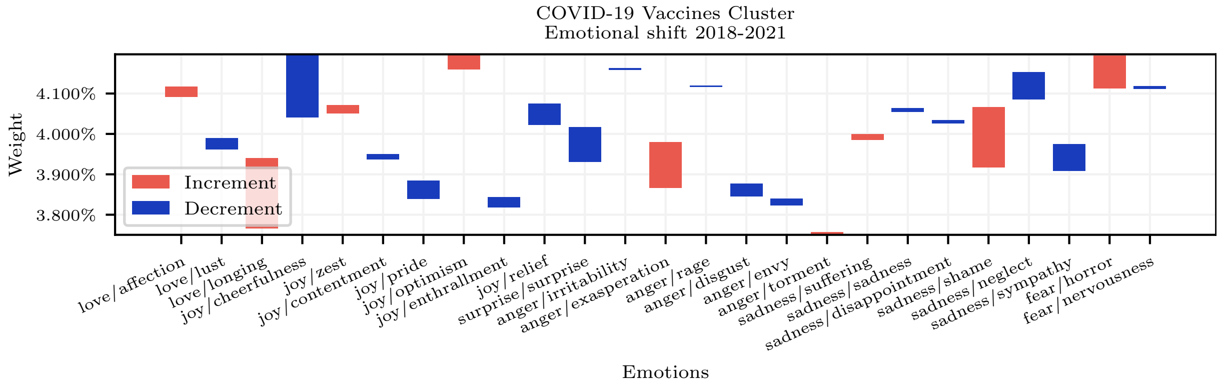


Figure 24: Relative emotional shift of words in the “COVID-19 vaccines” cluster to secondary emotions

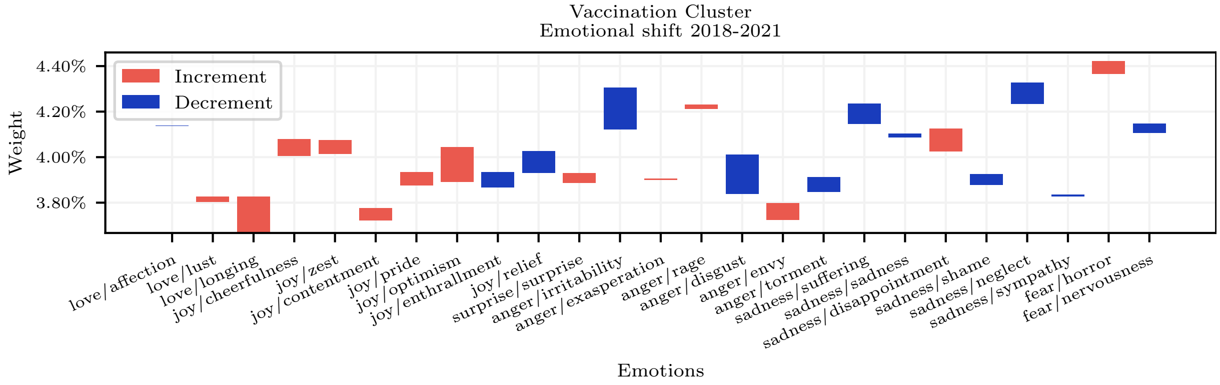


Figure 25: Relative emotional shift of words in the “vaccination” cluster to secondary emotions

5.6 COVID-19 Testing

In the COVID-19 testing cluster, the most prevalent secondary emotions are sadness/neglect, fear/horror, love/affection, anger/irritability, and sadness/suffering. The biggest emotional shift occurred in sadness/neglect, fear/horror, and love/affection. Regarding sadness/neglect, the dominant tertiary emotion is isolation, which increased compared to the previous period. Those changes are shown in Figure 26.

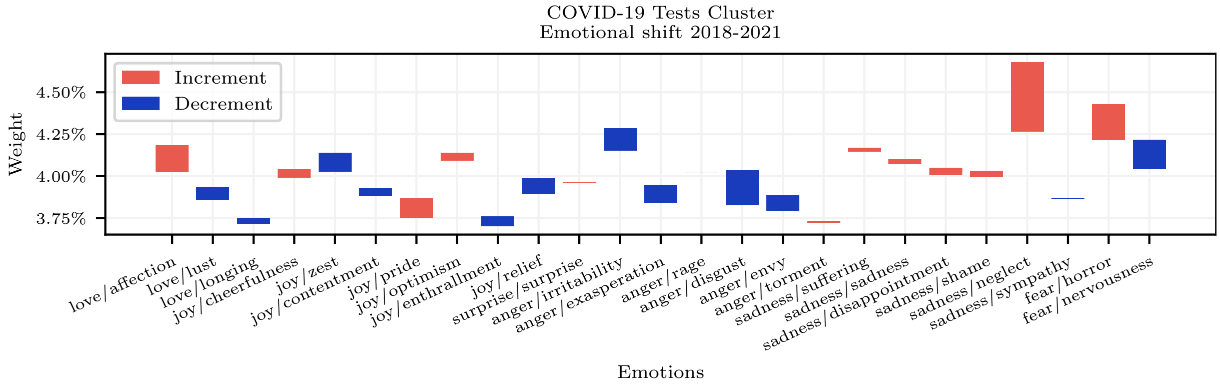


Figure 26: Relative emotional shift of words in the “COVID-19 testing” cluster to secondary emotions

5.7 Masks

The cluster of terms related to masks is characterized by dominant secondary emotions such as love/affection, anger/irritability, sadness/neglect, and fear/horror, as illustrated in Figure 27. However, the results of measuring the difference reveal that these words have shifted away from anger/irritability and towards love/affection and sadness/neglect.

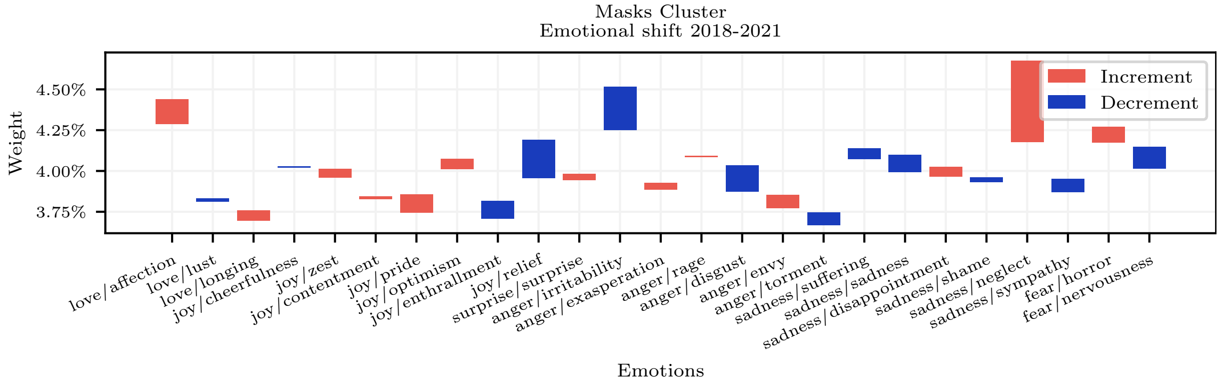


Figure 27: Relative emotional shift of words in the “masks” cluster to secondary emotions

5.8 Sports

In the sports cluster, there is a clear increase in the weight of the secondary emotions joy/pride and love/affection, although the extent of the increase varies depending on the specific sport.

5.8.1 Cycling

Within the cycling cluster, the dominant emotion is joy/pride, followed by anger/rage, sadness/sadness, and fear/horror. The emotion that saw the largest increase in weight was joy/pride. The changes are shown in Figure 28.

5.8.2 Soccer

Within the soccer cluster, the most dominant secondary emotion is love/affection, followed by anger/rage, sadness/sadness, and sadness/neglect. The largest increase occurred in the love/affection emotion, with sadness/shame also showing a notable increase. The changes are shown in Figure 29.

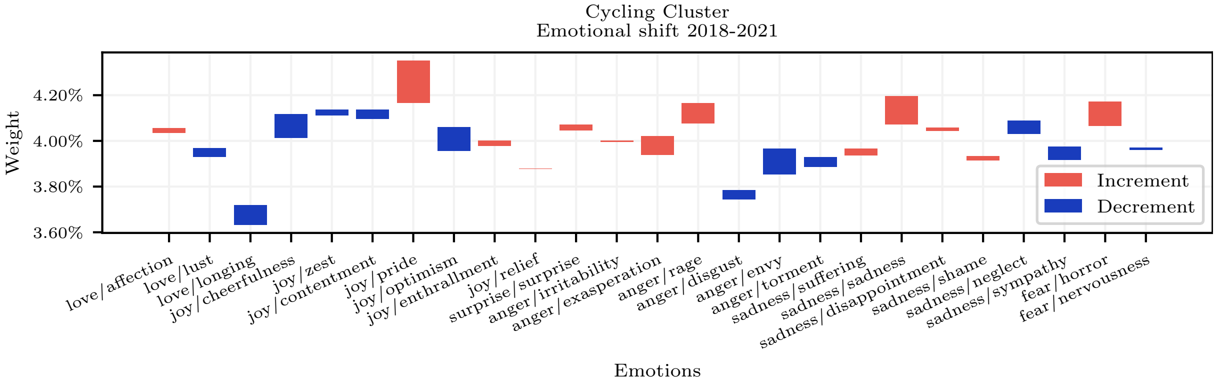


Figure 28: Relative emotional shift of words in the “Cycling” cluster to secondary emotions

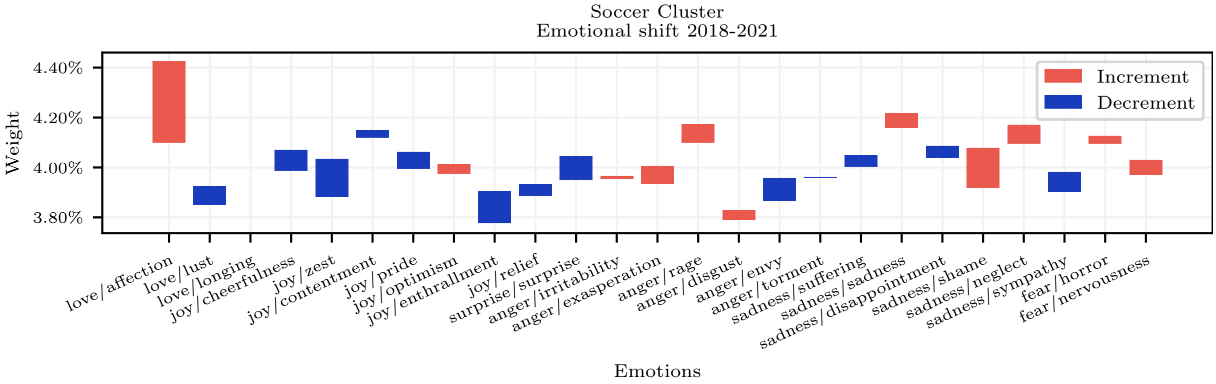


Figure 29: Relative emotional shift of words in the “Soccer” cluster to secondary emotions

5.8.3 Basketball

In the basketball cluster, the dominant secondary emotion is love/affection, followed by joy/pride, anger/rage, and sadness/sadness. The largest increase occurred in the love/affection emotion, while it moved further away from joy/contentment and anger/envy. Similar to the clusters for cyclists and tennis players, love/affection is the most prominent secondary emotion in the basketball cluster. The changes are shown in Figure 30.

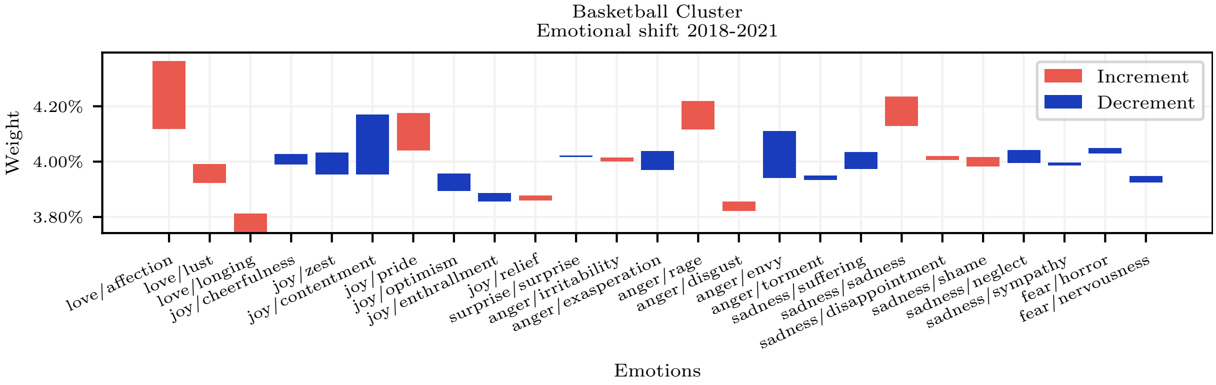


Figure 30: Relative emotional shift of words in the “Basketball” cluster to secondary emotions

5.8.4 Tennis

In the tennis cluster, the dominant secondary emotions are sadness/sadness, joy/pride, and love/affection. The largest increase occurred in the love/affection emotion. The changes are shown in Figure 31.

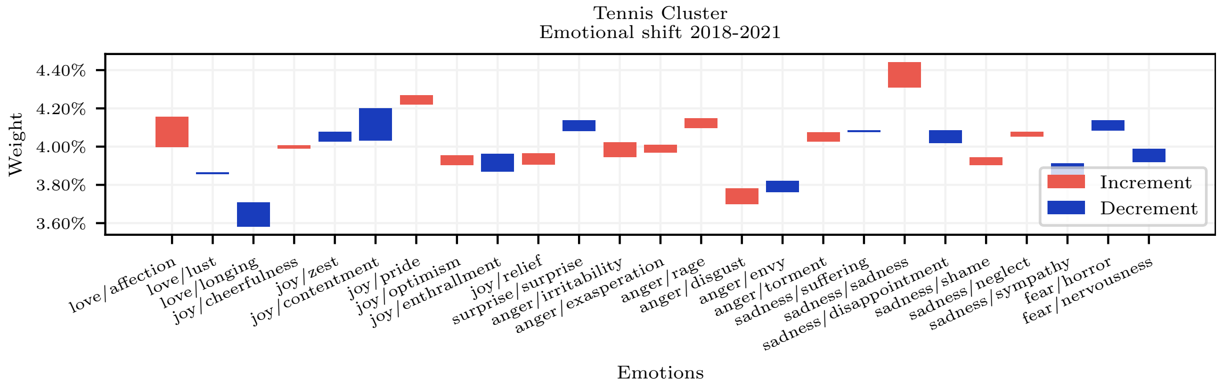


Figure 31: Relative emotional shift of words in the “Tennis” cluster to secondary emotions

5.9 Pornography

The cluster containing terms associated with pornography has love/lust as the dominant secondary emotion, followed by joy/cheerfulness and joy/zest. The largest increase occurred in the secondary emotion love/lust, suggesting that the words in the cluster were likely not previously used in this context. The changes are shown in Figure 32.

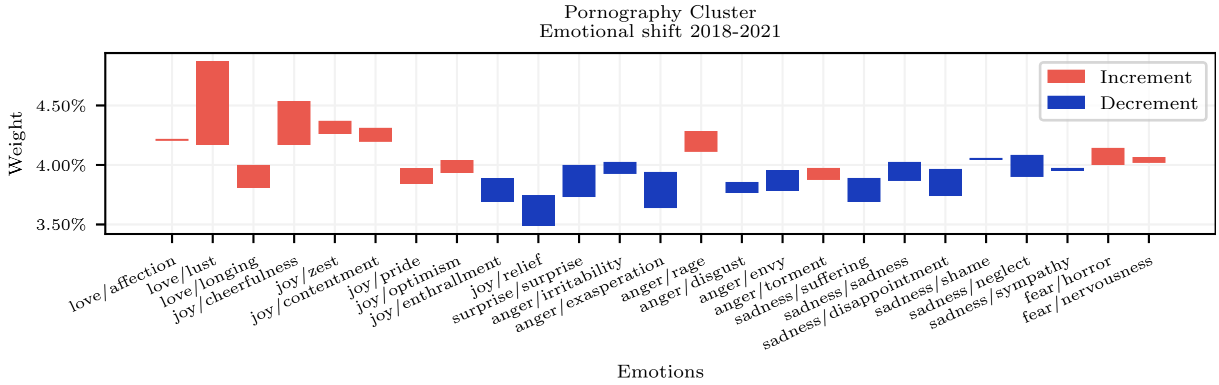


Figure 32: Relative emotional shift of words in the “Pornography” cluster to secondary emotions

5.10 K-Pop

As shown in Figure 33, in the K-Pop cluster the dominant emotions are love/affection and joy/zest. The biggest emotion shifts were an increase in anger/torment, love/lust and surprise/surprise, while there were big decrements in fear/horror, sadness/suffering and sadness/sadness.

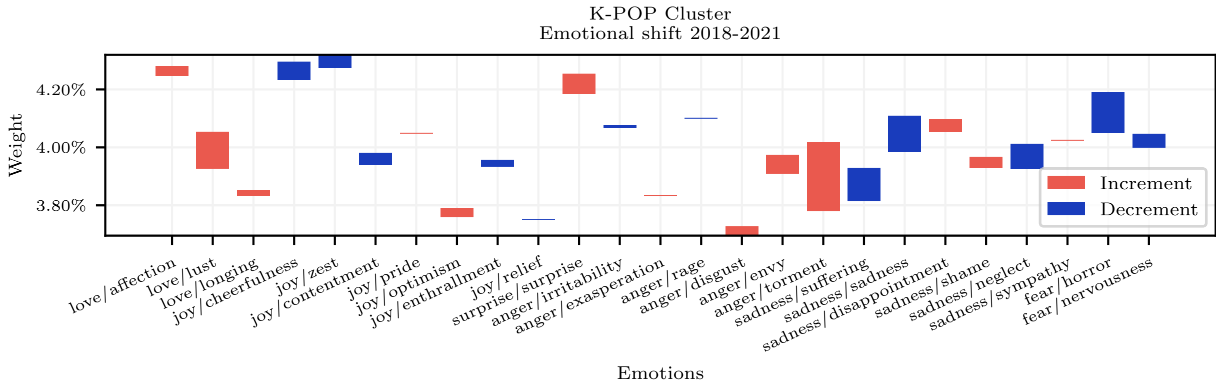


Figure 33: Relative emotional shift of words in the “K-Pop” cluster to secondary emotions

6 Conclusions

In this work, a diachronic word embedding model based on word2vec was created to perform semantic shift analysis of Spanish words. Our work contributes a new diachronic word embedding model that shows the semantic shift of words during significant events, such as the COVID-19 outbreak.

The diachronic word embedding was built using data collected from the open web. We were able to observe that between the years 2018 and 2021, there was significant semantic shift in several areas, such as videogames, cryptocurrency, COVID-19 topics, the entertainment industry, sports, social networks, pornography, and space missions, among others. It is important to emphasize that such topics were discovered using unsupervised clustering algorithms: the clusters emerged from the data, they were not defined manually.

Although both corpora are near in time, significant semantic shift was found in some topics. And those changes were not only caused by a once-in-a-lifetime event like the COVID-19 pandemic, but also by everyday events such as the release of a new videogame or the debut of an athlete or singer. This highlights the need to regularly retrain natural language processing models with fresh data. For instance, if a model that reflects current reality is desired, it is important to note that after a relatively short period of time, some areas of the data, such as new products, sports, and entertainment, may become outdated due to the aging of the training data, leading to the need for frequent retraining of natural language processing models. On the other hand, shifts in pornography-related clusters may make content filters less precise in the future. It is important to learn to tackle this issue, as it may not only affect adult content filters, but also mechanisms used to detect fake news or hate speech.

Most semantic shift studies have focused on shift through big time spans, like decades or even centuries. However, this study has shown that significant shifts can occur in relatively short periods, like just a few years. This highlights the importance of analyzing semantic shifts in shorter time frames, as it provides a more up-to-date understanding of how language is evolving in response to contemporary events and trends. Moreover, analyzing semantic shifts over shorter time frames can help researchers identify emerging trends and changing attitudes in real-time, potentially providing valuable insights for business, governments and other organizations that need to be aware of society's attitudes and preferences.

7 Future work

This study was conducted using word2vec. However, this word embedding model is not context-sensitive. The techniques proposed by [28] could be used to analyze semantic shifts using a model like BERT. Nevertheless, the cost of training BERT with a volume of data similar to that used in this research could be significantly high regarding what is intended to be analyzed. Therefore, it would be interesting to develop new techniques with a similar effectiveness to BERT, but with a lower training cost.

This work was also limited to two points. With collections such as CommonCrawl, it would be possible to carry out this type of study with greater granularity and over more years.

Regarding the emotional shift analysis, it was assumed that a direct translation of the words in the English hierarchy presented by [23] to Spanish words are possible. Future work is needed to determine how well that hierarchy maps to the emotions of native Spanish speakers, or if there are subtle differences that should be added into the taxonomy. Furthermore, correcting for the semantic shift of the words representing the emotions was not done.

8 Acknowledgments

We would like to thank Dr. Gabriela Marín Raventós and Dr. Kryscia Daviana Ramírez Benavides for their support in the final graduation project, of which this research is a part. Additionally, we extend our thanks to the CLEI anonymous reviewers for their valuable feedback.

References

[1] E. Rodriguez-Betancourt and E. Casasola-Murillo, "Analysis of semantic shift before and after covid-19 in spanish diachronic word embeddings," in *2022 XVLIII Latin American Computer Conference (CLEI)*, 2022, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CLEI56649.2022.9959896

[2] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. [Online]. Available: https://www.aclweb.org/anthology/C18-1117

[3] E. C. Traugott, "Semantic change," 03 2017. [Online]. Available: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-323

[4] T. Mikolov and otros, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[5] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 746–751. [Online]. Available: https://www.aclweb.org/anthology/N13-1090

[6] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically significant detection of linguistic change," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2015, p. 625–635. [Online]. Available: https://doi.org/10.1145/2736277.2741627

[7] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1489–1501. [Online]. Available: https://www.aclweb.org/anthology/P16-1141

[8] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *j-PSYCHO*, vol. 31, no. 1, pp. 1–10, 1966.

[9] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong, "Discovery of evolving semantics through dynamic word embedding learning," *CoRR*, vol. abs/1703.00607, 2017. [Online]. Available: http://arxiv.org/abs/1703.00607

[10] H. Gong, S. Bhat, and P. Viswanath, "Enriching word embeddings with temporal and spatial information," in *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1–11. [Online]. Available: https://aclanthology.org/2020.conll-1.1

[11] M. Asif, D. Zhiyong, A. Iram, and M. Nisar, "Linguistic analysis of neologism related to coronavirus (COVID-19)," *Social Sciences & Humanities Open*, vol. 4, no. 1, p. 100201, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590291121000978

[12] Y. Guo, C. Xypolopoulos, and M. Vazirgiannis, "How COVID-19 is changing our language : Detecting semantic shift in twitter word embeddings," 2021.

[13] C. S. Butler and A.-M. Simon-Vandenbergen, "Social and physical distance/distancing: A corpus-based analysis of recent changes in usage," *Corpus Pragmatics*, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s41701-021-00107-2

[14] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger, "The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora," in *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2021, pp. 13–20. [Online]. Available: https://aclanthology.org/2021.smm4h-1.3

[15] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/bengio03a.html

[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1532–1543. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1162.pdf

[18] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. [Online]. Available: https://aclanthology.org/E17-2068

[19] S. Ji, N. Satish, S. Li, and P. Dubey, "Parallelizing word2vec in multi-core and many-core architectures," 2016. [Online]. Available: https://arxiv.org/abs/1611.06172

[20] S. Gupta and V. Khare, "Blazingtext: Scaling and accelerating word2vec using multiple gpus," in *Proceedings of the Machine Learning on HPC Environments*, ser. MLHPC'17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3146347.3146354

[21] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://www.aclweb.org/anthology/D18-2029

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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).* Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[23] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach." *Journal of Personality and Social Psychology*, vol. 52, no. 6, pp. 1061–1086, 1987. [Online]. Available: https://doi.org/10.1037/0022-3514.52.6.1061

[24] M. Alshahrani, S. Samothrakis, and M. Fasli, "Word mover's distance for affect detection," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, 2017, pp. 18–23.

[25] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, 1970.

[26] "Corpus de referencia del español actual," Real Academia Española. [Online]. Available: http://www.rae.es

[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[28] S. Montariol, "Models of diachronic semantic change using word embeddings," Ph.D. dissertation, Université Paris-Saclay, Feb. 2021. [Online]. Available: https://tel.archives-ouvertes.fr/tel-03199801