Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 51-62 recibido 08-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Exploiting user-frequency information for mining regionalisms in Argentinian Spanish from Twitter

Explotando información de frecuencia de usuarios para minar regionalismos del español de Argentina en Twitter

Juan Manuel Pérez,[1,2] **Damián E. Aleman,**[1]
Santiago N. Kalinowski,[3] **Agustín Gravano**[4,2]

[1]Universidad de Buenos Aires, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
[3]Academia Argentina de Letras, Buenos Aires, Argentina
[4]Universidad Torcuato Di Tella, Buenos Aires, Argentina

{jmperez,daleman}@dc.uba.ar, s.kalinowski@aal.edu.ar, agravano@utdt.edu

Abstract: The task of detecting regionalisms (expressions or words used in certain regions) has traditionally relied on the use of questionnaires and surveys, heavily depending on the expertise and intuition of the surveyor. The emergence of social media and microblogging services has produced an unprecedented wealth of content (mainly informal text generated by users), opening new opportunities for linguists to extend their studies of language variation. Previous work on the automatic detection of regionalisms depended mostly on word frequencies. In this work, we present a novel metric based on Information Theory that incorporates user frequency. We tested this metric on a corpus of Argentinian Spanish tweets in two ways: via manual annotation of the relevance of the retrieved terms, and also as a feature selection method for geolocation of users. In either case, our metric outperformed other techniques based on word frequency, suggesting that measuring the amount of users that use a word is an informative feature. This tool has helped lexicographers discover several unregistered words of Argentinian Spanish, as well as different meanings assigned to registered words.
Keywords: Lexical dialectology, Social media, Spanish variants, Entropy.

Resumen: La tarea de detectar regionalismos (expresiones o palabras utilizadas en determinadas regiones) se ha basado tradicionalmente en el uso de cuestionarios y encuestas, dependiendo en gran medida de la pericia e intuición del investigador. El surgimiento de las redes sociales y los servicios de microblogging ha producido una riqueza de contenido sin precedentes (principalmente textos informales generados por usuarios), lo cual ha abierto nuevas oportunidades para el estudio de la variación lingüística. Estudios previos de la detección automática de regionalismos dependen sobre todo de la frecuencia de palabras. En este trabajo presentamos una métrica novedosa basada en la Teoría de la Información, que incorpora la frecuencia de usuarios. Ponemos a prueba esta métrica en un corpus de Tweets en español argentino de dos maneras: a través de la anotación manual de la relevancia de los términos recuperados, y también usándola como un método de selección de características para la geolocalización automática de usuarios. En ambos casos, nuestra métrica superó otras técnicas basadas en la frecuencia de palabras, lo que sugiere que medir la cantidad de usuarios que usan una palabra es una característica informativa. Esta herramienta ha ayudado a lexicógrafos a descubrir varias palabras no registradas del español argentino, así como significados nuevos de palabras ya registradas.
Palabras clave: Dialectología léxica, Redes sociales, Variantes del español, Entropía.

ISSN 1135-5948. DOI 10.26342/2022-69-4

1 Introduction

Lexicography has been aided and enriched in the past 30 years by tools and resources from Computational Linguistics, mainly in the form of corpora of selected texts (Atkins and Rundell, 2008). Statistical analyses of corpora usually result in evidence to support the addition of a word to a dictionary, its removal, or its marking as dated or as unused or as regional, among other decisions.

In the process of compiling dictionaries, differences emerge between dialects, where frequently certain words or meanings do not span all speakers. Since languages are ideal constructs based on the observation of dialects, it is of paramount importance to establish which words are likely shared by an entire linguistic community and which are used only by smaller groups. In the latter case, word usage descriptions can profit considerably from information as precise as possible, about geographical extension (region, province, district, city, even neighborhood), registry (colloquial, neutral, formal), frequency (current, past or a combination of both depending on the chronological span of the corpus), and other such variables.

Regionalisms (words used mainly in a particular subregion, such as *che* or *metegol* in Argentinian Spanish[1]) are commonly detected through surveys or transcriptions, using methods that depend more or less on the intuition and expertise of linguists (Almeida and Vidal, 1995; Labov, Ash, and Boberg, 2005). The results of this methodology are of great value to lexicographers, who need evidence to support the addition of a word into a regional dictionary, as well as the indication of where it is used. Information gathered with such methods has been used as lexical variables to compute similarities between dialects (Kessler, 1995; Nerbonne et al., 1996).

The emergence of social media and microblogging services has produced an unprecedented wealth of content, with a clear tendency towards informal or colloquial text generated by users. This fact has opened many opportunities for linguists due to the possibility of accessing geotagged contents, which provide valuable information about the location of users. In this sense, social media texts have been used to aid *lexical dialectology*, for example to establish "continuous" isoglosses (Gonçalves and Sánchez, 2014; Huang et al., 2016) or to study the diffusion of lexical change (Eisenstein et al., 2014), inter alia.

A problem closely related to lexical dialectology is *geolocation*, which maps words into regions or locations (Eisenstein, 2014). A possible way to evaluate dialectological models is to use them in geolocation algorithms; regionalisms can be seen as *location-indicative words* (Han, Cook, and Baldwin, 2012). Most previous work in word-centric geolocation algorithms (and lexical dialectology) relies on the observation of the frequency of a certain word, ignoring the number of users producing them. Also, to our knowledge very little work has been performed in Spanish on these topics.

In this work, we present an information-theoretic measure to detect regionalisms in social media texts, particularly on Twitter, and we test it against a dataset of tweets in Argentinian Spanish. Our contributions are twofold: a) we introduce a new metric based on Information Theory which can be seen as a mixture of *TF-IDF* and Information Gain; and b) we show that measuring the dispersion of users is a strong indicator of relevance, for both lexical dialectology and geolocation. We conduct our experiments on a dataset of tweets in Argentinian Spanish, with 81M tweets, 56K users, all balanced across the country's 23 provinces.

2 Previous Work

Most previous work in lexical dialectology consists in measuring the usage of words that are known a priori to be regional variants. These studies typically use information gathered from sources such as web searches (Grieve, Asnaghi, and Ruette, 2013) and manually-collected regionalisms (Ueda and Ruiz Tinoco, 2003; Kessler, 1995). Even papers that analyze data from Twitter (Huang et al., 2016; Gonçalves and Sánchez, 2014) still rely on words already known for the discovery of dialectal patterns.

Language evolves so quickly that it is important to detect these contrastive words automatically – or at least, to alleviate the efforts needed to detect them. Two types of approaches exist for this problem: *model-based* approaches and *metric-based* approaches (Rahimi, Baldwin, and Cohn,

[1] *Che*: interjection for getting the interlocutor's attention; *metegol*: mechanic game that emulates football (*futbolín*) (Academia Argentina de Letras, 2008).

	Total	Mean	SD
Words	647M	28.14M	6.64M
Tweets	80.9M	3.51M	0.91M
Users	56.2K	2.44K	0.04K
Vocabulary	7.5M	0.32M	0.04M

Table 1: Dataset summary. Total figures, along with province-level means and standard deviations.

2017). Model-based approaches use generative models to detect topics and regional variants (Eisenstein et al., 2010; Ahmed, Hong, and Smola, 2013). Typically, these are computationally expensive, which limits the amount of data that may be processed. Metric-based approaches compute statistics for each word or expression, and use them to create rankings (Cook, Han, and Baldwin, 2014; Chang et al., 2012; Jimenez et al., 2018; Monroe, Colaresi, and Quinn, 2008). These rankings are subsequently evaluated by checking external sources of regionalisms, such as dictionaries. In the following section, we compare our metrics to those proposed by Han, Cook, and Baldwin (2012): Term-Frequency Inverse Location Frequency (TF-ILF) and Information-Gain Ratio.

Text-based geolocation can be seen as the inverse problem of lexical dialectology: while dialectology maps regions into text, geolocation maps text into regions (Eisenstein, 2014). Thus, a reasonable way of assessing the performance of a method for discovering regional words is to use it as a feature-selection method for a geolocation classifier, as proposed by Han, Cook, and Baldwin (2012). In the present work, we use provinces as our unit of study (see Section 3), but finer grained geolocation could be performed by using an adaptive grid (Roller et al., 2012).

Rahimi, Cohn, and Baldwin (2017) propose a different approach to this problem. They train a multilayer perceptron with a bag-of-words as input to geolocate users. Intermediate layers serve as vector representations to perform lexical analysis by analyzing proximities in the embedding space.

Information Theory is the basis for many of these methods (Han, Cook, and Baldwin, 2012; Roller et al., 2012; Chang et al., 2012). Other uses of information theoretic measures include telling whether a hashtag is promoted by spammers by analyzing its dispersion in time and in users (Cui et al., 2012; Ghosh, Surachawala, and Lerman, 2011), and also to discover valuable features from user messages on Twitter for sentiment analysis and opinion mining (Pak and Paroubek, 2010). The metrics discussed in the next section use this concept of measuring the entropy of the users of a particular word.

3 Materials

The territory of Argentina is divided into 23 *provinces* and the autonomous city of Buenos Aires, with populations ranging from 127,000 (Tierra del Fuego Province) to 15 million (Buenos Aires Province), according to the 2010 National Census.[2] Provinces are further subdivided into *departments*, which in some cases are called *partidos* or *comunas*.

To gather our data, we first collected information of all departments in Argentina from the 2010 National Census and conducted a lookup through the Twitter API for users with location matching those departments. Even though location fields in Twitter are not very reliable (Hecht et al., 2011), given that we restrict our search to a fixed number of department names, we observe that most of the potential noise is reduced. We used the Python library *tweepy* to interact with the Twitter API.[3]

For each of the retrieved users, we successfully downloaded their entire tweetlines. Tweets were tokenized using *NLTK* (Bird, Klein, and Loper, 2009). Hashtags and mentions to users were removed; the remaining words were downcased; and identical consecutive vowels were normalized up to three repetitions ("woaaa" instead of "woaaaaaa"). Table 1 summarizes the collected dataset, and Figure 1 shows the distributions of tweets per user and tweet length.

It is well known that the Twitter vocabulary tends to be very noisy with lots of contractions, non-normal spellings (e.g., vocalizations), typos, etc. (Kaufmann and Kalita, 2010). For this reason, we decided to take into account only words occurring more than 40 times and used by more than 25 users (these values were chosen empirically). This removes about 1% of the total words and shrinks the vocabulary from 2.3 million words to around 135,000 words.

[2] `https://www.indec.gov.ar`

[3] `https://www.tweepy.org`

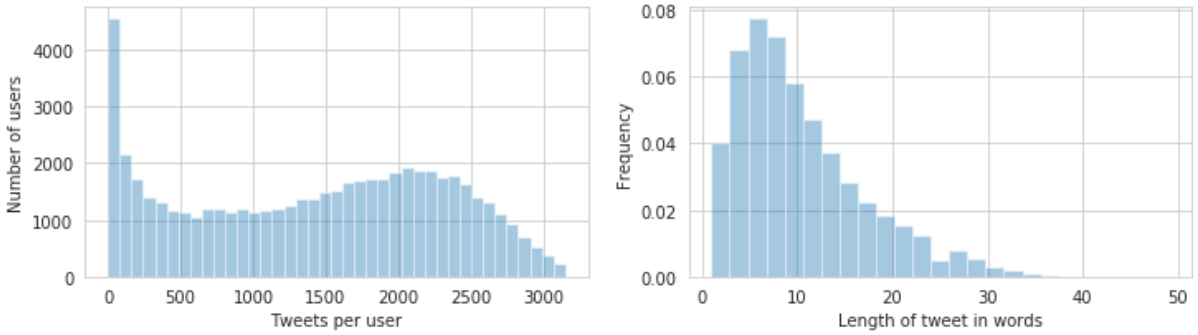


Figure 1: Dataset distributions: Number of tweets per user (left) and words per tweet (right).

4 Method

We can think of a regionalism as a word whose usage is not uniform across the territory – i.e., whose concentration is higher in a specific region. With this in mind, we aim to measure these *disorders* in word usage – or, more precisely, the *entropy* of words (Shannon, 1948).

In general, words with high entropy are more likely to be pronouns, connectors and other closed-class words, whereas their low-entropy counterparts are usually nouns, verbs, adjectives and adverbs with fuller semantic content (Montemurro and Zanette, 2002; Montemurro and Zanette, 2010). Also, words with high entropy (i.e., high disorder) can be regarded as used evenly across the country. On the other hand, low-entropy words are used with higher frequency in a few specific locations.

Let $l_1, l_2, \ldots l_N$ be our locations, and ω_1, ω_2, $\ldots \omega_M$ our vocabulary. If O_j refers to the event of occurrence of word ω_j, then $p(l_i|O_j)$ denotes the probability that word w_j occurred in location l_i.

We next define the *word-count entropy* as

$$H_{\text{words}}(\omega_j) = -\sum_{i=1}^{N} p(l_i|O_j) \cdot \log p(l_i|O_j). \quad (1)$$

Note that this measure does not take into account the actual frequency of words. For instance, if two words ω_1 and ω_2 occur only in one particular location, but ω_1 is much more frequent than ω_2, both words will still have the same entropy according to Equation 1.

In a similar fashion to *tf-idf* and inspired by Montemurro and Zanette (2010) and Han, Cook, and Baldwin (2012), we define measure $I_{\text{words}}(\omega)$ for word ω as follows:

$$I_{\text{words}}(\omega) = p(\omega) \cdot (\log N - H_{\text{words}}(\omega)), \quad (2)$$

where $\log N$ is the maximum possible value of $H_{\text{words}}(\omega)$ (Shannon, 1948), and $p(\omega)$ is the relative frequency of ω in the corpus ($0 \leq p(\omega) \leq 1$). In this way, $I_{\text{words}}(\omega)$ will be high for frequent words that accumulate in just a few locations.

Another important aspect of a word is the amount of people that use it (Cui et al., 2012). Assuming we now sample Twitter users, let U_j be the event that a particular user uses word ω_j. Then $p(l_i|U_j)$ denotes the probability that the location of a user is l_i given the fact that s/he uses word ω_j. We define the *user-count entropy* as

$$H_{\text{users}}(\omega_j) = -\sum_{i=1}^{N} p(l_i|U_j) \cdot \log p(l_i|U_j) \quad (3)$$

and the following metric of ω,

$$I_{\text{users}}(\omega) = q(\omega) \cdot (\log N - H_{\text{users}}(\omega)), \quad (4)$$

where $q(\omega)$ is the proportion of users who mentioned ω in the corpus ($0 \leq q(\omega) \leq 1$). Note that $I_{\text{users}}(\omega)$ will be high for words mentioned by several users who accumulate in just a few locations.

According to Zipf's Law, the counts of the most frequent words are orders of magnitude higher than the counts of the remaining words – a phenomenon that is also true when counting users of words. So the $p(\omega)$ and $q(\omega)$ terms in Equations (2) and (4) become a problem as words with high frequencies overcome their low entropies. To alleviate this, we performed a normalization on the word frequency as follows. Let M_ω be the most frequent word, that is,

$$M_\omega = \arg\max_{\omega \in W} \#\omega, \quad (5)$$

where $\#\omega$ denotes the total number of occurrences of ω in our dataset. Then, the *Nor-*

malized log-frequency of word occurrences is defined as

$$n_{\text{words}}(\omega) = \frac{\log(\#\omega)}{\log(\#M_w)}. \quad (6)$$

Words with very high frequency differ little in their values of $n_{\text{words}}(\omega)$. We define analogously the *Normalized log-frequency* of user mentions n_{users}. Hence, we rewrite Equations (2) and (4) and arrive at the final definition of our two metrics as

$$I_{\text{words}}(\omega) = n_{\text{words}}(\omega)(\log(n) - H_{\text{words}}(\omega)) \quad (7)$$

$$I_{\text{users}}(\omega) = n_{\text{users}}(\omega)(\log(n) - H_{\text{users}}(\omega)) \quad (8)$$

We call the first metric *Log-Term Frequency Information Gain (LTF-IG)* and the second one *Log-User Frequency Information Gain (LUF-IG)*. Summing up, **words with high values of LTF-IG or LUF-IG are cantidates for being regionalisms** – words that occur much more often in a certain region than in the rest of the country.

We subsequently sort all words in our dataset relative to these metrics, thus obtaining two word rankings: *Word-Count Ranking* and *User-Count Ranking*. The words that appear in the first positions of a ranking are those with high values for the metric, and thus more likely to be regionalisms.

4.1 Lexicographic Validation

With these rankings, a team of lexicographers from Academia Argentina de Letras performed a linguistic validation of the first thousand words according to each metric. This qualitative analysis consisted in a detailed study, word by word, to determine if the word in question is part of the lexical repertoire of a community of speakers.

Proper and place names (toponyms) were excluded –as is usual in lexicography– although many words in this class had high values for our metrics. Potential toponyms were automatically highlighted to facilitate their manual exclusion by lexicographers.

To perform the linguistic validation, lexicographers were provided with tables containing counts for each word and province: number of users, number of occurrences and normalized frequency (occurrences per million words). Also, samples of tweets containing these words were provided when necessary. The goal of this manual validation was to identify not only words used exclusively or mainly in a region, but also words used there with a different meaning.

As a result of this process, every word in the top-1000 of each ranking was annotated with '1' if it had lexical relevance as a regionalism, or '0' if it had not. Lastly, lexicographers performed a characterization of the words marked as regionalisms, according to the linguistic phenomenon they represent. The outcome of these procedures is described in Section 5.

4.2 Feature Selection for Geolocation

To indirectly assess the usefulness of our metrics, we used each as a feature-selection method to train geolocation classifiers. This means that, instead of using the entire bag-of-words as input for a geolocation algorithm, we consider a smaller subset of the vocabulary. This dimensionality reduction of the feature space is aimed at boosting the classifier performance.

This approach to geolocation can be described as "word-centric", as it uses lexical information from tweets to predict a location (Zheng, Han, and Sun, 2018). But we emphasize that we are interested in *user* geolocation, not tweet geolocation. Thus, the units considered here are all the tweets from individual users. We randomly selected 10,000 users from our dataset – 7,500 for training and 2,500 for testing.

For reference, we compare our results to those obtained using the *Information Gain Ratio (IGR)* metric (Han, Cook, and Baldwin, 2012; Cook, Han, and Baldwin, 2014): if L is a random variable denoting the location of a given occurrence of word ω_i, then the *Information Gain* of ω_i is

$$\begin{aligned} IG(\omega_i) &= H(L) - H(L|\omega_i) \\ &\propto P(\omega_i) \sum_{j=1}^{m} P(c_j|w_i) \log P(c_j|w_i) \\ &\quad + P(\overline{w_i}) \sum_{j=1}^{m} P(c_j|\overline{w_i}) \log P(c_j|\overline{w_i}) \end{aligned}$$

where $P(\overline{\omega_i})$ denotes the probability that ω_i does not occur. Then, $IGR(\omega_i)$ is defined as

$$IGR(\omega_i) = \frac{IG(\omega_i)}{IV(\omega_i)} \quad (9)$$

Rank	Word	User
1	ushuaia	chivil
2	rioja	ush
3	chivilcoy	poec
4	bragado	**malpegue**
5	viedma	**aijue**
6	logroño	tolhuin
7	chepes	vallerga
8	oberá	**yarca**
9	**cldo**	blv
10	tdf	portho
11	riojanos	jumeal
12	breñas	**sinf**
13	choele	plottier
14	gallegos	kraka
15	tiemposur	fsa
16	fueguinos	**bombola**
17	chilecito	**yarco**
18	blv	sanagasta
19	ush	wika
20	merlo	obera

Table 2: Top 20 words for the two metrics. Words in bold have lexicographic interest as regionalisms.

where IG is normalized by

$$IV(\omega) = -P(\omega)\log P(\omega) - P(\overline{\omega})\log P(\overline{\omega})).$$

We also calculate IGR with respect to the user frequencies of a word (which we abbreviate "user frequencies" for the sake of simplicity), in a similar way to Equation 4. As a baseline for our feature selection methods, we also calculate *Term-Frequency Inverse Location Frequency (TF-ILF)*, which consists in sorting our terms first by Location Frequency (in ascending order) and then by Term-Frequency (in descending order).

Summing up, five feature selection methods are tested as feature selection for geolocation: *TF-ILF*, *LTF-IG*, *LUF-IG*, basic *IGR*, and *User IGR*. We train Multinomial Logistic Regressions using the top N% words as features, and test against the 2.5K held out users. Performance is assessed using accuracy and mean distance between capital cities of each province – a fairly good estimate, since most of the population concentrates around those cities.

5 Results

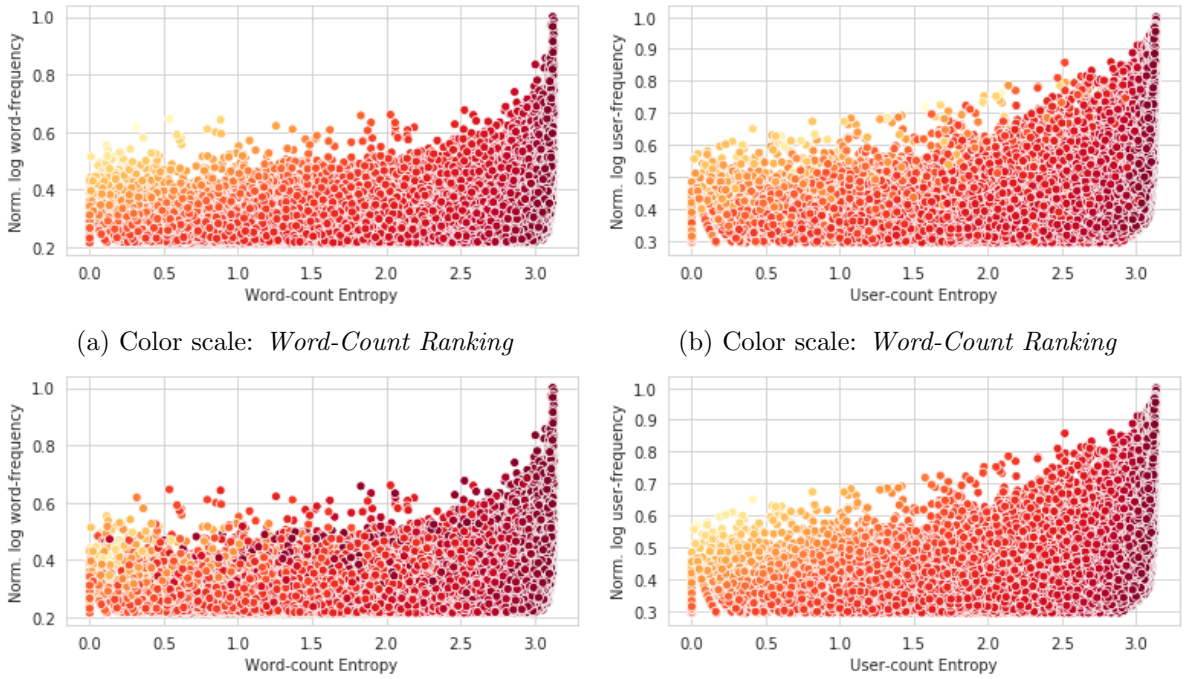
Table 2 shows the top-20 words calculated with each metric. Many are toponyms: *chivil, ush, blv, tolhuin, kraka, sanagasta, wika* refer to towns, cities and local clubs. Also, some words refer to gentilics (*riojanos, fueguinos*), or local institutions (*POEC*). Some of these words emerge as regionalisms: *yarca/yarco, aijue, sinf, cldo, bombola, malpegue*. We observe that the two rankings even share many words: *User-Count* and *Word-Count* have an overlap of 63% in the top thousand words.

Figure 2 shows four three-dimensional scatter plots. A dot in these plots corresponds to an individual word in our corpus, and is placed along the horizontal axes according to its word- or user-count entropy ($H_{\text{words}}(\omega)$ and $H_{\text{users}}(\omega)$, respectively). Along the vertical axes, each dot is located following its corresponding word or user frequency ($n_{\text{words}}(\omega)$ and $n_{\text{users}}(\omega)$). Additionally, each dot is colored according to the position of the word in one of our rankings using a chromatic scale, such that the lighter the dot, the higher the word's rank. For clearer visualization, word rankings are also shown in logarithmic scale.

Figure 2a shows that words higher in the *Word-Count Ranking* (in lighter color) tend to appear closer to the upper-left corner of the plot – that is, such words are more frequent and their mentions are concentrated in fewer regions. Figure 2d shows a very similar thing, now with respect to the number of users that mention the words: words higher in the *User-Count Ranking* are mentioned by a larger number of users from fewer regions. These two figures display a gradient from the upper-left corner (words ranked higher, in lighter color) to the lower-right corner (words ranked lower, in darker color).

Figure 2b uses horizontal and vertical axes corresponding to users (H_{users} and n_{users}), but colors each word with respect to the *Word-Count Ranking*. Here we can observe a slight perturbation in the gradient: there are words far from the left-corner that have light colors. From this, we understand that there are words with high *Word-Count Ranking* that have low *User-Count Ranking*.

Likewise, Figure 2c uses *User-Count Ranking* to color the points, and word axes H_{user} and n_{user}. The perturbation in the gradient is clearer in this plot; many words appear high in the *Word-Count Ranking* (closer to the top-left corner, see Figure 2a) but low in *User-Count Ranking* (darker color).



(a) Color scale: *Word-Count Ranking*

(b) Color scale: *Word-Count Ranking*

(c) Color scale: *User-Count Ranking*

(d) Color scale: *User-Count Ranking*

Figure 2: Scatter plots showing words (dots) along three dimensions. Horizontal axes: word-count entropy H_{words} (left plots) or user-count entropy H_{users} (right plots). Vertical axes: normalized log word frequencies n_{words} (left plots) or user frequencies n_{users} (right plots). Color: log word rank according to *Word-Count* (top plots) or to *User-Count* (bottom plots); lighter color means higher rank.

To further inspect this phenomenon, we searched for words that have large differences in the logarithm of *Word-Count Ranking* and *User-Count Ranking*. The logarithm reduces the difference between words ranked very high (e.g., between the word at position 10,000 and another in position 20,000) and amplifies the difference when one of the ranks is low and the other is high. A close examination of these words and their tweets showed that they were produced by bots (news and metheorological accounts, or accounts using tools to gain more followers) or in small niches of fans of some celebrity. From the top-100 words sorted by this difference, only one ranks higher in users than in words.

Summing up, when a word has a high *User-Count Ranking*, it also tends to have a high *Word-Count Ranking*. The reverse is not true, however, as words produced by a small number of accounts would not rank well with respect to users. Thus, the *User-Count Ranking* successfully discards words coming from automatic agents, as already done in Cui et al. (2012).

Word	Word Rank	User Rank
rioja	2	2499
vto	27	28179
hoa	81	83717
contextos	88	71290
cardi	32	23756
agraden	107	75042
hemmings	59	40227
ushuaia	1	565
tweeted	43	21342
precipitación	66	31042

Table 3: Top 10 words with the largest gaps between log word rank and log user rank.

5.1 Lexicographic Validation

The first thousand words in the *Word-Count Ranking* were manually analyzed by the lexicographers, who marked 21.9% as likely regionalisms. Likewise, from the first thousand words in the *User-Count Ranking*, 30.2% were marked as being lexicographically relevant. **This validation suggests that considering user-frequency dispersion is more relevant when assessing a word as a regionalism.**

Lexical characterization is illustrated in Table 4, which displays a few examples of groups of regionalisms found thanks to this methodology. A special note is reserved for the group of *indigenisms*, where a number of words were found coming from the *Guaraní* language (for instance, *mitaí, angá, angaú, nderakore*) and also from *Quechua* (*ura*). It is worth mentioning that the regions of the words derived from *Guaraní* – spoken in Northeastern Argentina, Paraguay, Bolivia and Southwest of Brazil – coincide with the region delimited by Vidal de Battini (1964).

Word	Region	Meaning
Colloquialisms		
culiado	Córdoba	asshole
chombi	Mendoza	poor in quality
carnasas	Neuquén	not classy, inelegant
bolasear	Cuyo	to bullshit
aprontar	E. Ríos	to get ready
Indigenisms		
ura	Northwest	vagina (quechua)
mitaí	Guaranitic	boy
angá	Guaranitic	unfortunate
Regional realities		
piadinas	San Juan	roll (food)
tarefero	Misiones	yerba mate worker
POEC	Neuquén	high School exam
Interjections		
aijue	Formosa	surprise
yirr	Corrientes	joy
aiss	Formosa	annoy
jiaa	Corrientes	yeehay
Ortographic variations		
pesao	Northwest	pesado
ql	Northwest	culiado
uaso	Córdoba	guaso
Regional Morpheme		
raraso	Córdoba	very strange (raro)
tardaso	Córdoba	very late (tarde)

Table 4: Examples of regionalisms found in the manual analysis. Each group corresponds to a subjective category found by the lexicographers during the annotation process.

5.2 Feature Selection for Geolocation

Moving on to the results of our second validation procedure, Figure 3 displays the performance of the different feature selection methods when used to train a discriminative classifier. Horizontal axes represent the percentage of top words selected, and the vertical axes represent the mean distance error in 3a and the accuracy in the case of 3b.

LUF-IG obtains the best performance in the user geolocation task, and stabilizes in a plateau at roughly 3.75% of top words used. It outperforms its word-frequency version LTF-IG and both IGR metrics. Table 5 displays the results of using the full bag of words (baseline) versus using the different feature selection methods with 5,000 top words.

When comparing our metrics, we note that the ones based on user-frequencies obtain a better performance than their word-frequency counterparts. This is more apparent in the case of LTF-IG and LUF-IG, but can also be observed for IGR metrics.

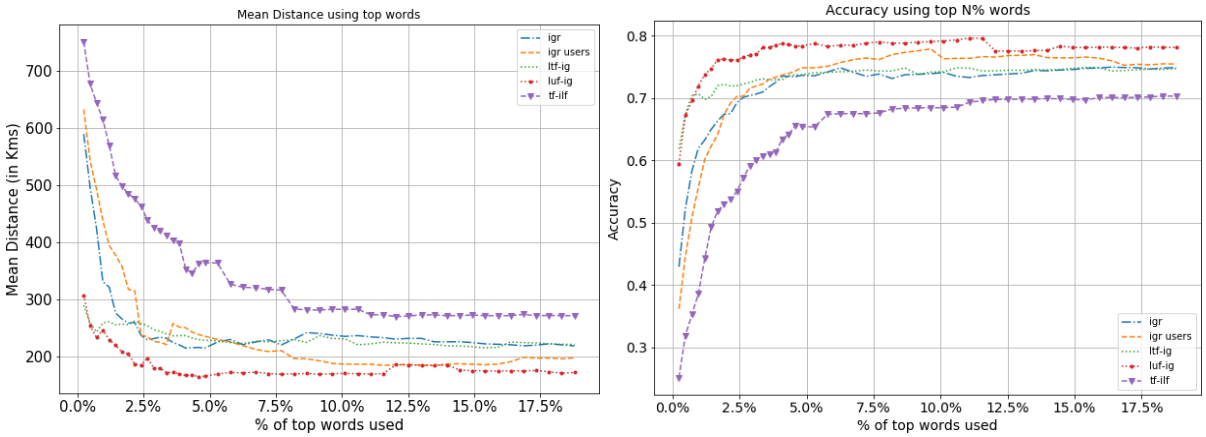
6 Discussion

Of the proposed metrics, *User-Count Metric* proved to be the most promising one. It successfully removed from the top of the ranking words likely to come from automatic agents or from small niches of users, and a manual lexicographic validation confirmed that this ranking contained more regionalisms than the *Word-Count Metric*. Further, using this metric as a feature selection method for geolocating users also showed a significative improvement over other metrics – both its word-frequency counterpart and IGR metrics from Han, Cook, and Baldwin (2012). This strongly suggests that measuring the dispersion of users of a certain word is a very informative indicator – both in lexicographic and in geolocation terms – backing what was already proposed in previous work to detect spam on Twitter (Cui et al., 2012).

The proposed metric was developed in the context of analyzing regional colloquialisms.

Features	Accuracy	Mean Distance
All	0.383	599.8
TF-ILF	0.654	363.3
IGR-Words	0.736	214.2
IGR-Users	0.748	234.7
LTF-IG	0.737	227.9
LUF-IG	**0.784**	**164.9**

Table 5: Performance of the different feature selection methods when using the top-5000 words.



(a) Mean distance error in user geolocation

(b) Accuracy in user geolocation

Figure 3: Comparison of the metrics when used as feature selection methods for geolocation. Vertical axes show the percentage of the top words used as features to train a Multinomial Logistic Regresion, and vertical axes display the performance of each respective classifier. Figure a uses mean distance error as y-axis (less is better) and Figure b uses accuracy (more is better)

This area of the lexicon is most elusive, since its impact on any printed medium arrives noticeably late – and in many cases it never reaches it at all. Colloquialisms are a class of words hardly found in any other media. Our best performing metric marked as relevant several words that were already listed in the *Diccionario del Habla de los Argentinos* (Academia Argentina de Letras, 2008), a fact that confirms the usefulness of both our metric and Twitter data in general for this task.

An outstanding subgroup of words found in the analysis are those coming from the *Guaranitic* region, in Northeastern Argentina. In particular, three words have already been proposed for addition to the aforementioned dictionary: *angá, angaú, mitaí*. This case is emblematic because it shows how this type of approach can help overcome the intrinsic limitations of doing regional lexicography. When lexicographers are native to only one of the different dialects of the region included in a projected dictionary, the probability of properly detecting and defining words of other dialects is slim or depends on mere chance. As the team of lexicographers expressed when confronted with these three words related to Guaraní heritage, those very robust normalized frequencies across a significant portion of the territory of Argentina would have otherwise remained unknown. Instead of including them in the next edition of the dictionary that attempts to describe all regional lexical items in the country, they would have remained unregistered, thus perpetuating a serious omission.

As our focus was in detecting lexical variations within provinces, we paid no attention to spatial granularity. If a better granularity were necessary in the analysis, adaptive partitioning could be used (Roller et al., 2012) to improve geolocation and to find localisms within provinces. Although previous work (Vidal de Battini, 1964) indicates that most provinces do not have large dialectal variations within them, this is something that would need to be explored and confirmed in future work.

Also, these techniques should be tested against other datasets, such as those used in (Roller et al., 2012; Han, Cook, and Baldwin, 2012), to further confirm that they outperform other feature selection methods.

7 Conclusions

In this work, we developed and compared two novel metrics useful for detecting regionalisms in Twitter based on Information Theory. One was based on the word frequency (*Log Term Frequency-Information Gain, LTF-IG*) and the other on the user frequency of a word (*Log user frequency-Information Gain, LUF-IG*). These metrics may be seen as a mixture of previous information-theoretic measures and classic *TF-IDF*.

We evaluated their performance in two ways. First, a team of lexicographers man-

ually assessed the presence of regionalisms in the first thousand words ranked by each metric. Second, we tested the metrics as feature-selection methods for geolocation algorithms, for which we also tested against metrics from previous works (Han, Cook, and Baldwin, 2012; Cook, Han, and Baldwin, 2014). In both evaluation types, the metric built upon user frequencies (*LUF-IG*) yielded the best results, suggesting that the number of users of a word is very informative – perhaps even more than simple word frequency.

This method has aided lexicographers in their task, allowing them to propose the addition of a number of words into the *Diccionario del Habla de los Argentinos*. The work behind this particular dictionary relies on a collaborative effort based on the intuition of scholars and lexicographers that identify regionalisms used mainly (seldom exclusively) within Argentina's borders by carefully parsing over a diversity of sources. Therefore, using Twitter to automatically detect regionalisms does not limit itself to avoiding most of this manual work, which, in and of itself, would already be a sizeable contribution. Since a considerable portion of the lexical repertoire of a community does not make its way across to published materials (which make most of the 300 millions words included to date in, for example, CORPES XXI (Real Academia Española, 2013)), the possibility of creating lists of words that are likely to be regional, based on actual utterances written by users, opens a way of shedding light onto entire pockets of lexical items that would remain otherwise chronically underrepresented in dictionaries. Even when a regional word is published, and then included in corpora, the task of appropriately isolating it remains largely unchanged, given that the word has to be previously identified in order to then take advantage of the statistical information available.

This work defines Argentinian provinces as the regional units of analysis, but this could be changed in order to repeat the analysis at different granularity levels. In this way, it might be possible to study intra-provincial dialectal differences (e.g., at the department level, see Section 3), although the limited precision of the geolocation of Twitter users may complicate this task. And it would definitely be possible to detect contrastive words across larger regions, for example to study Spanish in all its geographical variants.

A further challenge triggered by this work is the detection of regions with different dialectal uses (Gonçalves and Sánchez, 2014) but using features obtained in a semisupervised fashion with these metrics. This would allow to assess the validity of the dialectal regions of Argentina proposed by Vidal de Battini in 1964 (Vidal de Battini, 1964). Spatial and temporal information could be also explored, particularly finer-grained locations. Regarding geolocation, the proposed metrics should also be tested against other datasets to evaluate its performance as a feature selection method.

Acknowledgments

This work was funded in part by CONICET, Universidad de Buenos Aires, and Universidad Torcuato Di Tella. We thank Edgar Altszyler, Mariela Sued, and Federico Plager for helpful discussions, and our anonymous reviewers for valuable suggestions.

References

Academia Argentina de Letras. 2008. *Diccionario del habla de los argentinos*. Emecé Editores.

Ahmed, A., L. Hong, and A. J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. ACM.

Almeida, M. and C. Vidal. 1995. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):50.

Atkins, B. S. and M. Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Chang, H.-w., D. Lee, M. Eltaher, and J. Lee. 2012. @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society.

Cook, P., B. Han, and T. Baldwin. 2014. Statistical methods for identifying local dialectal terms from GPS-tagged documents. *Dictionaries: Journal of the Dictionary Society of North America*, 35(35):248–271.

Cui, A., M. Zhang, Y. Liu, S. Ma, and K. Zhang. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM 12, pages 1794–1798, New York, NY, USA. ACM.

Eisenstein, J. 2014. Identifying regional dialects in online social media. In *School of Interactive Computing Faculty Publications*. Georgia Institute of Technology.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Ghosh, R., T. Surachawala, and K. Lerman. 2011. Entropy-based classification of retweeting activity on Twitter. *arXiv preprint arXiv:1106.0346*.

Gonçalves, B. and D. Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

Grieve, J., C. Asnaghi, and T. Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American speech*, 88(4):413–440.

Han, B., P. Cook, and T. Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012*, pages 1045–1062.

Hecht, B., L. Hong, B. Suh, and E. H. Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246. ACM.

Huang, Y., D. Guo, A. Kasakoff, and J. Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Jimenez, S., G. Dueñas, A. Gelbukh, C. A. Rodriguez-Diaz, and S. Mancera. 2018. Automatic Detection of Regional Words for Pan-Hispanic Spanish on Twitter. In *Ibero-American Conference on Artificial Intelligence*, pages 404–416. Springer.

Kaufmann, M. and J. Kalita. 2010. Syntactic normalization of Twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 60–66. Morgan Kaufmann Publishers Inc.

Labov, W., S. Ash, and C. Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Montemurro, M. A. and D. H. Zanette. 2002. Entropic analysis of the role of words in literary texts. *Advances in complex systems*, 5(01):7–17.

Montemurro, M. A. and D. H. Zanette. 2010. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153.

Nerbonne, J., W. Heeringa, E. Van den Hout, P. Van der Kooi, S. Otten, W. Van de Vis, et al. 1996. Phonetic distance between Dutch dialects. In *CLIN VI: proceedings of the sixth CLIN meeting*, pages 185–202.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Rahimi, A., T. Baldwin, and T. Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using

mixture density networks. *arXiv preprint arXiv:1708.04358*.

Rahimi, A., T. Cohn, and T. Baldwin. 2017. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*.

Real Academia Española. 2013. Banco de datos (CORPES XXI) [online]. *Corpus del español del siglo XXI (CORPES)*.

Roller, S., M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ueda, H. and A. Ruiz Tinoco. 2003. Varilex, variación léxica del español en el mundo: Proyecto internacional de investigación léxica. In *Pautas y pistas en el análisis del léxico hispano (americano)*. Iberoamericana Vervuert, pages 141–278.

Vidal de Battini, B. E. 1964. El español en la Argentina. Technical report, Argentina.

Zheng, X., J. Han, and A. Sun. 2018. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.