



Regionalized models for Spanish language variations based on Twitter

Eric S. Tellez^{1,2,5} · Daniela Moctezuma³ · Sabino Miranda^{1,2,4} · Mario Graff^{1,2} · Guillermo Ruiz^{1,3}

Accepted: 23 January 2023 / Published online: 2 March 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract

Spanish is one of the most spoken languages in the world. Its proliferation comes with variations in written and spoken communication among different regions. Understanding language variations can help improve model performances on regional tasks, such as those involving figurative language and local context information. This manuscript presents and describes a set of regionalized resources for the Spanish language built on 4-year Twitter public messages geotagged in 26 Spanish-speaking countries. We introduce word embeddings based on FastText, language models based on BERT, and per-region sample corpora. We also provide a broad comparison among regions covering lexical and semantical similarities and examples of using regional resources on message classification tasks.

Keywords Linguistic resources · Semantic space · Spanish Twitter

1 Introduction

Communication is, at its core, an understanding task. Understanding a message implies that peers know the vocabulary and structure; i.e., the receiver obtains what the sender intended to say. Language is a determinant factor in any communication. Even people who speak the same language can find difficulties communicating information due to slight language variations due to regional variations, language evolution, cultural influences, and informality, to name a few.

A dialect is a language variation that diverges from its origin due to several circumstances. Dialects can differ regarding their vocabulary, grammar, or even semantics. The same sentence can be semantically different among dialects. In

Daniela Moctezuma, Sabino Miranda, Mario Graff, and Guillermo Ruiz have contributed equally to this work.

✉ Daniela Moctezuma
dmoctezuma@centrogeo.edu.mx

Extended author information available on the last page of the article

contrast, people of different dialects may need help understanding sentences with the same meaning. This effect is notoriously complex for figurative language since it contains cultural and ideological references. Studying these dialects can help us understand the cultural aspects of each population and the closeness between them. In this sense, dialectometry studies the regional distribution of dialects. Similarly, dialectometric analysis has as its objective, through a computational approach, to analyze this distribution and provide a quantitative linguistic distance between each dialect or region of it Donoso et al. (2017). Hence, the research in dialectology tries to understand language differences, innovations, or variations not only in space but also in time through several phenomena. Natural Language Processing (NLP) tools can analyze written communications automatically. However, the support for regional languages is still in its early stages in NLP, particularly for languages different from English.

Another important aspect to discuss is about if dialects are directly related to space or geographical boundaries (Hoff, 2020). As Penny et al. (2000) discuss that language people who live in different territory do not speak the same way because the neighboring locality plays a key role. Also, the social and historical language variations are discussed in Penny et al. (2000). In some way, with the analysis done we hope to contribute a little to this exciting discussion. A more linguistic comparison from Spanish variants in America could be found in Cotton and Sharp (1988).

On the other side, social media is a crucial component of our lives; Facebook, Twitter, Instagram, and Youtube are among the most used social networks that allow interaction among users in written form and other media. In particular, Twitter is a micro-blogging platform where most messages are intentionally publicly available, and developers and researchers can access these messages through an application programming interface (API). Twitter's messages are known as *tweets*. Each tweet contains text and additional metadata like the user that wrote it and the geographic location where it was published. In the case of social media, where the source of the messages is informal, the errors are another source of variability in the language. This kind of messaging may impose extra difficulties than in formal written documents.

Nonetheless, Twitter messages' quality and socio-demographic representativeness have been continuously questioned (Crampton et al., 2013). Some authors have shown that despite the over-representation of some social groups, social media usage can still be of enormous usefulness and quality (Huang et al., 2016). Language and geographical information are crucial to understanding the geographies of this online data and how some information related to economic, social, political, and environmental trends could be used (Graham et al., 2014).

It is easy to observe the relevance of having quality and realistic data for language variations analysis or model generation with this background. Sometimes, it is difficult to have a regional-specific corpus. A large corpus is necessary to learn language models and achieve confident analysis; more data often imply better models. With this kind of resources acquired, for instance, from social media platforms, many potential research and applications have been made in many research areas, such as health (Paul & Dredze, 2011), environmental issues

(Mooney et al., 2009), emotion (Suhasini & Srinivasu, 2020), mental health (Fingfeld-Connett, 2015), gender (Vashisth & Meehan, 2020), and misogyny (Frenda et al., 2019), among others.

There are several examples of linguistic resources attending specific tasks or applications, such as the study of Down syndrome in Escudero-Mancebo et al. (2022). Also, on CKennedy et al. (2022), a corpus is proposed for hate-based rhetoric, or hate recognition. On Gruszczyński et al. (2022), the authors provide a corpus of up to 13.5 million tokens of Polish texts between 1601 and 1772. As can be seen, a variety of corpus has been generated by the community to deal with general or more specific NLP tasks.

On the other hand, the Spanish language variations are also studied in the NLP research community. For instance, in Gonçalves and Sánchez (2014), authors present a crowdsourcing language diatopic variation using Twitter data with geolocation, employing tweets messages in Spanish for more than two years over the globe. The analysis was made with a set of pre-established words and counting each and its variations worldwide. In this sense, to know what regions are close to each other, the authors used the *k-means* clustering algorithm over these words frequencies and principal component analysis (PCA) to reduce data into two-dimensional space for visualization. The clustering approach identifies large macro-regions sharing language characteristics.

Unfortunately, for the Spanish language, there are few works, contrary to some language variations from Europe, e.g., English dialects (Hovy et al., 2020), French dialects (Lamontagne & McCulloch, 2022) and Arabic (Alshutayri & Atwell, 2017), to name some.

One of the possible assumptions using Twitter is that the behaviors of English users generalize to other language users. In Hong et al. (2011) is presented a study using 62 million tweets over more than 100 different languages over four weeks. Applying an automatic language detection algorithm, they found that most of the data were in English (51%), and 39% was for other languages such as Japanese, Portuguese, Indonesian, and Spanish.

The geographical region of a language helps to know how this language is used in a particular society. For instance, Spanish is a largely used language; nevertheless, it is used differently according to the country or even a more specific geographical location. Hence, the language could analyze at the regional level (Huang et al., 2016; Rodriguez-Diaz et al., 2018). In Rodriguez-Diaz et al. (2018), the authors study Spanish language variations in Colombia. The analysis used unigram features, and the authors stated that it was challenging to compare Spanish variations against regions identified by other authors using classical dialectometry. Hence, in conclusion, the authors said that automatic detection of *dialectones* is an adequate alternative to classical methods in dialectometry for automated language applications. A more current effort to deal with the Spanish Language on Twitter was presented by Huertas-Tato et al. (2022), which provides a powerful tool to take advantage of transformers generated with the native language. The authors test their solution with several classical NLP tasks.

The number of articles related to other languages is also reduced. In Alshutayri and Atwell (2017), the Arabic dialects were classified using the WEKA tool¹ reaching an accuracy of 79% on their classification results. The used dataset contained 210,915K tweets from some Arabic dialects and classified them considering their geographic location.

Mocanu et al. (2013) survey the linguistic landscape in the world using Twitter. This landscape includes linguistic homogeneity or variations over countries that consider the touristic seasons. The method employed to identify language is the Chromium Compact Language Detector by Google; the authors also used the location of the devices reported by tweets. As a result, it was possible to observe distributions of language on Twitter over several countries by month of the year and where touristic flow is evident.

Also, emoticons or emojis are effective communication symbols. Their usage has been studied in the literature. For instance, in Park et al. (2013), authors analyzed the semantic, cultural, and social aspects of their use on Twitter. Kejriwal et al. (2021) studied the use of emojis in terms of linguistic use and countries. The authors collected tweets from 30 different languages and countries, and the authors found that emojis usage strongly correlates between language and country level, which means that emojis are used according to language and region. Another example of studying Emojis is presented in Li et al. (2019).

Other efforts have been made to exploit the regionalized models for a specific language variation; for instance, in Jimenez et al. (2018) where methods to identify regional words and provide their meaning is studied.

Our contribution is a set of regionalized resources for different variations of the Spanish language. We created and characterized regional vocabularies and regional semantic representations of them (i.e., word embeddings). Also, we learn and test language models based on BERT. We built these resources from an extensive collection of public tweets from 2016 to 2019, written in 26 countries with a large basis of Spanish-speaking people. Regarding messages, we provide a sample of Twitter message identifiers divided by region such that researchers can retrieve them easily. Finally, we show some usage examples of our resources.

The rest of the manuscript is organized as follows. Section 2 describes our Twitter Spanish Corpora (TSC), used to generate our regional resources. Section 3 compare lexical traits among the corpora. Section 4 is dedicated to presenting our semantic resources and their affinity analysis that includes visualizations and experimental evidence that support the use of regional word embedding models on regional tasks. Our resources based on language models are presented and compared in Sect. 5. Finally, Sect. 6 summarizes and discusses the implications of the TSC.

¹ <https://www.cs.waikato.ac.nz/ml/weka/>.

Table 1 Datasets' statistics after filtering by retweets and ensuring at least five words per tweet

Country	Code	α	β	Number of users	Number of tweets	Number of tokens
Argentina	AR	0.7563	1.8594	1376K	234.22 M	2,887.92 M
Bolivia	BO	0.7509	1.8913	36K	1.15 M	20.99 M
Chile	CL	0.7555	1.8874	415K	45.29 M	719.24 M
Colombia	CO	0.7562	1.8993	701K	61.54 M	918.51 M
Costa Rica	CR	0.7447	1.8595	79K	7.51 M	101.67 M
Cuba	CU	0.7640	1.8677	32K	0.37 M	6.30 M
Dominican Republic	DO	0.7544	1.8832	112K	7.65 M	122.06 M
Ecuador	EC	0.7538	1.8968	207K	13.76 M	226.03 M
El Salvador	SV	0.7494	1.9066	49K	2.71 M	44.46 M
Equatorial Guinea	GQ	–	–	1K	8.93K	0.14 M
Guatemala	GT	0.7498	1.9175	74K	5.22 M	75.79 M
Honduras	HN	0.7486	1.8941	35K	2.14 M	31.26 M
Mexico	MX	0.7557	1.8895	1,517K	115.53 M	1635.69 M
Nicaragua	NI	0.7445	1.8535	35K	3.34 M	42.47 M
Panama	PA	0.7559	1.8952	83K	6.62 M	108.74 M
Paraguay	PY	0.7511	1.8815	106K	10.28 M	141.75 M
Peru	PE	0.7583	1.8966	271K	15.38 M	241.60 M
Puerto Rico	PR	0.7498	1.8433	18K	0.58 M	7.64 M
Spain	ES	0.7648	1.9036	1278K	121.42 M	1908.07 M
Uruguay	UY	0.7516	1.8346	157K	30.83 M	351.81 M
Venezuela	VE	0.7614	1.8959	421K	35.48 M	556.12 M
Brazil	BR	0.7681	1.9389	1604K	27.20 M	142.22 M
Canada	CA	0.7652	1.9331	149K	1.55 M	21.58 M
France	FR	0.9372	1.9324	292K	2.43 M	27.73 M
Great Britain	GB	0.7687	1.9129	380K	2.68 M	34.62 M
United States of America	US	0.7666	1.8929	2,652K	40.83 M	501.86 M
Total				12 M	795.74 M	10,876.25 M

We show the origin country, the country code in ISO 3166-1 alpha-2 format reported by the Twitter API, the number of tweets, and the number of different users in the collected period

2 Twitter corpora of the Spanish language

With 489 million native speakers in 2020,² Spanish is one of the languages with a higher native speaking basis, just ranked behind Chinese Mandarin in terms of the number of native speakers. Twenty-one countries have the Spanish language as the official language (by law or *de facto*).³ Our corpora selected these regions,

² <https://blogs.cervantes.es/londres/2020/10/15/spanish-a-language-spoken-by-585-million-people-and-489-million-of-them-native>.

³ https://en.wikipedia.org/wiki/List_of_countries_where_Spanish_is_an_official_language.

see Table 1; and we also considered five additional regions (US, CA, GB, FR, and BR) with well-known migration, business, and tourism activities of Spanish speakers. The number of Twitter users varies with each country; since each country has different social, political, security, health, and economic conditions, we will avoid generalizations.

As mentioned, we collected publicly published tweets between 2016 and 2019 using the Twitter stream API. Also, we limited our collection to geotagged messages marked by Twitter as written in Spanish. We decided to let out the corpus messages from the year 2020 and posteriors to avoid disturbances in social media regarding the COVID-19 pandemic. Please recall our objective is to build resources based on the language itself and not analyzing the pandemic event. Twitter stream API allows tweet retrieval in two ways. The first consists of using a language marker (*lang=es*, for Spanish) and a list of tracking words linked to the specified language. In this case, we can use Spanish *stopwords*⁴ to maximize the download process. The second strategy consists in using a language marker (*lang=es*, for Spanish) and geographical coordinates, these kinds of tweets are named geotagged Tweets. We specify worldwide coordinates to get tweets from everywhere. We use only geotagged tweets from the last strategy. These geotagged tweets have information such as country code corresponding to the country where the tweet was published, among other metadata. We rely on the information provided by Twitter about the country associated with each tweet.

To ensure a minimum amount of information in each tweet, we discard those tweets with less than five tokens, i.e., words, emojis, or punctuation symbols, following the strategy of Mikolov et al. (2013) for analyzing and learning from very large collections. We also removed all retweets to avoid duplication of messages and reduce foreign messages commented on by Spanish speakers. After this filtering procedure, we retain close to 800 million messages.

Table 1 shows statistics about our corpora describing aspects such as country, number of users, number of tweets, and number of tokens. The table shows that Spain, the USA, Mexico, and Argentina are countries with more users. Furthermore, they are also those with more tweets in the Spanish language, but the USA falls considerably in this aspect. A similar proportion is observed in the number of tokens column. However, Argentina has the highest number of tokens, above Mexico and Spain significantly.

The table also lists the coefficients for the expressions behind Heaps' and Zipf's laws. Both laws are broadly surveyed in the literature; for instance, Gelbukh et al. (2001) and Schütze et al. (2008, Chapter 5) describe them and study their implications from a general perspective. In a nutshell, the laws describe how the vocabulary grows in text collections written in non-severe-agglutinated languages. Heaps' law n^α describes the sub-linear growth of the vocabulary on a growing collection of size n . Zipf's law represents a power-law distribution where a few terms have very high

⁴ Words that are so common in a language, such as articles, prepositions, interjections, and auxiliary verbs, among other typical words.

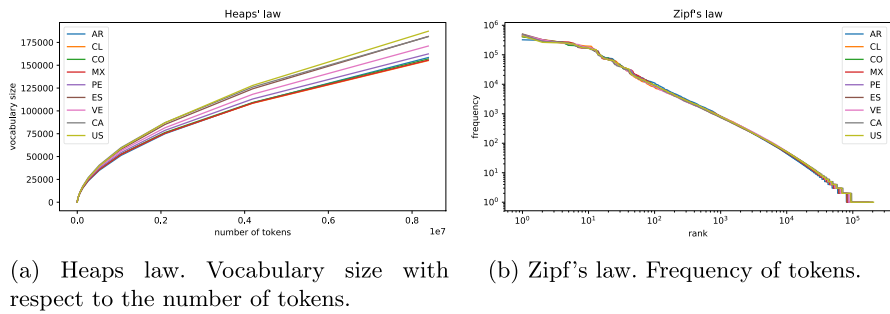


Fig. 1 The vocabulary growth and distribution of frequencies of 10^7 tokens over a sample of our Twitter's Spanish language corpora

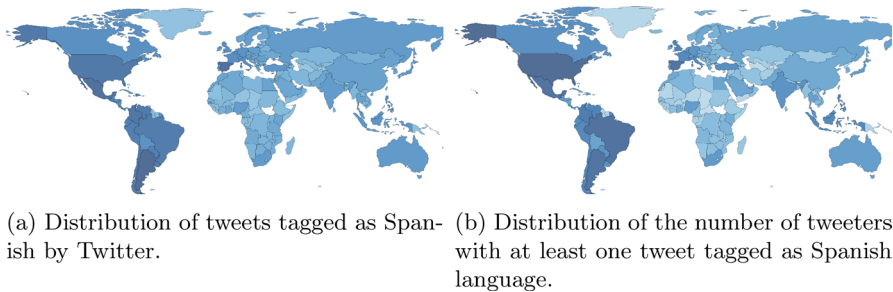


Fig. 2 Distribution of tweets and tweeters labeled as Spanish-speaking users around the world. Colors are related to the logarithmic frequencies in data collected from 2016 to 2019 with the public Twitter API stream. Darker colors indicate a high population; the logarithmic scale implies that only significant frequency differences produce color changes

frequencies, and many words occur with a shallow frequency in the collection. The expression that describes Zipf's law is $1/r^\beta$, where r is the rank of the term's frequency.

Figure 1a illustrates the Heaps' law in a small sample of regions of interest. One can observe its predicted sub-linearity and that Mexico has the lowest growth in its vocabulary size according to the number of tokens. On the contrary, the US corpus shows faster vocabulary growth, possibly explained due to the mix of languages in many messages.

Figure 1b shows Zipf's law under a log-log scale and its quasi-linear shape. We can see slight differences among curves, more noticeable on both the left and right parts of the plot. The left part of the curves corresponds to those terms with very high frequency, and the right side is dedicated to those terms being rare in the collection. Notice that all these curves are similar but slightly different; this is not a surprise since we analyze variations of the same idiom, i.e., the Spanish language.

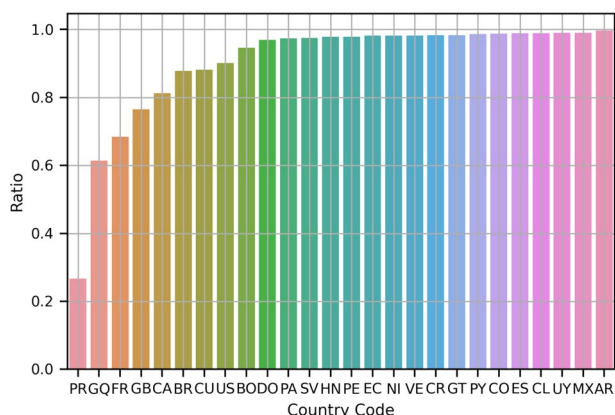


Fig. 3 Ratio between the number of tweets produces by local tweeters and the total tweets in the country

2.1 Geographic distribution

Figure 2a illustrates the number of collected Spanish language tweets over the world. The color intensity is on a logarithmic scale, which means that slight variations in the color imply significant changes in the number of messages; countries with the darkest blue have the highest number of tweets in Spanish. This figure shows how American countries (in the south, central, and north) and the Iberian Peninsula have, as expected, more tweets in the Spanish language than the rest of the world.

Figure 2b shows the distribution of tweeters (users) per country. As in the previous image, we present a logarithmic scale in the intensity of color to represent the number of users. The differences between this figure and Fig. 2a are low, as expected, and follow the same distribution. Note the high intensity of American countries.

Also, it is essential to know whether the tweets come from persons living in the country or travelers. It might be impossible to completely answer this by looking only at the tweets. However, an approach that can provide insight into the locality of the tweets is to measure the ratio between the tweets produced by local users and the total number of tweets produced in each country, where each user is assigned to the country where more tweets have. For example, suppose a tweeter has 100 tweets in Mexico and 10 in Spain. In that case, that user is considered Mexican, and those published in Spain (considered there a tourist) are not considered, which means only the tweets produced in Mexico are counted.

Figure 3 presents the ratio between local tweets and the total number of tweets; it can be observed that Puerto Rico (PR) has the lowest ratio, meaning that most tweets come from foreigners. The second lowest is Equatorial Guinea (GQ), where more than 60% of its tweets are local. Then, there is a block of countries where Spanish is not the primary language. After the block comes Cuba, with 88% of its tweets produced by locals. The United States of America (US) has a ratio of 90%. From the Dominican Republic (DO) to the right have a percentage higher than 95%.

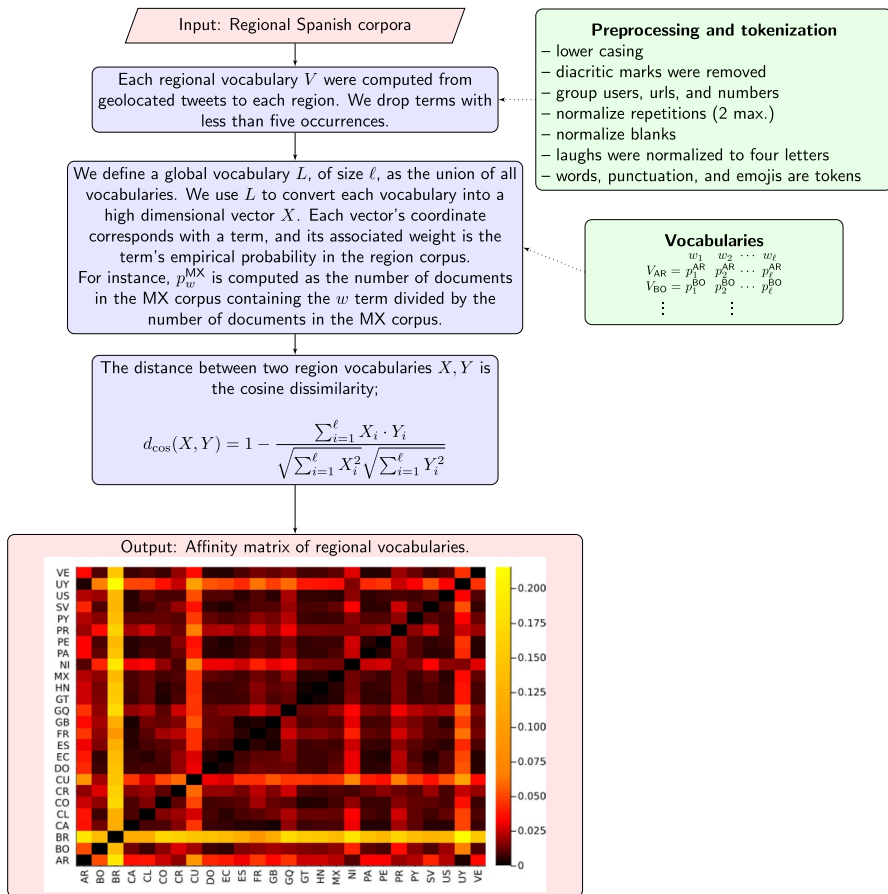


Fig. 4 Affinity matrix among Spanish regions' vocabularies

The median ratio in these countries is 98%. This effect could indicate most of the tweets are produced by residents.

3 Lexical resources

This section describes and analyzes our Spanish Twitter Corpora (STC) in the lexical aspect, specifically from the vocabulary usage perspective. This analysis complements that given of the Heaps' and Zipf's laws and the information given in Table 1.

Fig. 5 Spanish-language lexical similarity visualization among country's vocabularies through a two-dimensional UMAP projection using the Cosine among vocabularies. The points were colorized using a 3D UMAP projection (normalized and interpreted as RGB). Both projections use three nearest neighbors, which emphasizes local features

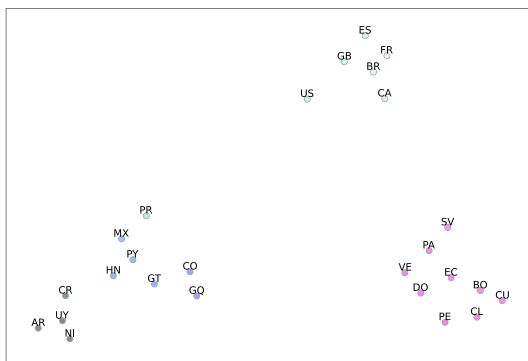


Figure 4 describes the procedure applied to obtain an affinity matrix of our Spanish corpora.⁵ For this purpose, we extracted the vocabulary of each corpus, i.e., a matrix that describes the similarities among corpora. The vocabulary was computed on the entire corpus after text normalizations described in the diagram. We also removed those terms with less than five occurrences in the corpus to remove the tail of the term-frequency distribution, similarly to Mikolov et al. (2013); Bojanowski et al. (2017). The remaining terms are used to create a vector that represents the regional corpus.

The affinity matrix is computed using the cosine distance described in the flow diagram. Note that we select the cosine distance as metric due to the reminiscences of the traditional bag of words with our vocabulary representation, see (Schütze et al., 2008) for more information about conventional vector models for information retrieval. The heatmap represents the actual values in the matrix. This matrix is crucial for the rest of this analysis since it contains distances (dissimilarities) among all pairs of our Spanish corpora. Values close to zero (darker colors) imply that those regions are pretty similar, and lighter ones (close to one) are those regions with higher differences in their vocabularies. For instance, the affinity matrix can show us how Mexico (MX) is more similar to Honduras (HN), Nicaragua (NI), Peru (PE), and the USA (US). This behavior could be the geographical location of the countries, and therefore, a large migration or cultural interchange is made. On the other hand, Brazil (BR) and Equatorial Guinea (GQ) are among the most atypical countries with low similarities with the other countries.

Figure 5 illustrates the similarity between Twitter country vocabularies. Here we rely on Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes & Healy, 2020), a non-linear dimension reduction technique that approximates the k nearest neighbor graph structure of a dataset in the projected low dimension. We applied UMAP projections (2D for spatial projection and 3D for coloring

⁵ Under our context, an affinity matrix is a pair-wise matrix of distances among different regions using a dissimilarity function. The i th row contains the distance of the i region vs. all regions; it has a zero diagonal.

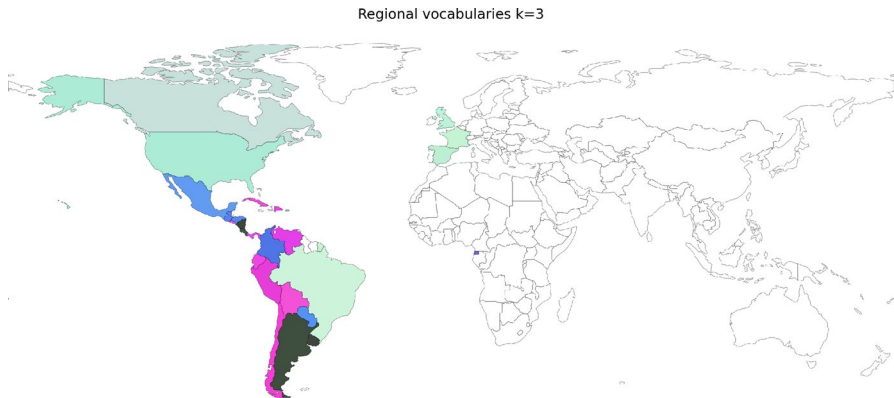


Fig. 6 Regional Vocabulary in RGB representation

points) using the affinity matrix as input. Please recall that the affinity matrix represents similarities between regional vocabularies using the cosine distance as the metric. Please remember that the UMAP algorithm uses the affinity matrix to generate the low-dimensional projection. To our best knowledge, this is a novel approach to visualizing similarities among vocabularies. Still, we can find several uses of dimensional reduction techniques like tSNE (Wada & Iwata, 2018) for visualizing word meaning similarities in a single multi-language model. UMAP is a more recent non-linear dimensional reduction technique that typically performs faster than other non-linear alternatives like tSNE or ISOMAP with remarkable stability on projections (McInnes & Healy, 2020). Interested readers on alternatives are referenced to the recent literature (Anowar et al., 2021).

The figure shows how close or far each Spanish variation is among the entire corpora. UMAP is parameterized by the number of nearest neighbors (knn) in the affinity matrix. The number of neighbors accepts values between $k = 2$ and n , i.e., the number of elements in the collection. Small values of k capture local characteristics of the graph's structure, while large k values capture global structures.

The figure shows the projection using $3nn^6$; we can see four well-defined clusters here. For instance, Uruguay (UY) is very close to Argentina (AR) in three figures, and this is the case in other countries, like Mexico (MX), Colombia (CO), and the United States (US); or Venezuela (VE), and Ecuador (EQ).

While some of these clusters support the idea that geographical similarities imply language similarities, there are notorious exceptions. Figure 6 shows a colorized map using the same colors encoding of Fig. 5. While it is possible to observe similarities and divisions among North America, Central and South America, and European countries, there are essential differences. For instance, Colombia (CO), a South American country, has more similarities to Central American language variants.

⁶ The value $k = 3$ was chosen after several tests, larger values capture more global characteristics, and $k = 2$ produce many sparse clusters (local characteristics).

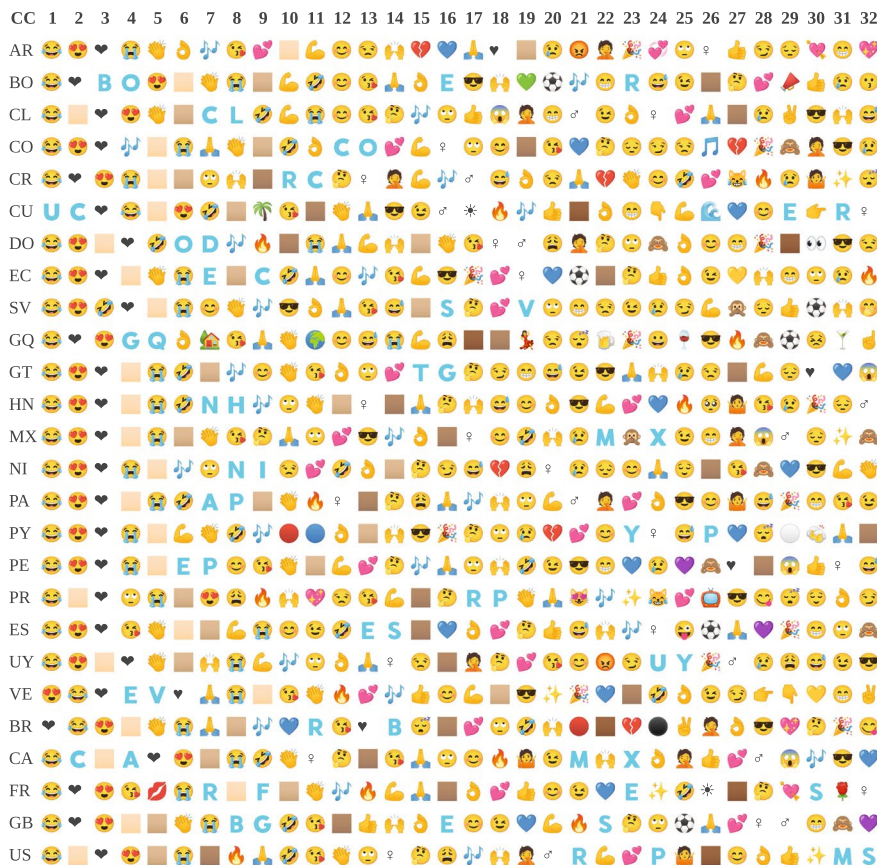


Fig. 7 Most popular emojis per Spanish-speaking country

Regarding our lexical features, Cuba and the Dominican Republic are close to Venezuela, Bolivia, and Ecuador. It is interesting to recall that these similarities are present in Twitter and may vary from other data sources. Still, it could be helpful to take advantage of this knowledge.

Our collections are small for some countries, e.g., GQ, PR, and CU, which can introduce some possible issues in our analysis. For instance, it is possible to declare similarities that are non-meaningful. Please recall that the UMAP projection uses the k nearest neighbor graph (that takes the affinity matrix of Fig. 4 as input). Even when other regions do not select these regions as neighbors (please recall we used $k = 3$), these regions will have direct neighbors that can have enough data. For instance, Fig. 4 shows Cuba with a light row column, which means that most regions are seen relatively far, but it connects Bolivia. Bolivia has more strong connections that positioned Cuba on the map. Another effect occurs with Brazil that even when it has a large corpus, it contains a lot of messages mixing Spanish and Portuguese. It is pretty different from most regions but

visible closer to FR and GB, and therefore, BR will be placed near them on the 2D and 3D projections, see Fig. 5.

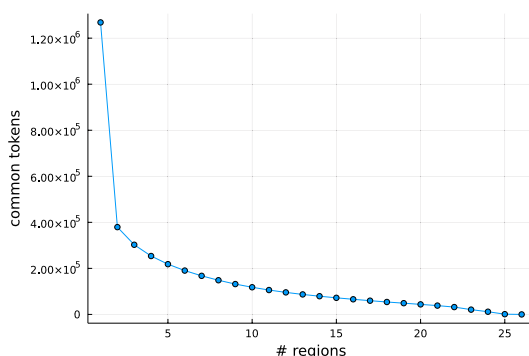
In addition, emojis are graphical symbols expressing an emotion or popular concepts. Hence, they are a lexical resource that can also imply an emotional charge. Emojis compensate for the lack of facial expressions and other expressive ways of face-to-face conversations. Therefore, emojis are popular on social networks like Twitter since they are concise and friendly ways to communicate (Dresner & Herring, 2010). The use of emojis is also dependent on the region, as illustrated in Fig. 7. The figure shows the 32 most used emojis in each country; skin tone markers were separated from composed emojis and counted in an aggregated way. Note that the most popular emojis have consensus in almost all regions. In top rank, we found the *laughing face*, the *in love face*, and the *heart* (love). Another symbol that deserves attention is the color-skin mask, which marks emojis with a skin hue. Regarding frequencies, lighter color-skin marks are more popular than darker ones; this information could have different meanings. For example, users identified as white people, or perhaps it is tricky to select the proper one with Twitter clients. The real reason behind this finding is beyond the scope of this manuscript but deserves attention.

4 Semantic analysis and regional word embeddings

This section discusses the creation of regional semantic representations (word embeddings) for our Spanish language corpora and also analyze similarities between regions using visualization techniques. Word embeddings are vector representations of a vocabulary that capture the semantics of words learning how words are used in a large text corpus. Algorithms learn a high dimensional vector for each token using a distributional hypothesis: words used in similar contexts have similar semantics. Therefore, if two vectors are close, both are semantically related; the contrary also becomes true, as two distant vectors are different semantically. In summary, embeddings are a popular and effective way to capture semantics from a corpus (Yang et al., 2018). There exist several techniques to learn word embeddings, for instance, Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2017; Bojanowski et al., 2017), and Glove (Pennington et al., 2014). Our resources are FastText models, to support out-of-vocabulary words, which are common in social network data. FastText is both a word representation generator and a text classification tool. It is an open-source library well-known for its broad language coverage.⁷ For instance, Grave et al. (2018) trained word embeddings for 157 languages using Wikipedia (800 million tokens) and Common Crawl (70 billion tokens); these models include support for the Spanish language. Nonetheless, there is a lack of country-level language support to our knowledge. Our resources are the first broad effort on this matter, making it possible to take advantage of regionalisms and Spanish dialects.

⁷ <https://fasttext.cc/>.

Fig. 8 Number of common tokens shared by different countries or regions

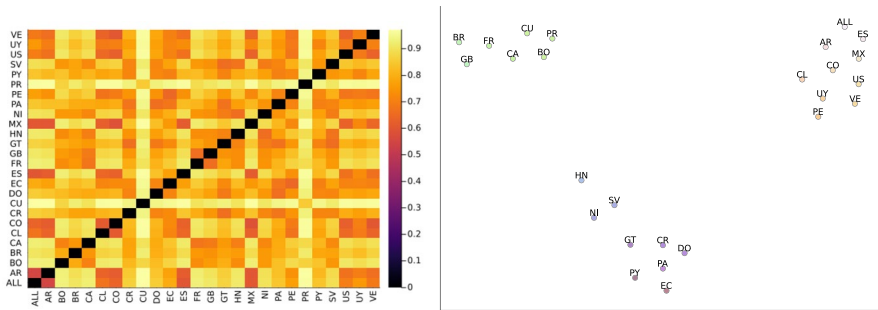


We use our Twitter corpora, divided by country, and apply our preprocessing step, described in Sect. 3, as input of the FastText algorithm. Before, we filtered out several messages: we removed messages with URLs, messages with less than seven tokens, and those produced by applications that use a template to write the tweet (e.g., Foursquare). We also removed retweeted messages. These decisions emphasize that messages contain useful textual information and do not reference external data. The rationale of retweets and external data removal is that they may be of a different user, and we cannot be sure that the linked resource is located in the same region as the original message. These filters reduced our corpora in half (close to 400 million messages).

Regarding the minimum size of seven tokens, the idea is to preserve some context for each token in the message; please recall that word embeddings learn the distributional semantics of each word using its surrounding words. At the same time, we are unaware of a proper study about the phrase's minimum length required to learn word embeddings. However, based on the FastText implementation that uses sliding windows of size five by default as context, preserving messages with at least seven tokens is a tradeoff between maintaining a large dataset and filtering out very short messages.

As commented, we created 26 word-embedding models, one per country, and learned 300 dimension vectors, which is almost a standard for pre-trained embeddings. We used the default values for the rest of the hyper-parameters of FastText. In addition, for comparing purposes, we use the entire corpora as a single corpus to create global word embeddings; the latter is the strategy of most pre-trained word embeddings. This embedding is used to show that regional word embeddings perform differently for regionalized tasks.⁸

⁸ These 27 embeddings are available in <https://ingeotec.github.io/regional-spanish-models/>



(a) Affinity matrix of our semantic representations. (b) Two dimensional UMAP projection of semantic representations.

Fig. 9 Semantic similarities of our Spanish regional word embeddings. Countries are specified in their two letter ISO code. On the left, an affinity matrix where darker cells indicate higher similarities (small distances). On the right a two dimensional UMAP projection, near points indicate similarity

4.1 Word-embedding similarity

Our semantic analysis requires an affinity matrix, as the lexical one given in the previous section. Therefore, we need a representation and a similarity measure to compare word embeddings. Please note that regular word embeddings produced with neural networks will generate vectors that cannot be mixed. Please recall that in the first stage of learning each neural network, its parameters are randomly initialized. An optimizing algorithm is then used to minimize a loss function on the dataset, adjusting parameters and iterating until some objective is achieved. These two procedures, random initialization and optimizing for different datasets, make that two neural network models produce no proximal vectors for the same word, i.e., under the cosine distance. Despite vectors having identical numerical structures, e.g., 300 dimensions, and components showing similar distributions, we cannot evaluate distances between points predicted in different models.

We propose using an intermediate representation and a distance function that captures similarities between these embeddings to measure the similarity between different countries. The core idea is to represent each embedding with a flattened version of the k nearest neighbor graph under a reduced set of tokens, i.e., tokens appearing in most word embeddings. Therefore, the similarity becomes linked to the neighborhood of each word (semantically similar words). The procedure to create this representation is the following:

- Select a common set of tokens; each token appears in at least five countries. This filtering reduces the vocabulary from more than a million tokens to nearly 200 thousand tokens (*vocsiz*). This selection corresponds to an inflection point in the tokens curve, (see Fig. 8). The core idea is to reduce the final representation dimensionality and increase the similarity between related words.
- Our representation requires constructing a k nearest neighbor graph for each country. We use dense vectors of the word embeddings closed to the common

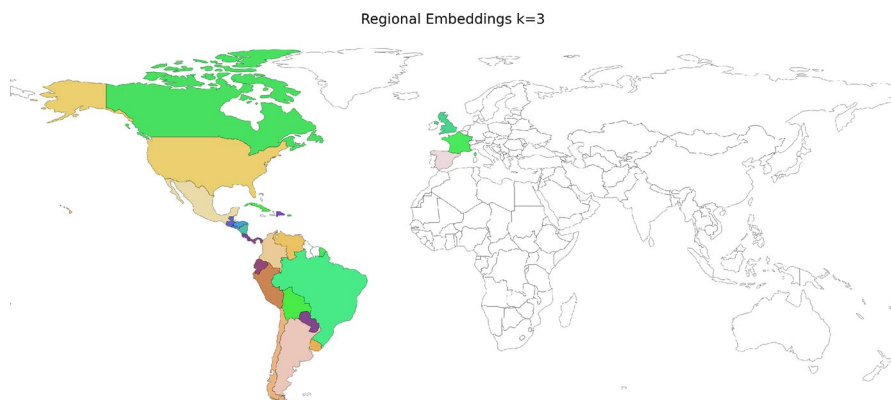


Fig. 10 Geographic visualization of regional embeddings. The 3D UMAP projection is encoded as RGB

tokens set, and we select $k = 33$ after probing several choices. This k value captures several similar terms and remains specific enough to let out different tokens. We used the cosine distance on the dense vectors.

- Finally, each country is represented as a high-dimensional vector that uses k entries per vocabulary word, one per neighbor in the common token set. Each token is then represented by its k nearest neighbors and weighted inversely to its distance.⁹ Note that each word embedding is represented with a very sparse high-dimensional vector, i.e., $vocsize^2$ possible entries, and more than 3.8 million non-zero components.

The set of Spanish embeddings is compared with the cosine distance on the sparse vectors. We computed the affinity matrix shown in Fig. 9a using the procedure described above. As in the previous affinity matrix, darker colors represent a high similarity between the regionalized embeddings and the contrary with lighter colors.

Figure 9b shows a two-dimensional UMAP projection of our semantic representation affinity matrix. The projection uses a 3 nearest neighbors graph as input; please recall that few neighbors capture local structures. The colors are computed by 3D reduction by applying the UMAP dimensional reduction to the same input; the resulting components create an RGB color set using a simple translation and scale procedure to compose values between 0 and 1. Both distances and colors describe a few well-defined groups. Please note that our ALL model is quite different from most points and similar to those regions with extensive collections. This effect is available in word embedding models constructed on non-regional corpora; they learn semantic traits of most represented regions. On the contrary, it is necessary to mention that countries with few messages

⁹ We use the weighting form $0.5 + \frac{1}{1+d(u,v)}$ for embedding vector u and its neighbor's vector v , the lower the distance, the higher the weight.

Table 2 Train distribution of the emoji-15 datasets. Since it is a 50-50 hold-out partition, the test set follows a similar distribution. We removed countries with a low number of examples

emoji	AR	BR	CA	CL	CO	CR	CU	DO	EC	ES	FR	GB	GT	HN	MX	NI	PA	PE	PY	SV	US	UY	VE
😊	3,749	275	158	1,263	4,760	425	26	458	745	2,640	53	95	666	708	10,160	736	827	739	684	278	2,536	446	359
❤️	18,454	1,775	124	2,911	6,237	821	41	348	1,123	8,394	153	255	653	609	9,508	491	880	968	1,800	219	3,216	1,478	784
👉	2,114	92	34	1,016	1,455	242	20	167	253	1,825	41	58	376	157	2,754	159	313	160	666	124	597	190	71
👉	6,785	433	80	2,995	2,486	159	33	369	886	7,061	64	105	197	159	4,175	52	388	577	761	222	1,281	1,248	530
❤️	4,278	100	28	390	1,407	161	7	123	200	1,164	26	37	134	91	1,438	105	260	115	428	31	667	283	142
😊	798	26	8	301	300	42	3	56	83	776	4	18	48	48	549	17	58	115	92	53	192	85	120
😊	3,609	105	35	1,803	1,475	171	32	175	364	3,827	74	56	257	129	3,395	115	236	817	369	192	829	377	293
😊	1,184	70	30	827	986	132	44	280	238	1,017	23	23	123	163	2,011	151	293	248	377	78	608	167	110
😊	15,999	932	111	2,190	6,824	510	28	509	837	8,000	107	136	491	514	8,711	341	1,010	999	1,996	268	2,232	1,194	686
😊	3,081	89	24	920	1,718	155	50	316	291	780	18	22	163	124	2,185	175	322	213	259	127	738	359	236
😊	5,935	211	70	1,764	1,482	98	29	119	347	10,785	203	99	171	72	4,290	63	136	374	190	181	1,808	719	493
😊	2,777	136	60	1,098	1,412	150	5	110	291	1,320	12	25	158	52	2,428	59	155	227	250	90	769	301	252
😊	2,144	89	38	699	1,039	153	8	102	296	1,507	31	39	151	129	2,646	131	204	271	239	69	781	227	135
😊	13,873	436	125	1,967	4,461	581	25	604	787	3,935	157	135	388	364	6,752	321	1,057	979	1,832	200	2,799	939	530
😊	6,751	275	154	2,756	4,173	440	47	614	771	4,781	99	111	421	339	7,380	211	741	1,135	937	384	1,941	951	734

(e.g., CU, PR, and BO) could need more data to support the learning procedure; nonetheless, we decided to maintain them in the projection to learn about their similarity, yet under this advisor. On the other hand, we remove GQ since their small vocabulary produces numerical errors while computing its corresponding k nearest neighbor graph and UMAP projection.

Figure 10 shows the colormap computed from the previous dimensional reduction to colorize a world map obtaining a kind of map of semantic similarity of the Spanish language under our construction characteristics. We can observe how green colors group non-Spanish speaking countries (CA, FR, UK, and BR), except for PR, CU, and BO, in Fig. 9. Note that they correspond to our corpus with fewer messages and that Fig. 3 also indicates a high number of foreign messages. It is necessary to take these results with reservation since can be issues related to their sizes, as explained in Sect. 3 for the same set of collections.

Another large cluster is found with countries of all of Latin America (bottom of Fig. 9). Here we see at least two subclusters with meaningful geographical meaning HN, NI, SV, and GT. The other groups include DO, CR, and PA. Note that EC and PY are also included here. Finally, we found a cluster containing countries around the world. Note that the ALL word embedding is also placed in this cluster. This cluster seems to be composed of countries with larger collections and other countries that are related to them. We found the AR, ES, CL, CO, UY, PE, US, VE, and MX here. Interestingly, we see the US here and not in the green cluster that agglutinates countries not having the Spanish language as an official or *de facto* language.

The semantic similarities between word embeddings can be of interest and can be the object of further research, but their practical usage is also of interest. For instance, it is possible to know what countries can be mixed or interchanged without affecting the regional semantics significantly. A proper topic analysis could help clarify some of these clusters, but it is beyond the scope of this manuscript.

Table 3 Performance statistics of all benchmarks (countries)

Country code	Min acc	Max acc	Local rank	Top-5				
				1	2	3	4	5
AR	0.478	0.490	3	UY	PY	AR	PE	CO
BR	0.461	0.488	1	BR	ALL	DO	PY	CR
CA	0.293	0.353	18	CL	ALL	CO	MX	US
CL	0.426	0.449	1	CL	US	MX	AR	ES
CO	0.425	0.437	2	US	CO	VE	EC	GT
CR	0.369	0.388	9	US	VE	ALL	MX	CO
DO	0.338	0.381	13	US	CO	VE	CL	ALL
EC	0.380	0.414	9	MX	US	CL	ALL	CO
ES	0.475	0.486	1	ES	AR	MX	US	VE
FR	0.419	0.442	4	ALL	GT	EC	FR	PA
GB	0.347	0.376	23	ALL	AR	ES	VE	MX
GT	0.349	0.388	13	MX	US	ALL	CO	ES
HN	0.335	0.367	18	PE	EC	BR	CR	UY
MX	0.423	0.434	1	MX	GT	CR	US	CO
NI	0.337	0.372	18	VE	CO	CL	MX	US
PA	0.366	0.393	10	US	CL	VE	CO	PE
PE	0.380	0.420	9	MX	ALL	US	AR	CO
PY	0.424	0.442	1	PY	US	BR	PE	UY
SV	0.323	0.395	18	US	CO	MX	CL	VE
US	0.404	0.424	1	US	MX	CO	ES	CL
UY	0.435	0.457	1	UY	US	CO	CL	VE
VE	0.385	0.434	4	MX	CO	ES	VE	US

Top-5 models are also listed and the rank position of the local model on solving the current benchmark

4.2 A regional task example: predicting emojis with Emoji-15

Regional information can be used to improve understanding of formal and informal messages, using typical terms and expressions in some regions but not necessarily used in others. Our regional models can help improve some NLP tasks having these characteristics. We introduce the Emoji-15 classification task, a simple multiclass classification problem that predicts the emoji for given messages among 15 possible ones. This task involves identifying emotions and sentiments without a particular topic.

Its creation methodology is simple. We selected 15 popular emojis (see Sect. 3); we do not select the top 15 emojis per region, but a subset that gives some diversity in emotions. Also, we explicitly avoided the most popular emoji and skin tones from our selection. The selected emojis are listed in the first column of Table 2. We selected the datasets for training and test sets from 2020's January and February; therefore, the corpus resources, training, and test sets are disjoint. We ensured that tweets contain at most one of these emojis (even when

Table 4 Average rank of all regional models along all countries datasets

Model	Voc size	Avg rank
US	292,465	4.23
CO	324,635	6.05
MX	438,136	6.27
CL	282,737	6.91
VE	271,924	7.00
ALL	1,696,232	8.45
PE	178,113	8.64
UY	200,032	8.73
EC	147,560	8.95
AR	673,424	9.41
ES	571,196	10.95
PY	124,162	11.14
BR	127,205	11.27
CR	103,086	12.50
PA	111,635	13.36
GT	95,252	13.64
DO	108,655	14.91
GB	82,418	18.00
NI	68,605	18.18
FR	69,843	18.91
CA	63,161	19.00
SV	73,833	19.14
HN	60,580	20.36

Models with low average ranks are better

they can have other emojis). Messages were also selected to be geotagged to one of our objective Spanish-speaking countries. We followed the same filtering procedure and preprocessing as made for the word embedding; note that we also masked emoji's occurrences. That emoji was used as a label for the classification task.

We obtained a number of examples that were divided into a 50-50 holdout (proportion of label messages remain similar in train and test set); see Table 2 for more details. We removed four countries (BO, CU, GQ, and PR) from this task due to the low number of retrieved messages. For instance, we kept the statistics of Cuba in the table to show the lower limit cutting. The idea is to solve all-region benchmarks with all-region models and quantify their performance and the pertinence of local models on local tasks.

The train partition was used to create one model per country and one for the entire set of messages (called ALL). Table 3 shows the accuracy performance scores of all models vs. all test databases. We can observe that some regions are more challenging to predict than others, e.g., CA achieves a maximum score of 0.35 while AR achieves 0.49. One can observe that most accuracy scores are low

emoji	AR	BR	CA	CL	CO	CR	CU	DO	EC	ES	FR	GB	GT	HN	MX	NI	PA	PE	PY	SV	US	UY	VE
😊	3,749	275	158	1,263	4,760	425	26	458	745	2,640	53	95	666	708	10,160	736	827	739	684	278	2,536	446	359
💖	18,454	1,775	124	2,911	6,237	821	41	348	1,123	8,394	153	255	653	609	9,508	491	880	968	1,800	219	3,216	1,478	784
👉	2,114	92	34	1,016	1,455	242	20	167	253	1,825	41	58	376	157	2,754	159	313	160	666	124	597	190	71
👉👉	6,785	433	80	2,995	2,486	159	33	369	886	7,061	64	105	197	159	4,175	52	388	577	761	222	1,281	1,248	530
💖💖	4,278	100	28	390	1,407	161	7	123	200	1,164	26	37	134	91	1,438	105	260	115	428	31	667	283	142
😊😊	798	26	8	301	300	42	3	56	83	776	4	18	48	48	549	17	58	115	92	53	192	85	120
😊😊😊	3,609	105	35	1,803	1,475	171	32	175	364	3,827	74	56	257	129	3,395	115	236	817	369	192	829	377	293
😊😊😊😊	1,184	70	30	827	986	132	44	280	238	1,017	23	23	123	163	2,011	151	293	248	377	78	608	167	110
😊😊😊😊😊	15,999	932	111	2,190	6,824	510	28	509	837	8,000	107	136	491	514	8,711	341	1,010	999	1,996	268	2,232	1,194	686
😊😊😊😊😊😊	3,081	89	24	920	1,718	155	50	316	291	780	18	22	163	124	2,185	175	322	213	259	127	738	359	236
😊😊😊😊😊😊😊	5,935	211	70	1,764	1,482	98	29	119	347	10,785	203	99	171	72	4,290	63	136	374	190	181	1,808	719	493
😊😊😊😊😊😊😊😊	2,777	136	60	1,098	1,412	150	5	110	291	1,320	12	25	158	52	2,428	59	155	227	250	90	769	301	252
😊😊😊😊😊😊😊😊😊	2,144	89	38	699	1,039	153	8	102	296	1,507	31	39	151	129	2,646	131	204	271	239	69	781	227	135
😊😊😊😊😊😊😊😊😊😊	13,873	436	125	1,967	4,461	581	25	604	787	3,935	157	135	388	364	6,752	321	1,057	979	1,832	200	2,799	939	530
😊😊😊😊😊😊😊😊😊😊😊	6,751	275	154	2,756	4,173	440	47	614	771	4,781	99	111	421	339	7,380	211	741	1,135	937	384	1,941	951	734

Fig. 11 Loss and accuracy during training on the Masked Language Model task. The batch size is of 128 tweets

but far from a uniform distribution (15 classes). The table shows the local model's position in each country benchmark (*local rank* column). The best-performing model for a benchmark will rank as 1, the second-best as 2, and similarly for the rest. Note that small local rank values indicate that the local model (for that region) is efficient for its corresponding benchmark. The best five models for each country benchmark are also listed; we can observe how many geographically near regions perform well in their geographic neighborhoods. The average rank of the local model is 8.09 while the median is 6.5; these values support the idea that local models are useful on tasks where regional information can be used. Also, one can observe that not always more data (ALL model) is the best, in this case, it could be said that the geographical aspect is more relevant.

The average rank of a single model along all benchmarks indicates how well this model generalizes. Table 4 shows the performance of all models, along with all benchmarks, as its average rank. We can observe that some country models are outstanding, like the US model. In this sense, it is remarkable that models like the US or CO (both using a vocabulary of 300k tokens) perform better than huge ones. On the other hand, the ALL model is competitive; however, it is not the best (global 6th regarding average rank). Please recall that we created the ALL model by merging the entire corpora into a single corpus, which is the typical construction; for instance, the ALL model contains close to 1.7 million tokens in its vocabulary.¹⁰

While these results apply to the regional task of predicting the most popular emojis, the evidence points out that local models are competitive options for solving tasks requiring local traits as emoji predictions (see Table 3). Even more, some

¹⁰ Note that our vocabulary in Sect. 3 has more than 1.2 million tokens, and here we mentioned a larger one; this is the vocabulary recognized by the fastText parser. However, both vocabularies were computed using the same corpus. It is similar for other word embeddings listed in Table 4.

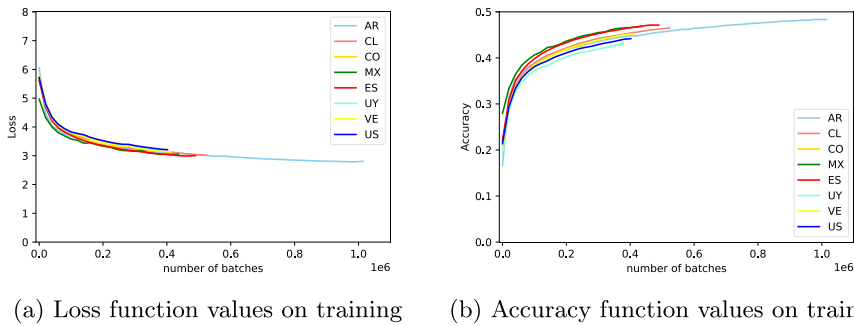


Fig. 12 Comparison of the accuracy of the trained models on all the regions on MLM and emoticon prediction tasks

regional models perform better than large ones, as shown in Table 4, which is remarkable.

5 Language Models

Language Models (LM) are more sophisticated than word embedding models since they go beyond word semantics to context semantics and text generation. In more detail, Word2Vec, FastText, and Glove generate fixed embeddings for each word independently if the word can take different meanings depending on the context. For example, the word *orange* can be a fruit or a color, depending on the context. Language Modeling is the task of predicting the next word given some context so they perform well in distinguishing homonyms.

In that sense, BERT (Devlin et al., 2019) is an LM that has gained considerable attention lately. It is a model that uses a series of encoders to generate embeddings for each word depending on its context. BERT differs from alternatives like ELMo (Peters et al., 2018) because the same pre-trained model can be fine-tuned for different tasks. The pre-train on BERT uses the masked language model (MLM) task where each input sentence contains a *mask token* on 15% random words. Then, BERT was trained on a second task, the next sentence prediction (NSP) task, where the input has two sentences, with a *separation token* in between, and the task was to predict if the second sentence followed the first.

Our resources include regional pre-trained BERT-like models using the MLM task over tweets for the countries AR, CL, CO, MX, ES, UY, VE, and the US, i.e., larger ones. First, we applied the same preprocessing as detailed in Sect. 3 to our corpora. The pre-training was the same as the original BERT, where 15% of the tokens on each sentence were marked with a [MASK] token, and the model must predict them. We used the corresponding regional tweets from 2016 to 2019 to pre-train each model. All the models had a series of two encoders with four attention heads each and output 512-dimensional embedding vectors. This configuration corresponds with the small-size model following the official BERT implementation

Table 5 Predictions of the masked words over different regions. The color intensity indicates the probability of prediction

el [MASK] subio de precio							
AR	CL	CO	MX	ES	UY	VE	US
dolar	chofer	que	cel	que	video	internet	que
bondi	metro	tiempo	video	movil	que	video	video
que	0	0	uber	tiempo	profe	0	juego
0	que	agua	numero	bus	0	precio	q
auto	bus	se	celular	dia	horoscopo	telefono	0
video	internet	ano	tuit	metro	tema	que	pueblo
autoestima	papa	dia	que	coche	nombre	dolar	arbitro
colectivo	mio	man	telefono	agua	pibe	queso	tipo
tiempo	precio	mundo	dinero	cafe	grupo	pan	no
tren	profe	celular	tiempo	pelo	amor	tlf	mundo
la [MASK] subio de precio							
AR	CL	CO	MX	ES	UY	VE	US
foto	wea	gente	foto	gente	foto	gente	gente
que	gente	vida	vida	que	vida	plata	que
lluvia	micro	noche	gasolina	foto	gente	semana	foto
luna	foto	que	cancion	vida	historia	caja	prensa
musica	mama	ley	gente	noche	profe	foto	policia
caja	luna	historia	noche	camara	abuela	carne	tipa
gente	mina	semana	ropa	cara	cancion	vida	camara
camara	senora	lluvia	app	bateria	que	cola	ley
coca	profe	luna	pagina	musica	vieja	arepa	cara
heladera	vieja	musica	morra	semana	madre	pasta	0
me gusta tomar [MASK] en la manana							
AR	CL	CO	MX	ES	UY	VE	US
mates	desayuno	fotos	cafe	cafe	mates	cafe	cafe
teres	once	cafe	fotos	algo	mate	fotos	decisiones
cafe	cafe	cerveza	decisiones	nota	algo	ron	ropa
helado	agua	agua	agua	decisiones	agua	clases	clases
sol	clases	peliculas	alcohol	cerveza	cafe	decisiones	manana
fernet	sol	decisiones	cerveza	cola	sol	hoy	fotos
mate	almuerzo	hoy	clases	sol	vino	anis	sol
birra	cerveza	ropa	hoy	todo	alcohol	cartas	agua
algo	hoy	frio	tequila	uno	helado	agua	hoy
birras	helado	musica	comida	alcohol	cerveza	0	comida
vamos a comer [MASK]							
AR	CL	CO	MX	ES	UY	VE	US
asado	sushi	mierda	tacos	mierda	algo	pizza	=
pizza	todo	jaja	mucho	ya	pizza	mierda	hoy
algo	hoy	!	jaja	mas	hoy	manana	connmigo
helado	mierda	rico	pizza	esto	jaja	rico	pizza
noquis	algo	.	0	mucho	helado	hamburguesa	manana
pizzas	pizza	=	manana	jaja	oreja	hoy	todo
hoy	!	helado	hoy	hoy	todo	arepa	mierda
empanadas	jaja	pizza	taquitos	tio	nada	.	bien
facturas	ctm	hoy	emo	usr	uno	todo	playa
milanesas	emo	asi	algo	bien	eso	torta	tacos
estoy en la ciudad de [MASK]							
AR	CL	CO	MX	ES	UY	VE	US
mierda	santiago	bogota	mexico	madrid	mierda	venezuela	=
argentina	vina	colombia	monterrey	espana	casa	caracas	mexico
cordoba	conce	cali	guadalajara	verdad	uruguay	valencia	dios
rosario	chile	medellin	puebla	sevilla	hoy	merida	miami
mierdaa	valparaiso	barranquilla	cancun	barcelona	historia	barquisimeto	hoy
tucuman	stgo	dios	veracruz	mierda	nuevo	maracaibo	casa
hoy	concepcion	hoy	mty	hoy	filosofia	maracay	vacaciones
capital	valpo	antioquia	merida	valencia	ingles	margarita	pr
aca	maipu	cartagena	cdmx	futbol	montevideo	aragua	disney
sol	chillan	paz	todos	exámenes	clase	carabobo	mi

setup. We chose this setup based on the computational resources we had available. We name our model BILMA, for Bert In Latin America. We used a learning rate of 10^{-5} with the Adam optimizer; the models for CL, UY, VE, and the US were trained

for three epochs and AR, CO, MX, and ES for just one because of the size of their corpus. All the pre-trained models are available for download.¹¹

Figure 11 shows the loss and accuracy of the MLM task during the training. We can see that the BILMA model for AR was trained on double the number of batches; that was because Argentina has double the corpus size. The rest of the models were trained on a similar number of batches.

In Fig. 12a, we compare the models predicting the masked words on all the regions over the test set of tweets used in Sect. 4.2. The test was to predict the [MASK] tokens correctly. Some interesting points to highlight are the following. First, the Argentina model got very high scores on all the regions, even above their corresponding models for UY and CO. This might be because this model was trained for like double the data. Second, some models got better results in the ES region than theirs, like CL, CO, UY, VE, and the US. The US region got the worst outcomes for AR, CL, CO, ES, and UY models. Finally, CO and UY were the models with lower accuracy.

5.1 BILMA's performance on the Emoji-15 regional task

We applied our BILMA models to our Emoji-15 task (see Sect. 4.2). For this matter, we fine-tuned the pre-trained language models to predict the emoticon by adding two linear layers to the first token of each sentence (the start-of-sentence token), so the output of the fine-tuned models was a probability distribution of the assigned emoticon, independent of its position. We split the tweets into 90% train and 10% validation and trained until the accuracy stalled. After that, we evaluated the test set; the results are presented in Fig. 12b. We can conclude that all the models got better results in their corresponding regions from the results. The AR, MX, and ES models got good results over all the regions; meanwhile, UY and VE got low scores. The prediction scores are similar to those found in Table 3; however, our regional FastText models are slightly better than our fine-tuned BILMA models. Nonetheless, BILMA models learn how people write in different regions, as it is exemplified in the rest of this section.

5.2 Generating text with BILMA regional language models

As a qualitative and exemplification exercise, we present how each region model predicts the masked word for the same example phrase. In Table 5 we show the predictions for the masked token on a set of selected sentences. The color intensity indicates the confidence of the model to predict the word. The first two examples are *el/la [MASK] subio de precio* (the [MASK] raised in price),¹² here we can

¹¹ <https://ingeotec.github.io/regional-spanish-models/>.

¹² The article *el* indicates the masked word should be singular and masculine, and for *la* it should be singular feminine.

see differences in how each region name their public transportation, in AR they use *bondi*, *colectivo*, in CL *metro*, *micro*, *bus*, in MX *uber* and ES *metro*, *bus*; we can also see the differences in how they called the cellphone service, in CO and MX is *celular* and in ES is *movil*. The third example is *me gusta tomar [MASK] en la mañana* (I like to drink [MASK] in the morning)¹³, here we can note that in AR and UY people prefer to drink *mates* meanwhile in MX, ES, VE, and the US drink coffee. The fourth phrase is *vamos a comer [MASK]* (let us eat [MASK]) where we can see the differences in the cuisine of the countries with dishes like *asado*, *pizza*, *ñoquis*, *empanadas*, *milanesas*, *sushi*, *tacos*, *oreja*, *hamburguesa*, *arepa*, *torta*. The last sentence is *estoy en la ciudad de [MASK]* (I am in [MASK] city). The results include a list of some of the larger cities in each region. This exercise is a proof of concept to show that the models of different regions can predict very different words, i.e., regional information. Note that the diversity of the predictions include dialect differences (*celular* vs *movil*) but also topical (*tacos* vs *asado*) and will depend on the input sentence.

6 Conclusions

This manuscript proposes a set of regionalized resources for the Spanish language using Twitter as the data source. We collected messages from Twitter's public streaming API from 2016 to 2019; messages must be tagged as being written in Spanish and geotagged to one of the 26 countries that use Spanish as one of their primary languages. The vocabulary of each corpus was extracted, characterized, and compared their similarity, defining a distance metric between them. We also produce visualizations and insightful information about lexical and semantical similarities of the Spanish language variations in Twitter messages.

On the other hand, we created regional semantic models using FastText and produced some visualizations of the semantic similarities among regions. We also create regional language models called BILMA, based on the well-known BERT transformer architecture. We give empirical evidence of the usefulness of regional models in regionalized text classification tasks (Emoji-15 task) and how this more careful data segmentation can yield better performances than the typical more-data-is-better approach.

We provide access to our vocabularies, word embeddings, language models, and corpora sample through the project site (available in <https://ingeotec.github.io/regional-spanish-models/>). The necessary packages (BILMA) and all scripts used to generate our resources, open-sourced under the MIT license, are also reachable under the same site.

¹³ *tomar* could mean to take or to drink.

6.1 Limitations and further research

While the regional models seem promising tools for many tasks that require understanding regionalisms and idiosyncrasies, the use of multiple models can be cumbersome for real-world systems, not to mention the necessary computing units needed to handle many models. It is necessary to create models that can *shift* their region depending on a *regional context*. This approach requires further research.

Our Spanish corpora could be more balanced concerning countries. Some countries have too many elements, while others like GQ, CU, PR, and BO need to be bigger to have reliable semantic models (i.e., word embeddings and language models). More research and data collection are needed to improve resources in these regions.

Our region similarity comparisons are based on lexical and semantic properties of vocabularies computed and learned from Twitter messages. While it is not our goal, the presented projections could compare topics and other internal knowledge in the resources. Proper topic analysis is beyond the scope of this study and requires further research. Similarly, it is possible to mine our resources, i.e., language models, semantic, lexical features, corpus, etc., to perform data-driven social sciences research, i.e., research about gender inequality or race perception. All these topics require more research.

The methodology and implementation of this manuscript are open to improvements. For instance, countries with relatively few examples require different strategies to be competitive. It is also essential to find ways to collect better data and discard the bad ones, such that results become more reliable for small collections. In the same sense, comparing vocabularies and embeddings created from datasets with such disparate sizes require robust normalization methods that we barely sketched. Our long-term goal is to update our resources using more and more data and novel language models as they appear in the literature.

Appendix

BILMA language model usage

In order to use our BILMA models, we need to download one first, we will also need the vocabulary file.

To clone the repository, download the model and install dependencies, in a linux terminal just type the following commands:

```
git clone https://github.com/msubrayada/bilma
cd bilma
bash download-emoji115-bilma.sh
python3 -m pip install tensorflow==2.4
```

and now we have the python package and its dependencies, the model and the vocabulary file (shared to all BILMA models). In particular, this example downloads the MX model that was trained with one epoch on the MLM task and fine-tuned on the Emoji-15 task for 13 epochs.

We need to run a Python 3 console and load the BILMA model.

```
from bilma import bilma_model
vocab_file = "vocab_file_All.txt"
model_file = "bilma_small_MX_epoch-1_classification_epochs-13.h5"
model = bilma_model.load(model_file)
tokenizer = bilma_model.tokenizer(vocab_file=vocab_file,
                                  max_length=280)
```

this BILMA model has two outputs, the first with shape $(bs, 280, 29025)$ where bs is the batch size, 280 is the max length and 29025 is the size of the vocabulary. This output is used to predict the masked words. The second output has shape $(bs, 15)$ which corresponds to the predicted emoji.

The next step is tokenizing some messages as follows:

```
texts = [
    "Tenemos tres dias sin internet ni senal de celular en el pueblo.",
    "Incomunicados en el siglo XXI tampoco hay servicio de telefonía",
    "fija",
    "Vamos a comer unos tacos",
    "Los del banco no dejan de llamarme"
]
toks = tokenizer.tokenize(texts)
```

the prediction is made as follows:

```
p = model.predict(toks)
```

finally, the predicted emojis can be displayed with:

```
tokenizer.decode_emo(p[1])
```

this produces the output: $['😞', '😞', '😞', '😡']$, each emoji corresponds to the most probable one for each message in `texts`.

Cut off N

This section presents a methodology that addresses the minimum token frequency to be kept in the analysis.

The idea is to compute the confidence interval of a Bernoulli variable and select the minimum frequency f (number of times the token appears in the corpus) that sets the interval in a feasible region. Let p be the probability of seeing a particular token, assuming \hat{p} is Gaussian distributed. The confidence interval is $\hat{p} \pm \alpha \text{se}(\hat{p})$, where se is the standard error of \hat{p} , and α is the percent point function with parameter $1 - \frac{c}{2}$, where $1 - c$ represents the confidence, e.g., $\alpha \approx 2$ gives approximately a 95% confidence interval.

The following equations show that under the assumption made, the frequency f (number of times the token appears in the corpus) must be greater or equal to $\frac{N\alpha^2}{N+\alpha^2}$ that in the limit when N tends to infinity corresponds to α^2 .

$$\hat{p} - \alpha \text{se}(\hat{p}) \geq 0 \quad (1)$$

$$\hat{p} - \alpha \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{N}} \geq 0 \quad (2)$$

$$\sqrt{N}\hat{p} - \alpha\sqrt{\hat{p}(1-\hat{p})} \geq 0 \quad (3)$$

$$\sqrt{N}\hat{p} \geq \alpha\sqrt{\hat{p}(1-\hat{p})} \quad (4)$$

$$\sqrt{N}\hat{p} \geq \sqrt{\alpha^2\hat{p}(1-\hat{p})} \quad (5)$$

$$N\hat{p}^2 \geq \alpha^2\hat{p}(1-\hat{p}) \quad (6)$$

$$N\hat{p}^2 - \alpha^2\hat{p}(1-\hat{p}) \geq 0 \quad (7)$$

$$\hat{p}(N\hat{p} - \alpha^2(1-\hat{p})) \geq 0 \quad (8)$$

$$N\hat{p} - \alpha^2(1-\hat{p}) \geq 0 \quad (9)$$

$$N\hat{p} - \alpha^2 + \alpha^2\hat{p} \geq 0 \quad (10)$$

$$\hat{p}(N + \alpha^2) \geq \alpha^2 \quad (11)$$

$$\hat{p} \geq \frac{\alpha^2}{N + \alpha^2} \quad (12)$$

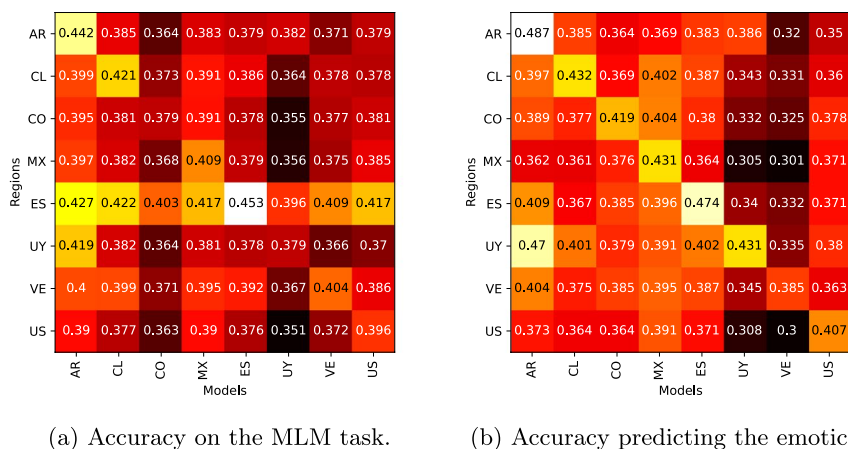


Fig. 13 Two-dimensional UMAP projections of regional vocabularies (left side) and word embeddings (right side) removing CU, PR, GQ, and BR (regarding Figs. 5 and 9). Colors also capture similarity using a 3D UMAP projection of the same data. Nonetheless, the similarity between figures is undefined

$$f \geq \frac{N\alpha^2}{N + \alpha^2} \quad (13)$$

Removing regions with relatively small datasets

A possible issue to attend to in our work is the need for more data in some regions, e.g., Equatorial Guinea (GQ), Cuba (CU), and Puerto Rico (PR). Another possible source of problems is the mix of languages in the collection, like Brazilian (BR) Portuguese; please recall that we use the country label provided by Twitter and that Spanish and Portuguese languages have significant lexical similarities that can introduce errors in automatic detectors. This appendix studies the effect of removing these regions in the similarity analysis presented in Sects. 3 and 4.

To analyze if these issues have an impact on our results and analysis, we proceeded to redo a section of our experiments, removing three small data countries, Cuba, Puerto Rico, and Equatorial Guinea. We also remove the Brazil corpus, which does not identify as Spanish-speaking but also shows essential divergences in its vocabulary similarities, see Fig. 4.

With the reduced corpora, we compute the UMAP-based visualization on both lexical and semantic representations; see Figs. 5 and 9 as original ones.

As the results of these new computations, Fig. 13a and b show that removing low-represented regions and others like BR does not produce a significant change in the structure of the projections, i.e., the close regions are almost preserved as well the clusters of regions. For instance, the lexical visualization kept almost all cluster structures during semantic projection. However, it transfers Honduras (HN)

from one group to another. Still, the HN's color remains close to those elements in its previous cluster, i.e., with NI (Nicaragua) and SV (El Salvador). The lack of enough data for some regions barely changes the entire perspective of the similarity matrix and the UMAP projection. On the other hand, the atypical vocabulary of BR also has a low impact since it is far from the rest of the regions, and these regions do not select them in their *knn* sets. Therefore, its removal barely affects the final projection.

Acknowledgements This work has been done through CONACYT (National Council of Science and Technology from Mexico) support with the Ciencia Básica grant with project ID A1-S-34811. Also, the authors acknowledge the support from CIMAT and “Laboratorio de Supercomputo del Bajío” through project 300832 from CONACyT. We also thank the reviewers’ comments and suggestions that help us improve the manuscript’s quality.

References

- Alshutayri, A., & Atwell, E. (2017). Exploring Twitter as a source of an Arabic dialect corpus. *International Journal Of Computational Linguistics (IJCL)*, 8, 37–44.
- Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, 40, 100378.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions Of The Association For Computational Linguistics*, 5, 135–146.
- CKennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaladar, S., Portillo-Wightman, G., Gonzalez, E., & Hoover, J. (2022). Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. In *Language Resources and Evaluation*. Springer.
- Cotton, E., & Sharp, J. (1988). *Spanish in the Americas*. Georgetown University Press.
- Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating “big data” and leveraging the potential of the Geoweb. *Cartography and Geographic Information Science*, 40(2), 130–139.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)*, (pp. 4171–4186). Association for Computational Linguistics.
- Donoso, G., & David S. (2017). Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, (pp. 16–25). Association for Computational Linguistics.
- Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, 20(3), 249–268.
- Escudero-Mancebo, D., Corrales-Astorgano, M., Cardeñoso-Payo, V., Aguilar, L., González-Ferreras, C., Martínez-Castilla, P., & Flores-Lucas, V. (2022). *Prautocal corpus: A corpus for the study of down syndrome prosodic aspects* *Language Resources and Evaluation*. Springer.
- Finfgeld-Connett, D. (2015). Twitter and health science research. *Western Journal of Nursing Research*, 37, 1269–1283.
- Frenda, S., Ghanem, B., Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36, 4743–4752.
- Gelbukh, A. & Sidorov, G. Zipf (2001) Heaps Laws’ Coefficients Depend on Language. In *Computational Linguistics And Intelligent Text Processing* (pp. 332–335).
- Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing dialect characterization through twitter. *PLoS ONE*, 9(11), e112074.


- Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in twitter. *The Professional Geographer*, 66(4), 568–578.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Gruszczyński, W., Łodzimierz, Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., & Woliński, M. (2022). *The Electronic Corpus of 17th-and 18th-century Polish Texts Language Resources and Evaluation*. Springer.
- Hoff, M. (2020). Cerca mía/o or cerca de mí? A variationist analysis of Spanish locative+ possessive on Twitter. *Studies in Hispanic and Lusophone Linguistics*, 13, 51–78.
- Hong, L., Convertino, G., & Chi, E. (2011). Language matters in twitter: A large scale study. *Proceedings Of The International AAAI Conference On Web And Social Media*, 5, 518–521.
- Hovy, D., Rahimi, A., Baldwin, T., & Brooke, J. (2020). Visualizing regional language variation across Europe on Twitter. In S. Brunn & R. Kehrein (Eds.), *Handbook of the changing world language map* (pp. 3719–3742). Springer.
- Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59, 244–255.
- Huertas-Tato, J., Martin, A., & Camacho, D. (2022). BERTuit: Understanding Spanish language in Twitter through a native transformer. <http://arXiv.org/2204.03465>
- Jimenez, S., Dueñas, G., Gelbukh, A., Rodriguez-Diaz, C., & Mancera, S. (2018) Automatic detection of regional words for pan-hispanic spanish on twitter. In: *Ibero-American Conference On Artificial Intelligence* (pp. 404–416).
- Joulin, A., Edouard, G., Piotr, B., & Tomas, M. (2017). 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*, (pp. 427–431). Association for Computational Linguistics.
- Kejriwal, M., Wang, Q., Li, H., & Wang, L. (2021). An empirical study of emoji usage on twitter in linguistic and national contexts. *Online Social Networks and Media*, 24, 100149.
- Lamontagne, J., & McCulloch, G. (2022). Phonological variation on Twitter: Evidence from letter repetition in three French dialects. *Journal of French Language Studies*, 32, 165.
- Li, M., Chng, E., Chong, A., & See, S. (2019). An empirical analysis of emoji usage on Twitter. *Industrial Management & Data Systems*, 119, 1748.
- McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.
- Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4), e61981.
- Mooney, P., Winstanley, A., & Corcoran, P. (2009). *Evaluating Twitter for use in environmental awareness campaigns*. Department of Computer Science: National University of Ireland, Maynooth.
- Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. *Proceedings Of The International AAAI Conference On Web And Social Media*, 7, 466–475.
- Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Proceedings of The International AAAI Conference on Web And Social Media*, 5, 265–272.
- Pennington, J., Richard, S., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).
- Penny, R., Penny, R., & Ralph, P. (2000). *Variation and change in Spanish*. Cambridge University Press.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (pp. 2227–2237). Association for Computational Linguistics.
- Rodriguez-Diaz, C. A., Jimenez, S., Dueñas, G., Bonilla, J. E., & Gelbukh, A. (2018). Dialectones: Finding statistically significant dialectal boundaries using twitter data. *Computación y Sistemas*, 22(4), 1213–1222.
- Schütze, H., Manning, C., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

- Suhasini, M., & Srinivasu, B. (2020). Emotion detection framework for twitter data using supervised classifiers. In *Data Engineering And Communication Technology*, (pp. 565–576).
- Vashisth, P., & Meehan, K. (2020). Gender classification using twitter text data. In *2020 31st Irish Signals And Systems Conference (ISSC)*, (pp. 1–6).
- Wada, T. & Iwata, T. (2018) Unsupervised cross-lingual word embedding by multilingual neural language models. CoRR. <https://arXiv.org/1809.02306>
- Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in Twitter election classification. *Information Retrieval Journal*, 21(2), 183–207.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Eric S. Tellez^{1,2,5} · Daniela Moctezuma³  · Sabino Miranda^{1,2,4} · Mario Graff^{1,2} · Guillermo Ruiz^{1,3}

Eric S. Tellez
eric.tellez@infotec.mx

Sabino Miranda
sabino.miranda@infotec.mx

Mario Graff
mario.graff@infotec.mx

Guillermo Ruiz
lgruiz@centrogeo.edu.mx

¹ Conacyt, Consejo Nacional de Ciencia y Tecnología., Av. Insurgentes Sur 1582, Col. Crédito Constructor., 03940 CDMX, Mexico

² INFOTEC, Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopol Norte, No.112 Col. Tecnopol Pocitos II, 20326 Aguascalientes, Aguascalientes, Mexico

³ CentroGEO, Centro de Investigación en Ciencias de Información Geoespacial., Circuito Tecnopol Norte, No.107 Col. Tecnopol Pocitos II, 20313 Aguascalientes, Aguascalientes, Mexico

⁴ UPIITA-IPN, Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2580 Col. Barrio la Laguna Ticomàn, Gustavo A. Madero, 07360 Mexico City, Mexico

⁵ CICESE, Centro de Investigación Científica y de Educación Superior de Ensenada, Carr. Tijuana-Ensenada, No.3918, Zona Playitas, 22860 Ensenada, Baja California, Mexico