BOBCATSSS 2022 • DEBRECEN

UNIVERSITY of DEBRECEN

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III
Medien, Information und Design

ANALYSIS OF SPOTIFY SPANISH SPOKEN PROFILES ON TWITTER

Juan-José Boté-Vericad

Facultat d'Informació i Mitjans Audiovisuals. Universitat de Barcelona. (Spain)

Abstract

Twitter is a social networking site where brands create profiles and interact with their audience. Brands also look for new audiences to consume their products or services. In some cases, they make different profiles for different countries or linguistics regions. The written language is the central Twitter expression. Multimedia elements such as images or videos help spread the message and interact with the audience. In this study, we analyze the different profiles of Spotify in the Spanish language addressed to other spoken Spanish countries.

Spotify has different profiles on Twitter addressing the content to other spoken Spanish language countries. These profiles are addressed to Spain and South American countries such as Argentina, Chile, Colombia, and Mexico. Finally, a profile under the name LATAM is addressed to the rest of the Spanish spoken countries in South America. All these profiles have different audiences and differences in the number of followers.

In all these countries, Spanish is spoken with different linguistic variations. As a result, the message is other. Consequently, audience interaction and engagement may vary from one profile to another depending on the written language used.

The analysis considers these Spanish linguistic variations. We perform a sentimental study of these Spanish spoken profiles, looking for differences in Spanish variations. We also combine the research with topic modelling and the use of hashtags. Spanish linguistic variations may influence the analysis but in the profile's engagement.

Our results show that while messages are similar in writing, engagement with the audience varies from profile to profile. We conclude that Spanish variations influence engagement and commercial companies should consider a similar strategy. We suggest not unifying under a unique spoken Spanish version to promote products and services in Spanish-speaking countries.

Keywords: Twitter, Spotify, Data mining, Spanish variations.

1 INTRODUCTION

This paper aims to analyse the Spanish variations on Twitter concerning the Spotify Spanish spoken profiles. As a deliver music company, Spotify has different audiences with different profiles on Twitter. It seems reasonable since the company has profiles for the United States, UK & Ireland and India. In the case of Spanish spoken countries, the company has profiles in Argentina (@SpotifyAGR), Chile (@SpotifyChile), Colombia (@SpotifyColombia), Mexico (@SpotifyMexico), Spain (@SpotifySpain) and finally LATAM (@Spotify_LATAM). This last profile encompasses other countries not represented in local accounts.

In scientific literature, linguistic variations in Twitter seem scarce. Most studies classify language variations using the geotagging (also the geolocated) approach to later create a classification. This means collecting a corpus of tweets based on the location, such as a study concerning the Arabic language. Omar et al., (2020) studied a corpus of 1,597,348 geolocated Twitter posts from 650,847 users using later principal component analysis to classify lexical features. They found different intensifiers of the Arabic language classifying tweets from other countries such as Egypt, Lebanon, Sudan and linguistic variations in the most popular six dialects in the Arabic language. Similar studies are concerning variations of the language in English (Huang et al., 2016; Miletic et al., 2020), Dutch variations (Dijkstra et al., 2021) and Spanish (Donoso & Sánchez, 2017; Gonçalves & Sánchez, 2014).

Another approach to collecting geotagged tweets is having a big corpus publicly available of tweets to use later machine learning algorithms. This is the case of the study of Petrović et al., (2010), who collected a corpus of 97 million tweets not restricted to English under the Creative Common license.

Miletic et al. (2020) explored 78.8-million-tweet to study variations in Canadian English, focusing on the dialects spoken in Toronto, Montreal and Vancouver. The collected tweets identified Twitter users in geographic areas of interest and later crawled the indexed user through their timelines. They tested different libraries of language identification to see their accuracy. They concluded that their corpus could allow the novel regional variants.

Similarly, Dijkstra et al. (2021) performed a study analysing Frisian, a spoken language in the northwest of the Netherlands. They identified 186 Twitter profiles and collected their tweets from 2010 to 2019 with 698,369 tweets. They concluded that collecting Frisian tweets from Twitter was complex because the corpora are small compared to Dutch speakers, and the automatic identification of Frisian and Dutch was not very successful because they are closely related languages. Consequently, they had to analyse manually 100,365 tweets to distinguish their target from other words.

2 METHODOLOGY

The methodology of this study consists of analysing 6 Twitter Spotify profiles addressed to Spanish spoken countries. The studied profiles are SpotifyArgentina (2013), SpotifyChile (2013), SpotifyColombia (2013), Spotify_LATAM (2013), SpotifyMexico (2011) and SpotifySpain (2009). Years between parenthesis are the year when the profile was created. Because of the extension of the proceeding, we have chosen a temporary limitation of 2021. To process and analyze the data, we used the Twitter Academic API and the current version of R software (Fig. 1).

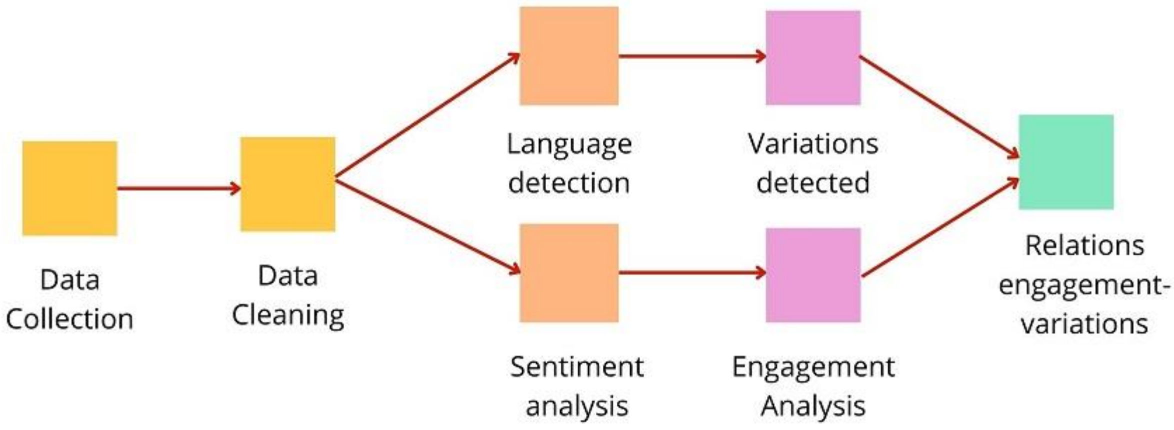

Figure 1. Datasets workflow

3 RESULTS

We have downloaded from Academic Twitter the tweets from all the accounts individually. To have a uniform number of profiles and tweets, we have made a temporary limit from 2013 to 2021. Some profiles, such as @SpotifySpain, were created in 2009. This profile started having activity in 2012 and beyond. Other profiles began their activity in 2013. In Table 1, there are also 3 exceptions. In 2018 three profiles, @SpotifyARG, @SpotifyChile and @Spotify_LATAM, had minimal activity, and the number of produced tweets was not enough to analyse, and we have discarded them. As we can see, the activity of these profiles raises significantly in 2018 in all profiles. The profile with more tweets is Mexico, with 100661 tweets and which activity has increased its activity in the last three years. This is 2019, 2020 and 2021.

Table 1. Descriptive table with the number of tweets produced by any profile.

	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total
Argentina	615	1060	2654	2179	153		2824	1968	9543	20996
Chile	144	769	1301	1057	102		142	252	6432	10199
Colombia	179	778	1925	1487	501	1164	2294	2416	27604	38348
LATAM	58	391	248	684	174		140	1359	422	3476
Mexico	1886	2900	2483	4502	2808	2615	7379	47422	28666	100661
Spain	315	1692	2651	1080	524	420	765	18254	1922	27623

Since it is supposed that all profiles are in Spanish, we have also tested the accuracy of 3 packages that detect the language based on probabilities and the corresponding ISO. These R packages are *textcat* (Hornik et al., 2020), *fastest* (Mouselimis, 2022), and *franc* (Csardi et al., 2021). Each package provides the probability of detecting a language. Most of the tweets were detected in Spanish, but some texts were not correctly detected. We provide in Table 2 an example of language detection from 10 tweets corresponding to @Spotify_AGR.

Table 2. Detecting language probability from tweets.

	ISO LANG	**Probability**
1	eu	0.1378300
2	fa	0.1101400
3	es	0.7650530
4	es	0.4139960
5	en	0.3723510
6	en	0.1868070
7	en	0.1498270
8	en	0.2664110
9	en	0.2867880
10	en	0.2145200

Tweets that are not detected in Spanish could be for various reasons. For example, in the text of the tweets, some adopted words or expressions in English were used, considering that the probability of detecting English is very low such as in the case of rows five to10 in Table 2. Then, improvements in the cleaning process need to be considered to avoid these issues later.

3.1 Representation of the Spanish variations

We have studied profile by profile differences in the words to represent Spanish variations, especially tweets with more engagement concerning others. We have selected the 20 more engaging tweets and the 20 more frequent words from each profile. The profile with the most variations is Argentina, where there are four words among the most used.

Table 2. Frequency of the most used words in @Spotify_ARG – Year 2021.

	Words	N		Word	N
1	temporada	5069	11	spotify	1918
2	notificación	4372	12	año	1845
3	recibir	4371	13	artista	1815
4	estrenar	2571	14	resumen	1814
5	preparadx	2570	15	acá	1810
6	avisar	2569	16	preparate	1805
7	disponible	2477	17	descubrí	1794
8	viajar	2472	18	escuchado	1776
9	realidad	2470	19	top	1775
10	preparatar	2467	20	podcaster	1770

When observing Table 2, 2 words are a variation of the Spanish: number 15 acá, which means here and number 17 descubrí, which is an imperative clause, discover!. The high frequency of these words does not mean that other words with linguistic variations are not used. Fig. 2 represents the word cloud in @SpotifySpain in 2021.

There are some common hashtags for all accounts. First, #Caso63 is common in the Twitter profiles from Argentina, Chile, Colombia and LATAM. The reason is that Caso3 is a fiction series played in Argentina, Chile, Mexico, and the USA. Second #NovedadesViernes is common to LATAM, Colombia and Spain, especially Spain, where this hashtag generated many engagements. The meaning is a novelty on Fridays, generating 600 retweets, 171 replies, 3548 likes, and 33 quotes. Quoting is not very much used in all profiles.

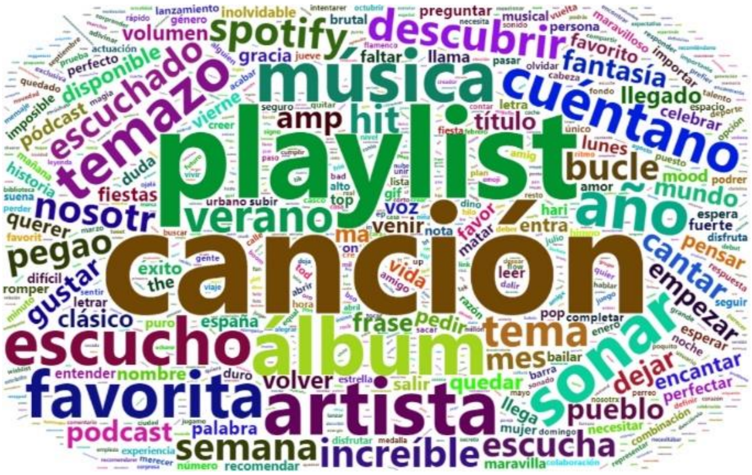

Figure 2. Word Cloud @SpotifySpain – year 2021

3.2. Sentiment Analysis

We have also performed a sentiment analysis from all profiles. Our sentiment analysis was based on the work of Mohamad & Turney (2020). They use a dictionary called "NRC" and tag each word in one of 10 emotions through an intensive search algorithm. In Fig 3, there is an example of @SpotifyColombia in the year 2021. Published tweets usually have this pattern where the use of positive language in addition to trust and joy. It took our attention concerning other countries a high level of negative punctuation that in other countries is low.

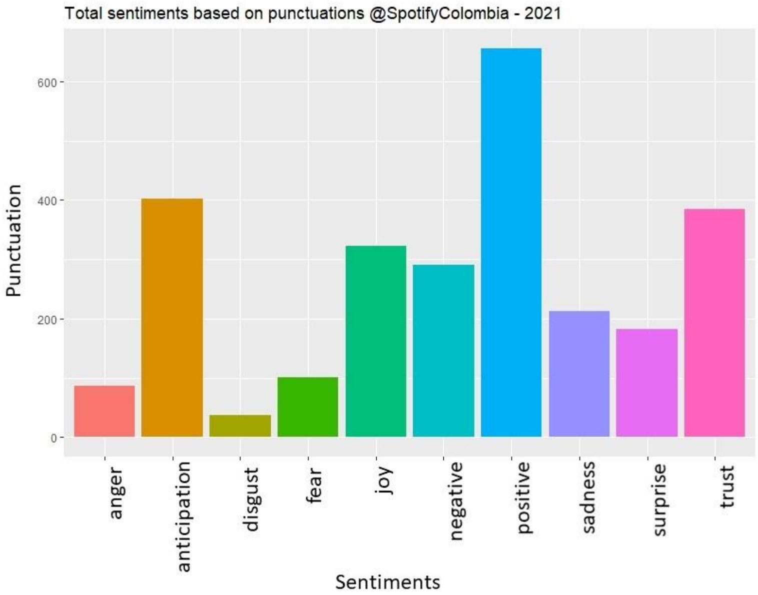

Figure 3. Sentiment Analysis @SpotifyColombia - 2021

3.3. Limitations

One of the limitations to identifying Spanish language variations in these profiles is that when detecting language on the tweets, the ISO identification code in R libraries is "es" for all countries corresponding to Spanish. Moreover, ISO language codes for Spanish Spoken countries are Argentina: es-AR, Chile: es-CL, Colombia: es-CO, Mexico: es-MX and Spain es-ES, which R libraries do not use. The situation is similar in the Python language. This is problematic when looking for linguistic variations as a piece of information. In addition, to our knowledge, there is no existing corpus from these countries to be compared with machine learning techniques.

4 CONCLUSIONS

Our results show different situations concerning Spanish linguistic variations. It should be considered that in Latin America, Twitter profiles do not use regularly geotagged tweets. Then, it is necessary to identify profiles from a concrete country and collect this data to have a corpus. Then, the best approach could be to use a machine learning approach compared with a corpus in the same country. For instance, in studying Spanish variations from Colombia, it is necessary to have a corpus of Colombia profiles. In our study, we collected data from different Twitter profiles from Spotify. We had to manually analyse the data and the hashtag to identify linguistic variations. Spotify seems to have a strategy for some hashtags where the profiles have a high level of engagement. Spotify profiles are focused on the international music industry but mainly on local music from each country. This allows Spotify as a delivery music company to reach better engagement than a global account. Moreover, the use of local expressions in the text or some local events in the hashtag allows the user to be identified with a local event. As a strategy, we recommend other companies who want to focus on the Spanish spoken market to perform a similar approach.

REFERENCES

Csardi, G., Wormer, T., Ceglowski, M., Rideout, J. R., & Johnson, and K. S. (2021). Franc: Detect the language of text (1.1.4) [Computer software]. https://CRAN.R-project.org/package=franc

Dijkstra, J., Heeringa, W., Jongbloed-Faber, L., & Van de Velde, H. (2021). Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in

Frisian. *Frontiers in Artificial Intelligence*, *4*. https://www.frontiersin.org/article/10.3389/frai.2021.644554

Donoso, G., & Sánchez, D. (2017). Dialectometric analysis of language variation in Twitter. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 16-25. https://doi.org/10.18653/v1/W17-1202

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *PLOS ONE*, *9*(11), e112074. https://doi.org/10.1371/journal.pone.0112074

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). The textcat package for n-gram based text categorization in r. Journal of Statistical Software, 52, 1–17. https://doi.org/10.18637/jss.v052.i06

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244-255. Scopus. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Miletic, F., Przewozny-Desriaux, A., & Tanguy, L. (2020). Collecting Tweets to Investigate Regional Variation in Canadian English. *Proceedings of the 12th Language Resources and Evaluation Conference*, 6255-6264. https://aclanthology.org/2020.lrec-1.767

Mohammad, S., & Turney, P. (2020). *NRC Emotion Lexicon*. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Mouselimis, L. (2022). Efficient Learning of Word Representations and Sentence Classification. https://cran.r-project.org/web/packages/fastText/index.html

Omar, A., Ethleb, H., & Hashem, M. E. (2020). Mapping Linguistic Variations in Colloquial Arabic through Twitter: A Centroid-based Lexical Clustering Approach. *International Journal of Advanced Computer Science and Applications*, *11*(11), 73-81. Scopus. https://doi.org/10.14569/IJACSA.2020.0111110

Petrović, S., Osborne, M., & Lavrenko, V. (2010). The Edinburgh Twitter corpus. *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, 25-26. https://aclanthology.org/W10-0513.pdf