# Colloquiality analysis on social networks: a case from Twitter

**André Luiz França Batista[1], Carlos Henrique da Silveira Campos[2], Daniel Ramos Pimentel[3]**
[1,2,3] *Department of Computer Science*
*Federal Institute of Triangulo Mineiro*
*Ituiutaba, MG, Brazil*
[1]andreluiz@iftm.edu.br, [2]carloscamposiki@gmail.com, [3]danielpimentel@iftm.edu.br

## Abstract

*Social network sites are present in a constant way in society. The manner people write in social network sites has their dynamism because of the speed information are replicated, the reach publications may have, network sites' peculiarities and the seek for fame. This generates linguistic constructions that are not in accord with the standard norm of the Portuguese language. This quantitative work aims to relate the use of colloquial constructions on* Twitter *with user's popularity, posts popularity and other specific factors of this social network. This analysis was made using regular expressions, dictionaries, and frequency distribution to identify colloquial constructions. A support system was developed to perform analysis, management, and mining.*

**Keywords: s**ocial network sites, Twitter, natural language, linguistics.

## 1. Introduction

Social networks have become an important element in modern society, with 2.65 billion active users in 2018 and a growth prospect to 3.1 billion by 2021 (Statista, 2020). Considering that the world population was 7.8 billion people in the same year (UNFPA, 2019) we have almost half of the entire globe using social networks.

Among these networks, Twitter registered 330 million active users in the first four months of 2019 (Statista, 2020). A tweet is a text message from a user containing no more than 280 characters that can contain other media such as images and videos (Twitter, 2020), where 500 million tweets are sent per day (Internet Live Stats, 2020).

In this work, we verify in a general way the incidence of colloquial structures in tweets and the relationship between colloquiality and popularity. Among the specific objectives of this study we can quote: (a) To identify occurrences of colloquiality in a tweet. (b) To verify if there is disparity in the colloquial level between tweets considering the number of followers. (c) To verify if there is a relationship between replication and popularity of tweets with occurrence of colloquialities. (d) Check for disparity in colloquial level between verified users, who have been authenticated by Twitter as legitimate, assigned to accounts that Twitter considers of public interest (Twitter, 2020), and unverified users. (e) Check for disparity in the colloquial level between tweets responses, which are targeted at one person and not all followers, and normal tweets. (f) Check for disparity in colloquial level between users with a large difference in total published tweets.

## 2. Theoretical foundation

Social networking features such as limitations and functionality create language conventions for its users. Twitter, for example, has a more severe character limitation per post than other social networks, so users use structures like contractions constantly to express themselves as Gouws, Metzler, Cai and Hovy (2011) show. The use of these structures in social networks in general, besides Twitter, is justified by users as a necessity for speed (YAGUI, 2018).

Perception and familiarity with the language is another relevant element in the use of colloquiality. Posts made by users in a language that is not their native language tend to use less these constructions as shown by Perez-Sabater (2012).

A concern with the popularity of social networks in the educational environment is the use of colloquial structures by these users outside the Internet. Souza and Depz (2012) showed that students who use social networks do not tend to use more informal constructions than non-users when in situations that require formality but end up using more in informal situations. Popularity and reach are relevant factors in social networks and internet in general.

Hagen, Uzener, Harrison and Katragadda (2016) showed a correlation between language and the popularity of online petitions. Their study uses computational tools to explore e-petitions, viewing them as persuasive texts with linguistic and semantic features that may be related to the popularity of petitions, as indexed by the number of signatures they attract. Hagen, Uzener, Harrison and Katragadda (2016) made use of a website data, to analyse linguistic features, such as extremity and repetition, and semantic features, such as named entities and topics, to determine whether and to what extent they are related to petition popularity.

There are several studies involving social networks, where various are an analysis of the network itself, such as Boyd, Golder & Lotan (2009), which examined the practice of retweeting, action of replicating to their followers a tweet from someone else. Other studies include linguistics in their analysis such as Liu, Li & Guo (2012) and Go, Bhayani & Huang (2009), who work with models for user mood identification. Nguyen, Gravel, Trieschnigg & Meder (2013) did a study linking language and user age on Twitter, using machine learning. Still, there is a lack of studies dealing with the language itself.

## 3. Methodological path

### *3.1 Data Collection*

The data collection stage in which the tweets sample was collected for analysis on users' colloquiality utilization.

For data collection, Twitter Search was used, which allows many tweets to be obtained. Twitter Search requires an API key which is obtained by creating an application on Twitter Developers (Twitter, 2020).

To use Twitter Search you must provide at least one keyword to search. More frequent nouns in the Portuguese language were used according to the Frequency Dictionary of Portuguese (Davies & Preto-Bay, 2008). Only nouns that did not include accents, tilde and cedilla were selected in order to have a less arbitrary data search. At each search cycle, interspersed by a minute of waiting so that the collection was not blocked by the Twitter API request limit, an average of 400 tweets were returned and then the current search noun was changed so that the number of distinct tweets remained high.

The tweets were then received, treated and persisted on the server. A returned tweet contains several metadata: from data about the tweet itself, such as ID number, retweets and favorites, to information about the user himself, such as followers' number and geolocation data.

The following data, extracted at the time of the search, for each tweet was persisted on the server:

- Tweet ID number: Unique number that identifies the tweet. Useful for a possible manual check of information.
- Text: Tweet text, 140 characters maximum.
- Username: Unique username that identifies the tweet author.
- Followers: Number of followers of the tweet author. The number of followers is directly linked to the user's popularity.
- Number of tweets: number of tweets performed by the author during their account lifetime.
- Favorites: Number of favorites or likes a tweet received. It is directly linked to the popularity of the tweet and, but not necessarily, the author.
- Verified Flag: Informs whether the user, author of the tweet, has been authenticated by Twitter or not. The verified flag is assigned to accounts that Twitter considers of public interest (Twitter, 2020).
- Flag Response: Informs if the tweet is a direct response to another tweet that is usually seen only by people following the tweet user who received the response and the user who answered simultaneously.

### *3.2 Text standardization*

A treated version of the tweet text is also persisted in which the at sign, hashtags and link are replaced by a single structure indicating such occurrences.

This substitution occurs because an at sign is a quote to another user (Twitter, 2020) and can be treated as a proper name. Therefore it is valid to replace all at signs by a single structure that will always be treated as a proper name since the specific name is not relevant.

The hashtag is used as keywords to facilitate searching for specific topics (Twitter, 2020). A hashtag is often replicated and cannot be analyzed since it does not reflect the use of colloquiality of the tweet author since he is hardly the author of the hashtag. Similarly a link can be replaced since the link itself does not provide any information about the use of colloquiality by the tweet author.

For example, the text of a tweet in its original form:

E a @REDCanids fecha o primeiro jogo e garante o 1x0! GG! #CBLoL https://t.co/VgVFUK1J5N

This tweet is in Portuguese and its translation into English means: "The @REDCanids finish the first game ensuring the 1x0! GG! #CBLoL https://t.co/VgVFUK1J5N". In the example we have an at sign"@REDCanids" that refers to a user, then we have the hashtag "#CBLoL" that refers to the event that the tweet is contextualized and finally we have the link "https://t.co/VgVFUK1J5N".

When receiving the tweet with the metadata, the links, at signs and hashtags are provided and detailed so the replacement could be done directly without having to identify when they occur. The treated example tweet text follows this:

E a <USER> fecha o primeiro jogo e garante o 1x0! GG! <HASHTAG> <LINK>

The structures for replacement were defined so that they were simple to view and so that they avoided false positives with original text.

There were 92727 distinct tweets collected between March and November 2017. All tweets were only persisted if they were informed by their metadata as belonging to the Portuguese language to avoid the inclusion of tweets with cognates with other languages, mainly Spanish.

### *3.3 Tokens production*

For textual analysis of tweets the text is first transformed into tokens. A token is a sequence of characters that can be a word, a numeral or even a combination of letters, numbers or special characters.

NLTK provides a specific tokenization function for Twitter, the TweetTokenizer that adapts and deals better with casual texts that include emoticons and other peculiarities like hashtags and at signs.

The following is an example of a tweet:

E a <USER> fecha o primeiro jogo e garante o 1x0! GG! <HASHTAG> <LINK>

After the tokenization we have a list of tokens that can be analyzed:

Table 1. List of tokens that can be analyzed

| | | | |
|---|---|---|---|
| E | o | garante | GG |
| a | primeiro | o | ! |
| <USER> | jogo | 1x0 | <HASHTAG> |
| fecha | e | ! | <LINK> |

It is worth noting that punctuations adjacent to words such as commas, interrogations and exclamations become a single token dissociating itself from the accompanying word. This facilitates analysis since we have the words and punctuations in their raw forms.

### *3.4 Tokens identification*

The identification of a token is done as shown in Figure 1. The token when entering the flowchart to be identified is first separated into its category. Category 1 includes tokens that contain only alphabetic character sequence. Category 2 includes tokens containing at least one special character such as punctuation and emoticon, category 3 must include digits from 0 to 9 and optionally alphabetic characters.

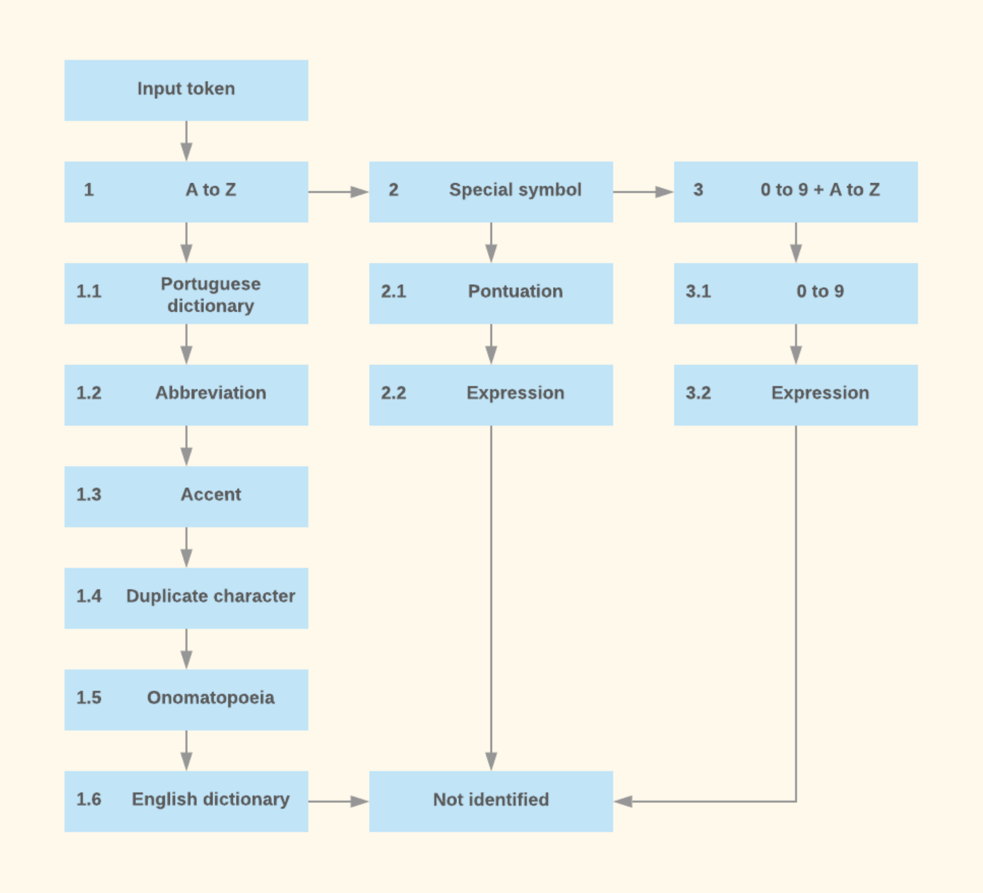


Figure 1: Flowchart of token identification.

Once in its category the token scrolls through the subcategories to be identified. When the token is identified the flowchart is interrupted and the subcategory of the token is defined. The order of the categories in the flowchart was defined in view of the frequency of distribution of their tokens from all the collected tweets. The distribution frequency tells you how often a distinct token appears in a text.

Thus, tokens that include only alphabetic characters, category 1, is the token category with the highest distribution rate. Tokens that include some special character, category 2, comes right after as the second most frequent category, finally category 3 comes last since their tokens are the least frequent among the three categories.

The arrangement of subcategories follows the same criterion. For example, subcategory 1.1 is the most frequent within category 1, according to the frequency of distribution of its tokens. The only exception is subcategory 1.6, which even though it has a higher distribution frequency rate than some subcategories in its category was placed last in order to reduce false positives with the English dictionary.

Once the token falls into its category and is not identified by any of its subcategories the token is then categorized as unidentified. The implemented preliminary version of this flowchart was able to identify 95.8% of all collected tweets tokens.

The tokens "<USER>", "<HASHTAG>" and "<LINK>" are categorized before entering the flowchart since they are peculiarities of the social network itself.

The description for each subcategory according to its category:

- **Category 1, A to Z:** Tokens containing alphabetic characters only.
  - **Subcategory 1.1, Portuguese Dictionary**: With the help of HunSpell the subcategory checks whether the token is valid in the Portuguese dictionary.
  - **Subcategory 1.2, Abbreviation:** A distribution frequency was made where tokens considered valid by the dictionary were removed keeping only the invalid ones. Checking the remaining tokens with higher distribution frequency it was possible to select 158 abbreviations of most common use on Twitter, these abbreviations were then persisted in the database. This subcategory checks whether the token belongs to this abbreviation list.
  - **Subcategory 1.3, Accent:** Generates variations of the token with accent, tilde and cedilla and checks whether these variations are valid in the dictionary. The "icon" token will not be identified in subcategories 1.1 and 1.2, here its variation "icon" will be tested and recognized as valid by the dictionary which will indicate an accent error in the tweet.
  - **Subcategory 1.4, Duplicate letter:** Remove duplicate letters from a token and check if it has become valid in the dictionary. For example the token "VERYYY" will be ignored in the previous subcategories, removing the duplicate letters will have the token "VERY" that will be recognized as valid by the dictionary and that will indicate a colloquial occurrence.
  - **Subcategory 1.5, Onomatopoeia:** Using regular expression it was possible to identify the most common onomatopoeia, according to the frequency of distribution. For example, the token "hahaha" will be identified in this category as an onomatopoeia that indicates laughter, thus indicating a colloquial occurrence in the tweet.
  - **Subcategory 1.6, English Dictionary:** Checks whether the character is valid in the English dictionary. This category is valid only for tokens that are at least five characters long to minimize the number of false positives.

- **Category 2, Special Symbol:** Tokens containing at least one special character.
  - **Subcategory 2.1, Score:** Recognizes punctuation and signs such as commas, quotation marks, question marks, exclamations, etc.
  - **Subcategory 2.2, Expressions:** Recognizes by means of regular expressions tokens that include special symbols such as dates ("01/01/2010"), ordinal values ("1st") and emoticons that have been listed as the most frequent by distribution rate.
- **Category 3, 0 to 9 and A to Z:** Tokens containing digits 0 to 9 accompanied by non-alphabetic characters.
  - **Subcategory 3.1, 0 to 9:** Tokens that include only numbers. Examples: "44", "152" etc.
  - **Subcategory 3.2, Expressions:** Recognizes through regular expressions structures with high distribution frequency containing letters and numbers only. Among these structures we have values with specific units like "10V", "100mm", "500GB" and other structures like scoreboards ("1a2", "2x2", "1vs1" etc.).

### *3.5 Colloquiality identification in Tweets*

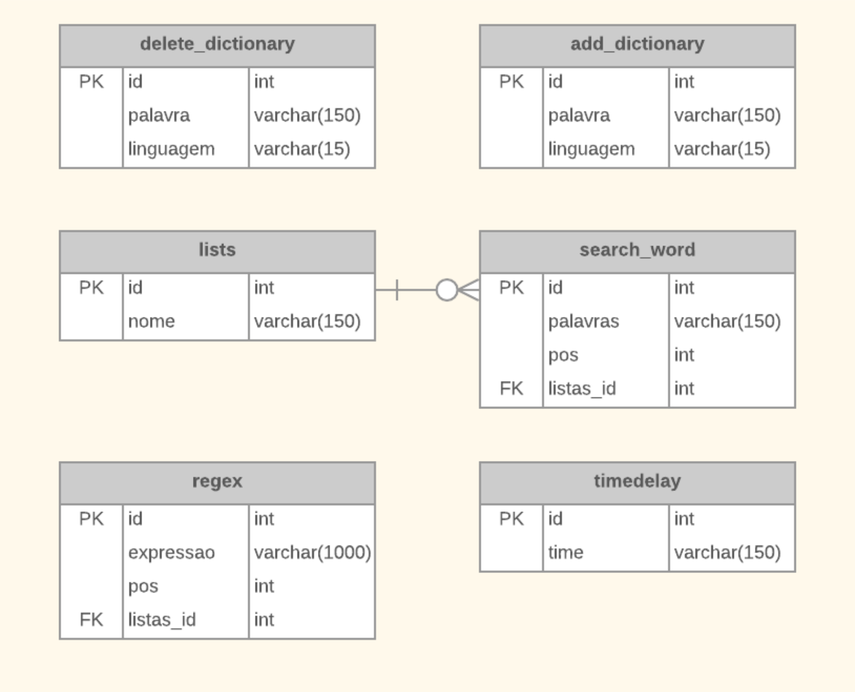
Once the tokens are identified, there is colloquiality in the tweet according to the identification of its tokens. The colloquial points observed in a tweet are:

- No capital letters at the beginning of the tweet.
- Absence of tweet punctuation.
- Presence of onomatopes.
- Presence of non-accented words.
- Presence of words with duplicate letters.
- Presence of foreign terms.
- Presence of emoticons.

With the identified colloquial tweets the calculations will be made to obtain the results according to the criteria described in the specific objectives.

### *3.6 Database Modeling*

Modelling the database where Tweets are persisted together with data used by the support system is shown in Figure 2.



| delete_dictionary | | |
|---|---|---|
| PK | id | int |
| | palavra | varchar(150) |
| | linguagem | varchar(15) |

| add_dictionary | | |
|---|---|---|
| PK | id | int |
| | palavra | varchar(150) |
| | linguagem | varchar(15) |

| lists | | |
|---|---|---|
| PK | id | int |
| | nome | varchar(150) |

| search_word | | |
|---|---|---|
| PK | id | int |
| | palavras | varchar(150) |
| | pos | int |
| FK | listas_id | int |

| regex | | |
|---|---|---|
| PK | id | int |
| | expressao | varchar(1000) |
| | pos | int |
| FK | listas_id | int |

| timedelay | | |
|---|---|---|
| PK | id | int |
| | time | varchar(150) |

Figure 2: Database modeling.

### *3.7 Support System*

A support system was developed to assist in colloquiality analysis. The Tweetllect system was developed in Portuguese and therefore the screens shown here are in this idiom. The main information on some screens is translated into English (inside a white textbox like next picture) but other, less relevant information remains in Portuguese. The system allows you to perform the analysis with the desired parameters, list tweets, view graphs to observe how the tweets are distributed according to their metadata, perform tweets mining and manage regular expressions, search lists, dictionaries, colloquial words, and unknown terms. Figure 3 shows the main system screen with the main system functions.

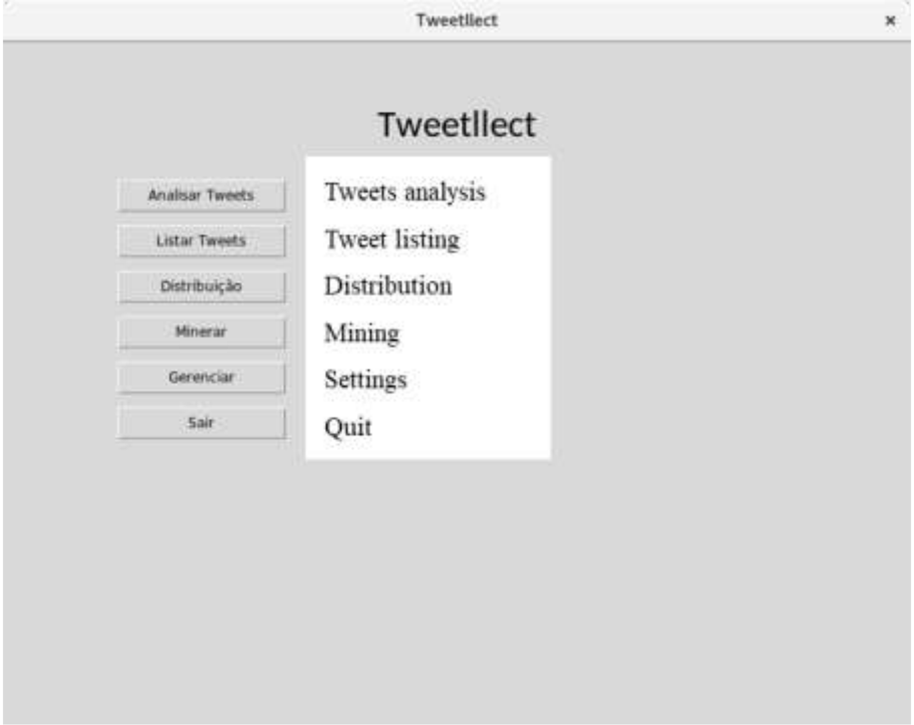
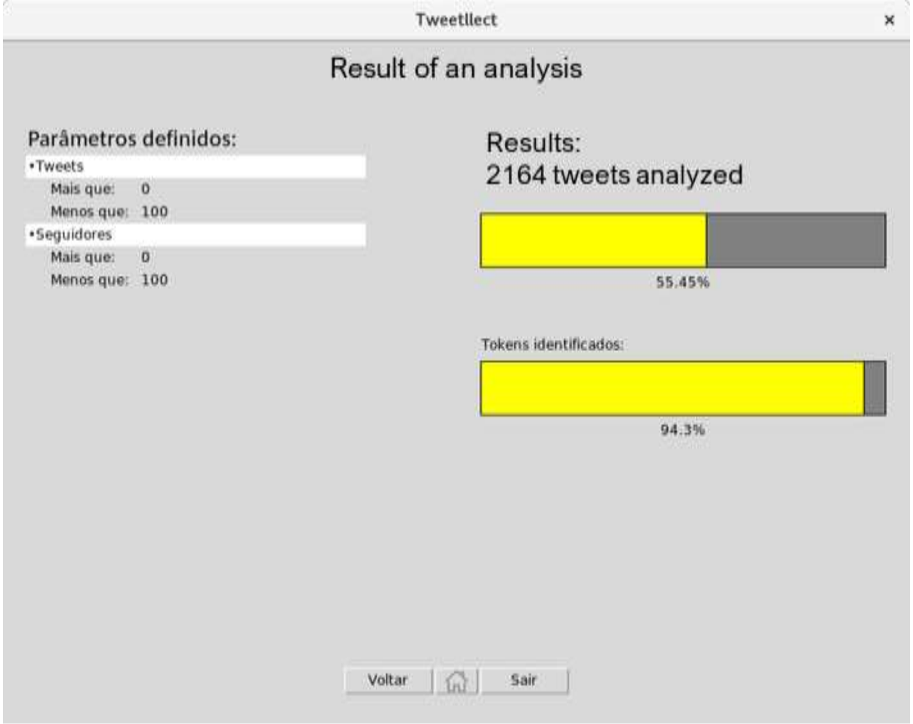


Figure 3: Support system main menu.

In the parameter setting screen for both tweets analysis and tweets listing, the parameters are, in relation to the user: number of followers, number of tweets and if it is a verified profile; in relation to the tweet: number of favorites, number of retweets and if it is a tweet reply. Any combination of parameters can be defined or a tweets analysis or search without parameters can be made for a result involving all tweets.

In Figure 4 the screen shows the result of an analysis. You can see the parameters set, the percentage of tweets with some colloquiality found, and the percentage of tokens identified.
In Figure 5 you can see the list of tweets. For each tweet the name and number of followers of the author, the text of the tweet and a button for more details are displayed.



Result of an analysis

Parâmetros definidos:

•Tweets
Mais que: 0
Menos que: 100
•Seguidores
Mais que: 0
Menos que: 100

Results:
2164 tweets analyzed

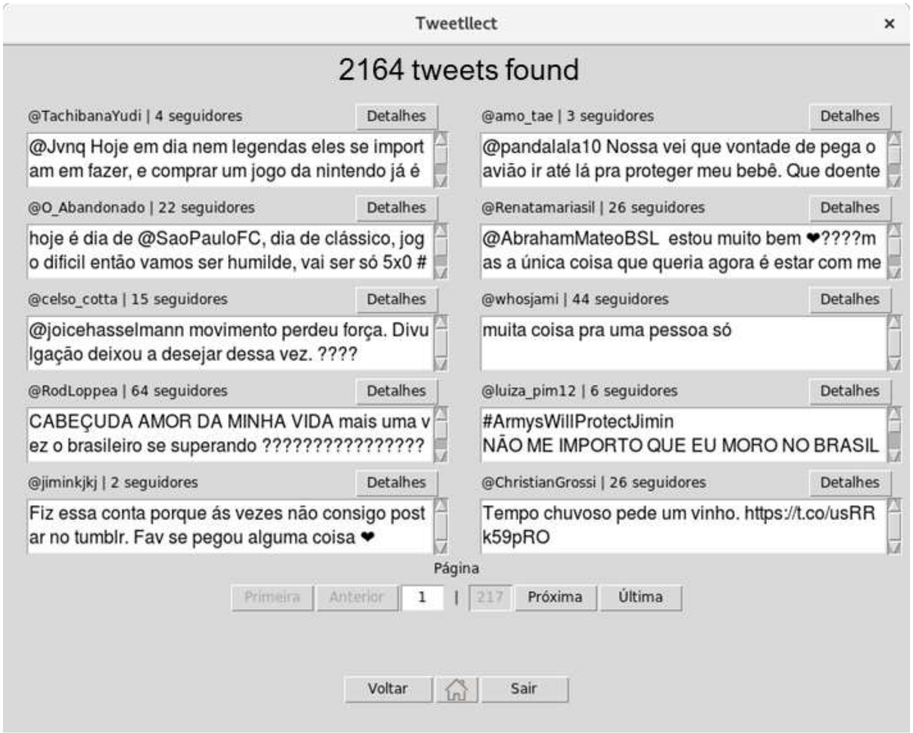
55.45%

Tokens identificados:

94.3%

Figure 4: Analysis result screen.

The tweets list screen shows the total tweets found given the input parameters and displays ten per page.



2164 tweets found

@TachibanaYudi | 4 seguidores
@Jvnq Hoje em dia nem legendas eles se import am em fazer, e comprar um jogo da nintendo já é

@amo_tae | 3 seguidores
@pandalala10 Nossa vei que vontade de pega o avião ir até lá pra proteger meu bebê. Que doente

@O_Abandonado | 22 seguidores
hoje é dia de @SaoPauloFC, dia de clássico, jog o difícil então vamos ser humilde, vai ser só 5x0 #

@Renatamariasil | 26 seguidores
@AbrahamMateoBSL estou muito bem ❤????m as a única coisa que queria agora é estar com me

@celso_cotta | 15 seguidores
@joicehasselmann movimento perdeu força. Divu lgação deixou a desejar dessa vez. ????

@whosjami | 44 seguidores
muita coisa pra uma pessoa só

@RodLoppea | 64 seguidores
CABEÇUDA AMOR DA MINHA VIDA mais uma v ez o brasileiro se superando ????????????????

@luiza_pim12 | 6 seguidores
#ArmysWillProtectJimin
NÃO ME IMPORTO QUE EU MORO NO BRASIL

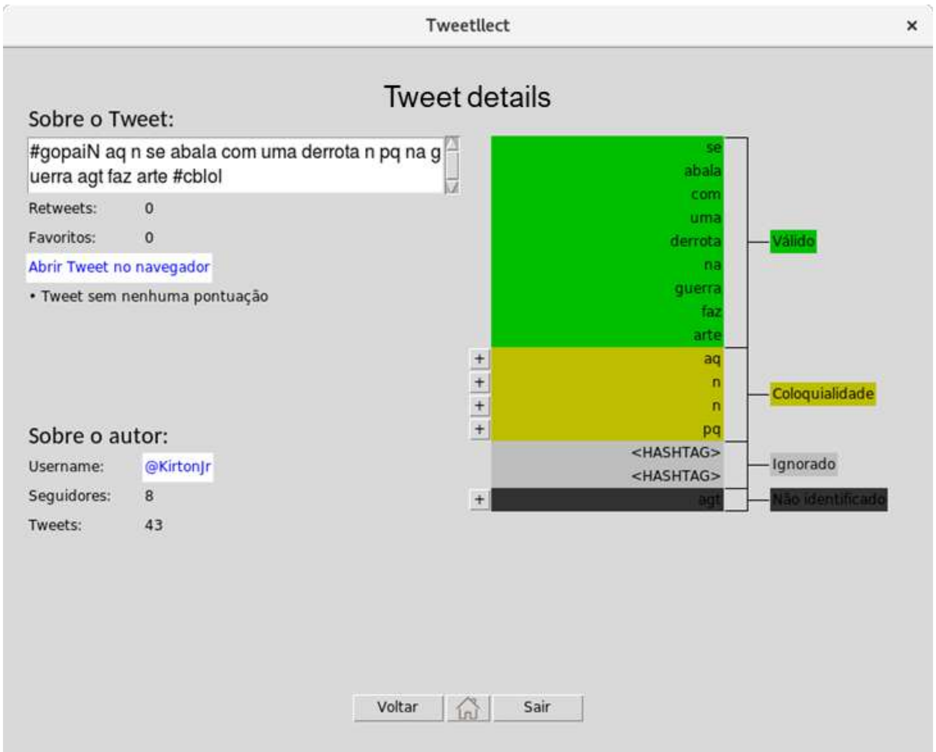
@jiminkjkj | 2 seguidores
Fiz essa conta porque ás vezes não consigo post ar no tumblr. Fav se pegou alguma coisa ❤

@ChristianGrossi | 26 seguidores
Tempo chuvoso pede um vinho. https://t.co/usRR k59pRO

1 | 217

Figure 5: Tweets listing screen.

The screen with a detailed tweet is shown in Figure 6. It lists all the parameters in relation to the tweet and the author. The tokens in this tweet are categorized. Tokens identified as colloquial can be added to the Portuguese or English dictionary while unidentified tokens can be defined as a colloquial term or can be entered in one of the dictionaries.



Tweetllect

Tweet details

Sobre o Tweet:

#gopaiN aq n se abala com uma derrota n pq na guerra agt faz arte #cblol

Retweets: 0

Favoritos: 0

Abrir Tweet no navegador

• Tweet sem nenhuma pontuação

Sobre o autor:

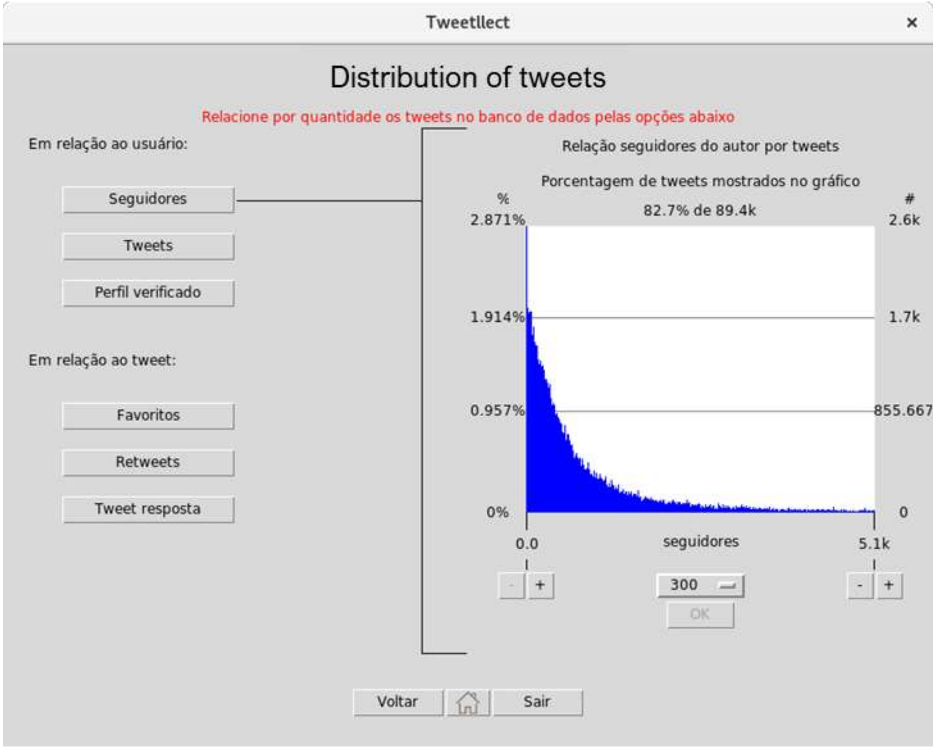
Username: @KirtonJr

Seguidores: 8

Tweets: 43

Figure 6: Tweet detail screen.

In Figure 6, valid tokens are grouped in green, tokens considered colloquial are grouped in yellow, ignored tokens are separated in gray, and unidentified tokens are colored in black. In the distribution screen (Figure 7) it is possible to check the tweets distribution, in relation to the author: by followers, by quantity of tweets and by verified or not profiles; already in relation to the tweet: by favorites, retweets, regular or reply tweet.



Tweetllect

Distribution of tweets

Relacione por quantidade os tweets no banco de dados pelas opções abaixo

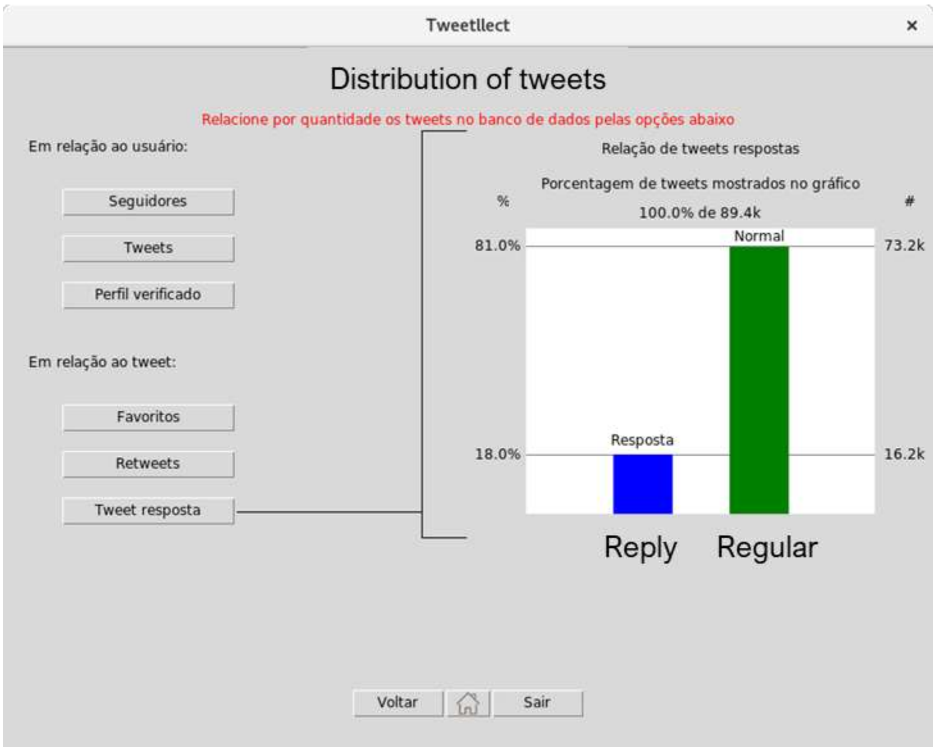
Em relação ao usuário:

Em relação ao tweet:

Figure 7: Screen showing the distribution of tweets by followers.

You can adjust the start and end point of the graph. In the example figure the range goes from zero to 5100 followers showing a total of 82.7% of all tweets stored in the graph, so 17.3% of tweets are out of the defined range of followers and it is necessary to expand the limits to show the totality of tweets.

Figure 8 shows the distribution of tweets as to whether it is a tweet reply or regular tweet. As it is a binary parameter the graph shows only two bars. The use of the bar graph is repeated in the parameter of verified or not verified profiles. For the other parameters the distribution is shown in the same way as in Figure 7.



Tweetllect

Distribution of tweets

Relacione por quantidade os tweets no banco de dados pelas opções abaixo

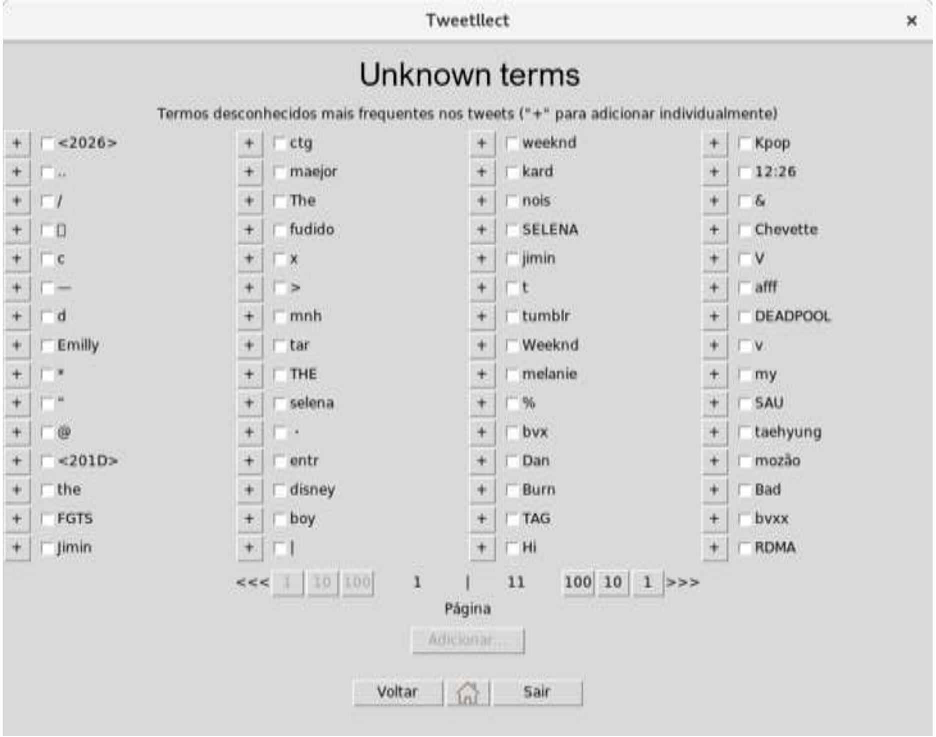
Relação de tweets respostas

Porcentagem de tweets mostrados no gráfico

100.0% de 89.4k

Figure 8: Screen showing the distribution by reply tweets or regular tweets.

Unknown terms, i.e., which are not identified by the algorithm, can be added in the Portuguese or English dictionary or defined as a colloquial term as seen in Figure 9. You can select several terms at once to speed up management. The terms are sorted according to distribution frequency, so the most frequent unknown terms are displayed first. There are ways to see tweets in which these terms are found for a better understanding of the context of their use



Tweetllect

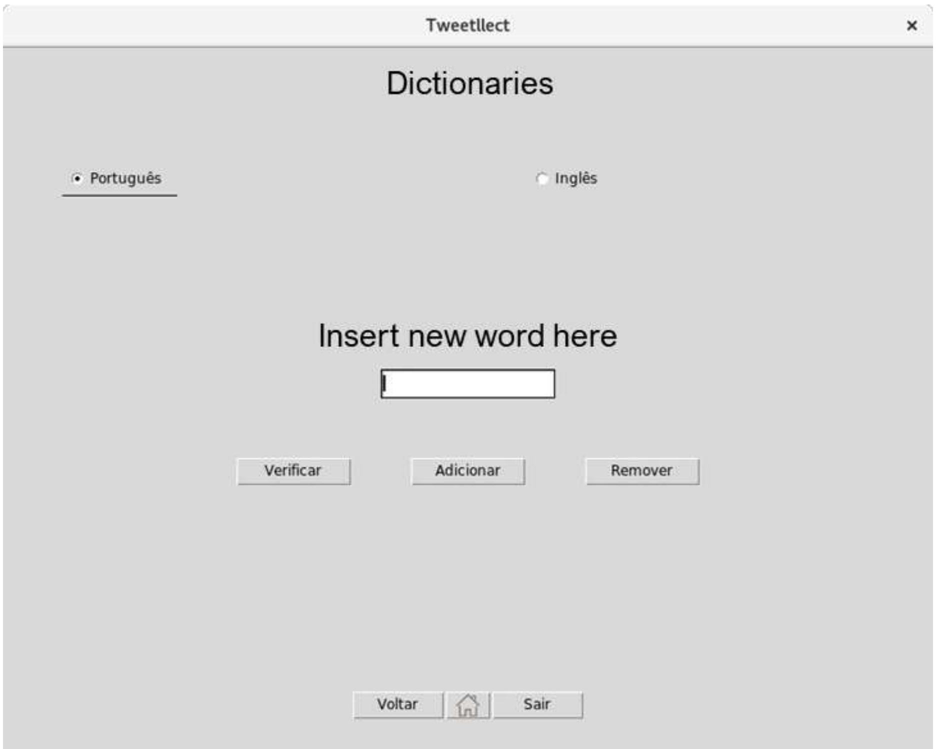
Unknown terms

Termos desconhecidos mais frequentes nos tweets ("+" para adicionar individualmente)

- + □ <2026>
- + □ ..
- + □ /
- + □ ▯
- + □ c
- + □ —
- + □ d
- + □ Emilly
- + □ *
- + □ "
- + □ @
- + □ <201D>
- + □ the
- + □ FGTS
- + □ Jimin
- + □ ctg
- + □ maejor
- + □ The
- + □ fudido
- + □ x
- + □ >
- + □ mnh
- + □ tar
- + □ THE
- + □ selena
- + □ ·
- + □ entr
- + □ disney
- + □ boy
- + □ |
- + □ weeknd
- + □ kard
- + □ nois
- + □ SELENA
- + □ jimin
- + □ t
- + □ tumblr
- + □ Weeknd
- + □ melanie
- + □ %
- + □ bvx
- + □ Dan
- + □ Burn
- + □ TAG
- + □ Hi
- + □ Kpop
- + □ 12:26
- + □ &
- + □ Chevette
- + □ V
- + □ afff
- + □ DEADPOOL
- + □ v
- + □ my
- + □ SAU
- + □ taehyung
- + □ mozão
- + □ Bad
- + □ bvxx
- + □ RDMA

Figure 9: Unknown terms management screen.

You can manage the dictionaries by adding, removing or checking if the word already exists (Figure 10).



Tweetllect

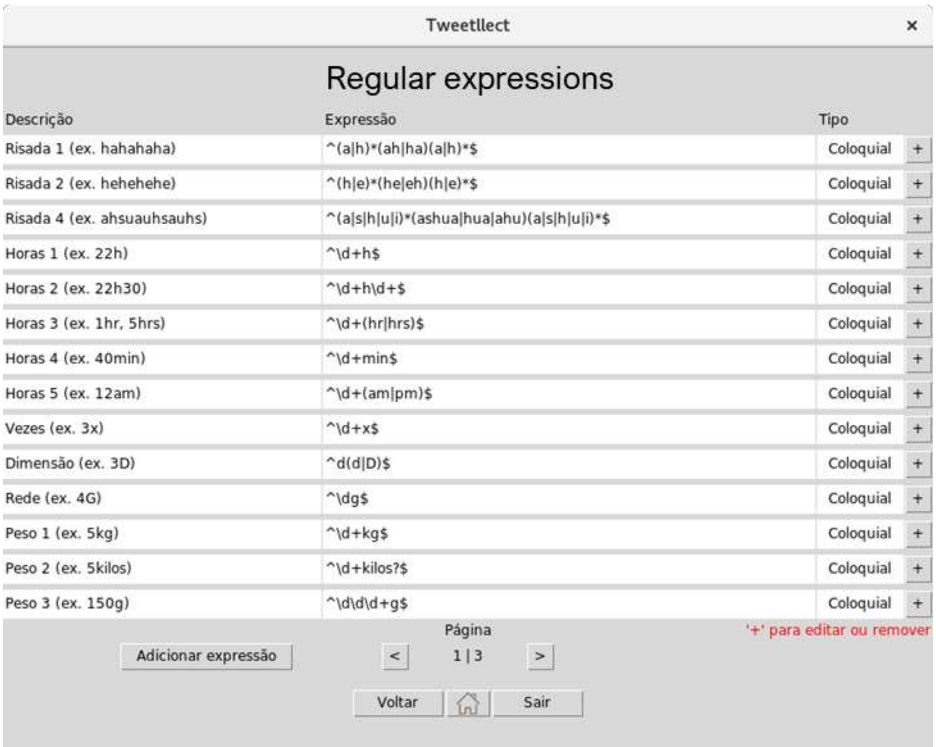
Dictionaries

● Português ○ Inglês

Insert new word here

Figure 10: Dictionary management screen.

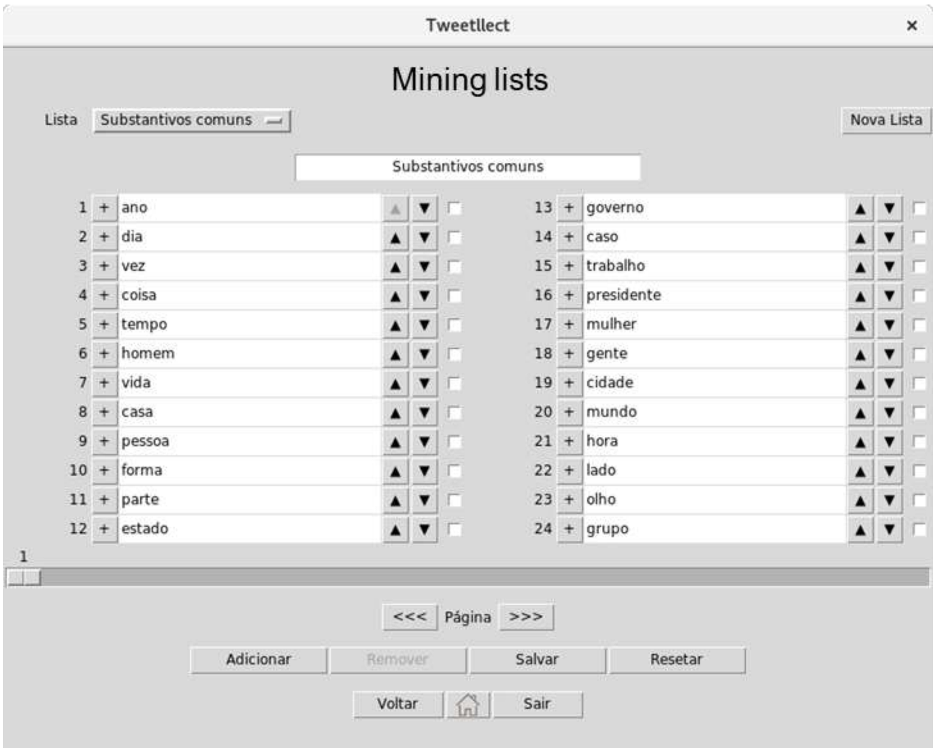
Figure 11 shows the regular expression management screen where you can add, remove, test, or edit the expressions. Tokens identified by these expressions can be defined as colloquial or valid.



| Descrição | Expressão | Tipo |  |
|---|---|---|---|
| Risada 1 (ex. hahahaha) | ^(a\|h)*(ah\|ha)(a\|h)*$ | Coloquial | + |
| Risada 2 (ex. hehehehe) | ^(h\|e)*(he\|eh)(h\|e)*$ | Coloquial | + |
| Risada 4 (ex. ahsuauhsauhs) | ^(a\|s\|h\|u\|i)*(ashua\|hua\|ahu)(a\|s\|h\|u\|i)*$ | Coloquial | + |
| Horas 1 (ex. 22h) | ^\d+h$ | Coloquial | + |
| Horas 2 (ex. 22h30) | ^\d+h\d+$ | Coloquial | + |
| Horas 3 (ex. 1hr, 5hrs) | ^\d+(hr\|hrs)$ | Coloquial | + |
| Horas 4 (ex. 40min) | ^\d+min$ | Coloquial | + |
| Horas 5 (ex. 12am) | ^\d+(am\|pm)$ | Coloquial | + |
| Vezes (ex. 3x) | ^\d+x$ | Coloquial | + |
| Dimensão (ex. 3D) | ^d(d\|D)$ | Coloquial | + |
| Rede (ex. 4G) | ^\dg$ | Coloquial | + |
| Peso 1 (ex. 5kg) | ^\d+kg$ | Coloquial | + |
| Peso 2 (ex. 5kilos) | ^\d+kilos?$ | Coloquial | + |
| Peso 3 (ex. 150g) | ^\d\d\d+g$ | Coloquial | + |

Figure 11: Regular expression management screen.

A mining list contains a few words that are used to mine tweets. These words can be specifically sorted if necessary. Figure 12 shows the management screen for these lists. The lists can be created, removed or edited.



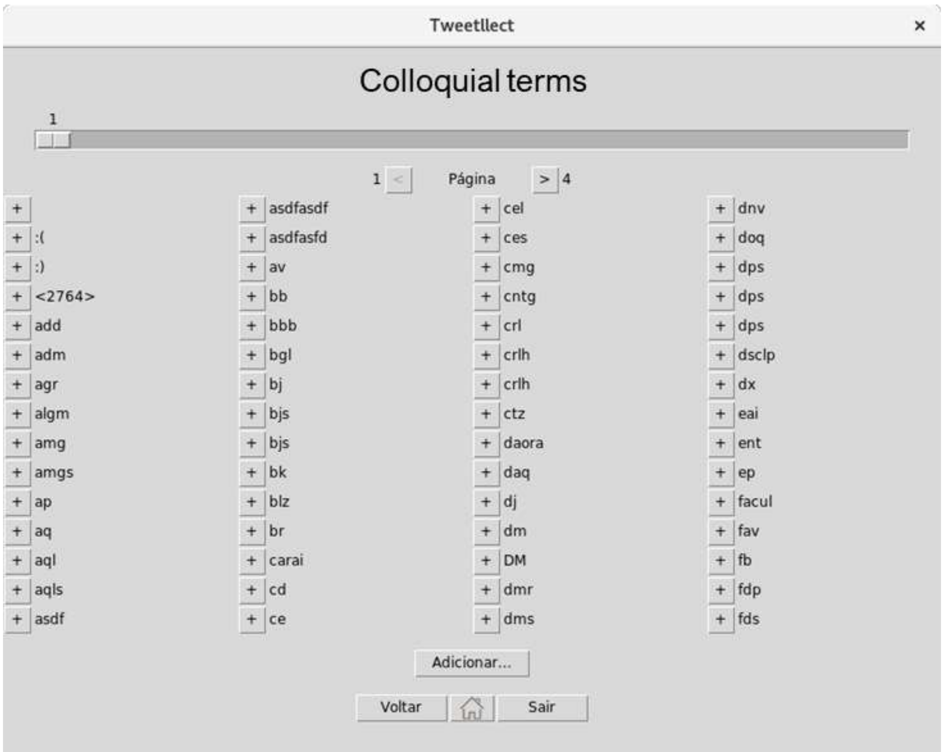
Lista: Substantivos comuns

Substantivos comuns

|  |  |  |  |
|---|---|---|---|
| 1 | ano | 13 | governo |
| 2 | dia | 14 | caso |
| 3 | vez | 15 | trabalho |
| 4 | coisa | 16 | presidente |
| 5 | tempo | 17 | mulher |
| 6 | homem | 18 | gente |
| 7 | vida | 19 | cidade |
| 8 | casa | 20 | mundo |
| 9 | pessoa | 21 | hora |
| 10 | forma | 22 | lado |
| 11 | parte | 23 | olho |
| 12 | estado | 24 | grupo |

Figure 12: Mining List Management Screen.

Figure 13 shows the colloquial term management screen, terms can be added, removed, or inserted into a dictionary.



Tweetllect

Colloquial terms

1

1 < Página > 4

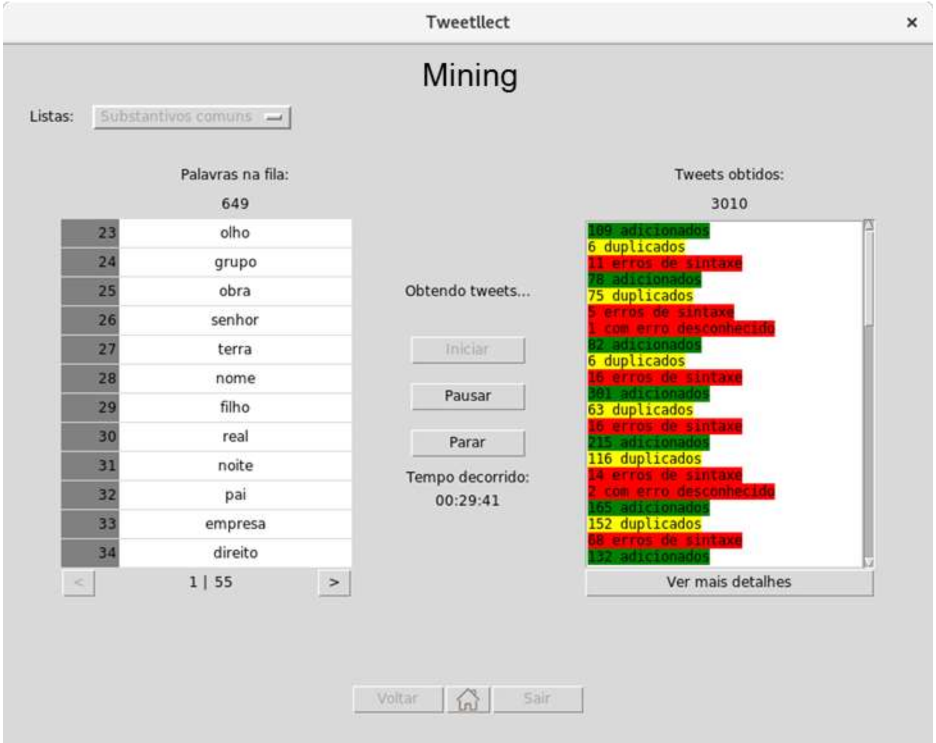
+
+ :(
+ :)
+ <2764>
+ add
+ adm
+ agr
+ algm
+ amg
+ amgs
+ ap
+ aq
+ aql
+ aqls
+ asdf
+ asdfasdf
+ asdfasfd
+ av
+ bb
+ bbb
+ bgl
+ bj
+ bjs
+ bjs
+ bk
+ blz
+ br
+ carai
+ cd
+ ce
+ cel
+ ces
+ cmg
+ cntg
+ crl
+ crlh
+ crlh
+ ctz
+ daora
+ daq
+ dj
+ dm
+ DM
+ dmr
+ dms
+ dnv
+ doq
+ dps
+ dps
+ dps
+ dsclp
+ dx
+ eai
+ ent
+ ep
+ facul
+ fav
+ fb
+ fdp
+ fds

Adicionar...

Voltar Sair

Figure 13: Colloquial term management screen.

The mining screen can be seen in Figure 14. When started mining the words in the list are used as keywords for tweets. Number of tweets obtained, which caused an error due to incompatibility of their content with the database or system coding, or which already exist in the database can be monitored on the side of the screen. You can see more details where all tweets are displayed individually and can be accessed.



Tweetllect

Mining

Listas: Substantivos comuns

Palavras na fila:

649

| | |
|---|---|
| 23 | olho |
| 24 | grupo |
| 25 | obra |
| 26 | senhor |
| 27 | terra |
| 28 | nome |
| 29 | filho |
| 30 | real |
| 31 | noite |
| 32 | pai |
| 33 | empresa |
| 34 | direito |

< 1 | 55 >

Obtendo tweets...

Iniciar

Pausar

Parar

Tempo decorrido:
00:29:41

Tweets obtidos:

3010

```
109 adicionados
6 duplicados
11 erros de sintaxe
78 adicionados
75 duplicados
5 erros de sintaxe
1 com erro desconhecido
82 adicionados
6 duplicados
16 erros de sintaxe
301 adicionados
63 duplicados
16 erros de sintaxe
215 adicionados
116 duplicados
14 erros de sintaxe
2 com erro desconhecido
165 adicionados
152 duplicados
68 erros de sintaxe
132 adicionados
```

Ver mais detalhes

Voltar Sair

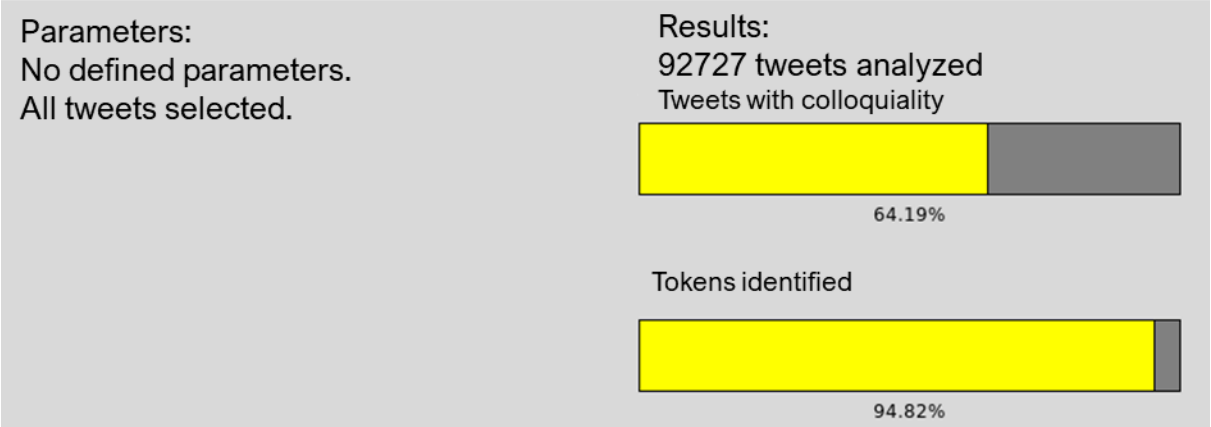
Figure 14: Tweets mining screen

## 4. Results Analysis

The results were obtained using the Tweetllect support system and using a total sample of 92727 tweets obtained between March and November 2017. According to the specific objectives the results will be discussed in this section.

### *4.1 Overall results*

We analyzed 92727 tweets, 64.19% of which have some colloquiality as seen in Figure 15. The identified token rate was close to 95%. This identification rate was constant ranging from 94% to 96% in all analyses.



Parameters:
No defined parameters.
All tweets selected.

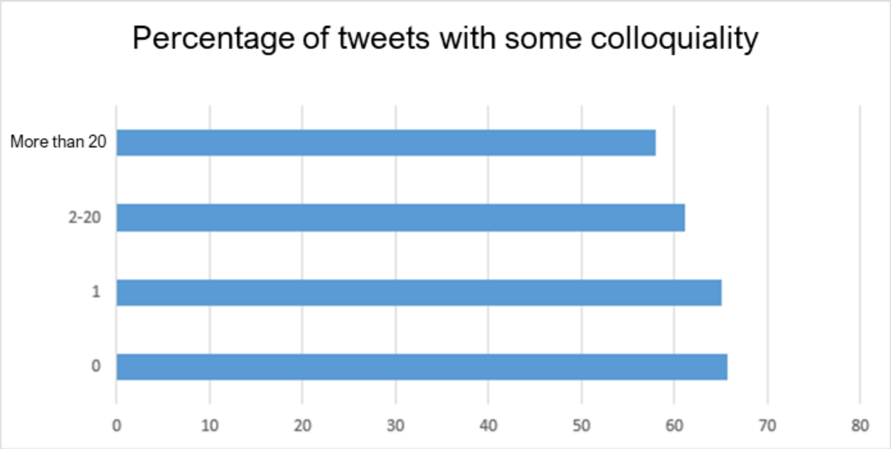
Results:
92727 tweets analyzed
Tweets with colloquiality
64.19%

Tokens identified
94.82%

Figure 15: Overall results.

### *4.2 Popularity and tweet replication*

How replication and popularity of a tweet considers its retweets and favorites. In Graph 1 we have the percentage of tweets with some colloquiality that have none, one, between 2 and 20 and finally more than 20 favorites.



Graph 1: List of favorites and colloquiality.

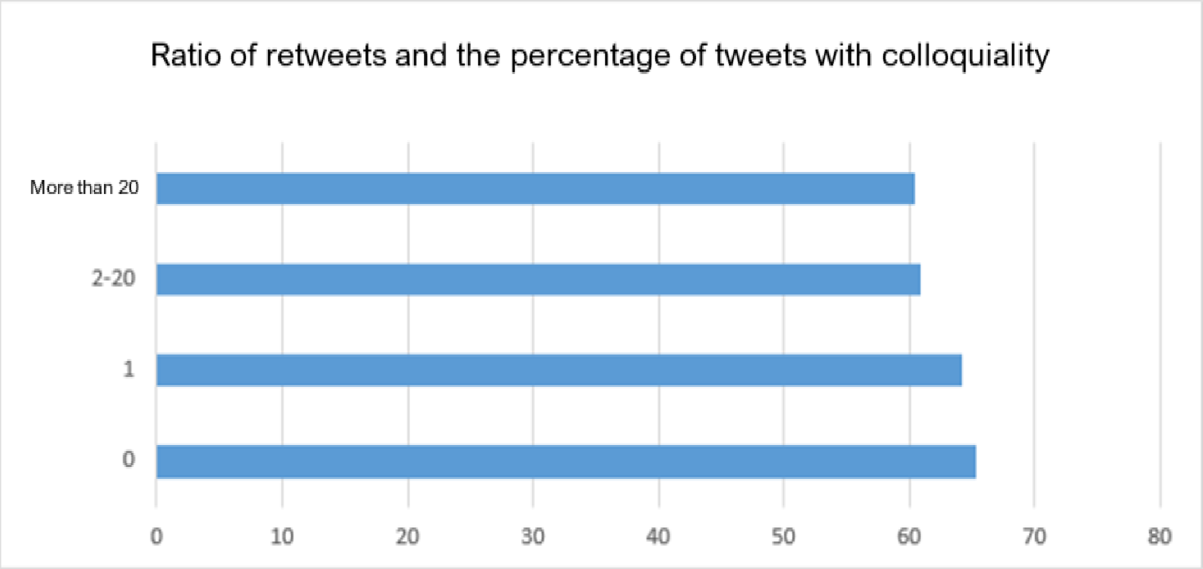
Tweets with fewer colloquial terms have a slight tendency to receive more favorites. About 65% of tweets with zero favorites have had some colloquiality already identified for tweets with 21 or more favorites this number drops to 58%.

Table 2 shows the results of Graph 1 along with information from the samples used. It is worth noting that the sample size for tweets with zero favorites is almost six times larger than the sample for tweets with more than 20 favorites but if we look at tweets with one favorite we have a close sample and a difference in the identification of tweets with colloquiality of 7.09%.

Table 2: List of favorites and colloquiality.

| Favorites | Sample size | Sample percentage (%) | Percentage of tweets with colloquiality (%) |
|---|---|---|---|
| 0 | 60.392 | 65,13 | 65,73 |
| 1 | 10.171 | 10,97 | 65,13 |
| 2-20 | 10.875 | 11,73 | 61,11 |
| More than 20 | 11.289 | 12,17 | 58,04 |
| Total | 92.727 | 100 | 64,19 |

Graph 2 shows the ratio of retweets and the percentage of tweets with some colloquiality. Similarly to the ratio of favorites and colloquiality (Graph 1), tweets with fewer colloquial errors receive more retweets.



Graph 2: Relationship of retweets and colloquiality.

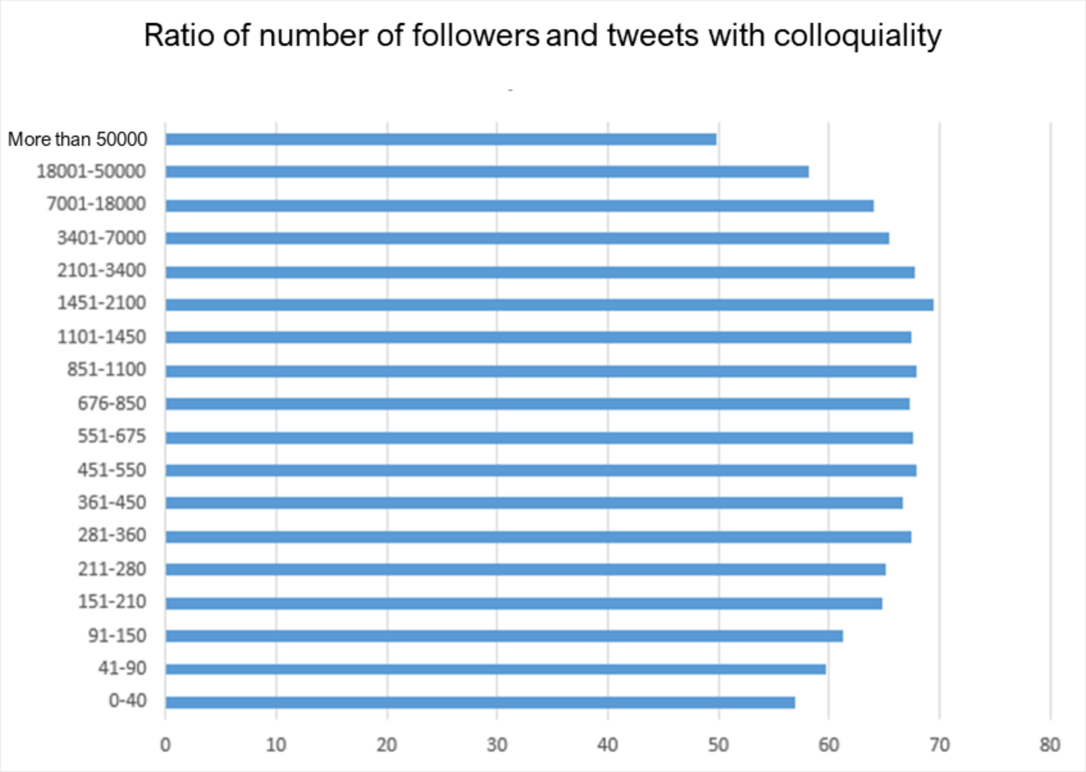
Tweets with zero retweets have an identified colloquiality rate of 65.37%, while tweets with more than 20 retweets 60.49%, or a difference of 4.88%. Again, the discrepancy between these two samples is high as can be seen in Table 3. Considering tweets with a retweet we only have a sample of near size and a difference of 3.67% in the identification of colloquiality when compared to tweets with more than 20 retweets.

Table 3: Relationship of retweets and colloquiality.

| Favorites | Sample size | Sample percentage (%) | Percentage of tweets with colloquiality (%) |
|---|---|---|---|
| 0 | 63.435 | 68,41 | 65,37 |
| 1 | 7.824 | 8,44 | 64,16 |
| 2-20 | 10.702 | 11,54 | 60,91 |
| More than 20 | 10.766 | 11,61 | 60,49 |
| Total | 92.727 | 100 | 64,19 |

### *4.3 User's Popularity*

As a user's popularity we consider the number of followers he has. Graph 3 shows several ranges of followers starting from zero to 40 followers and ending up with more than 50000 followers. The colloquiality rate starts low, 56.98%, for users who have 40 followers or less and grows up to the range of 281 to 360 followers. From there the rate remains constant, at about 67%, until it reaches the range of 7001 to 18000 followers where it begins to drop to a minimum of 49.79% for users with more than 50000 followers. What is noticeable is that accounts with few followers tend to have fewer colloquial errors, while users with followers ranging from hundreds to a few thousand lose this tendency but it returns to accounts with tens to hundreds of thousands of followers.



Graph 3: Ratio of number of followers and colloquiality.

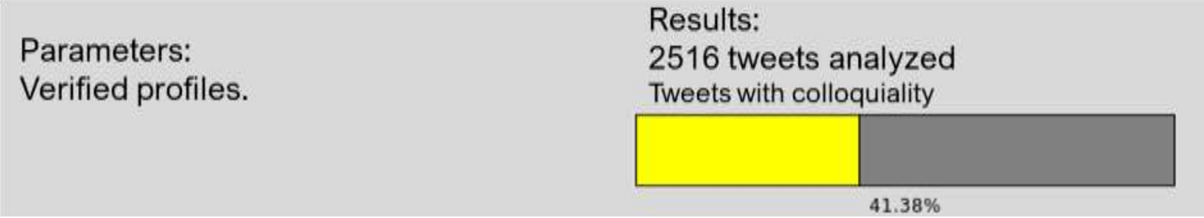
The following ranges were defined considering the sample size. Each interval averages 5000 tweets as shown in Table 4. This was done so that discrepancies in sample size would not influence the result.

Table 4: Relationship between the author's followers and colloquiality.

| Followers | Sample size | Sample percentage (%) | Tweet with colloquiality percentage (%) |
|---|---|---|---|
| 0-40 | 5135 | 5,54 | 56,98 |
| 41-90 | 5324 | 5,74 | 59,74 |
| 91-150 | 5599 | 6,04 | 61,21 |
| 151-210 | 5036 | 5,43 | 64,87 |
| 211-280 | 5396 | 5,82 | 65,14 |
| 281-360 | 5533 | 5,97 | 67,47 |
| 361-450 | 5154 | 5,56 | 66,63 |
| 451-550 | 4896 | 5,28 | 67,83 |
| 551-675 | 5049 | 5,45 | 67,64 |
| 676-850 | 5233 | 5,64 | 67,21 |
| 851-1100 | 5409 | 5,83 | 67,92 |
| 1101-1450 | 5158 | 5,56 | 67,45 |
| 1451-2100 | 5254 | 5,67 | 69,41 |
| 2101-3400 | 5183 | 5,59 | 67,76 |
| 3401-7000 | 5127 | 5,53 | 65,50 |
| 7001-18000 | 5132 | 5,53 | 64,11 |
| 18001-50000 | 3447 | 3,72 | 58,14 |
| More than 50000 | 5662 | 6,11 | 49,79 |
| Total | 92727 | 100,0 | 64,19 |

### *4.4 Verified accounts*

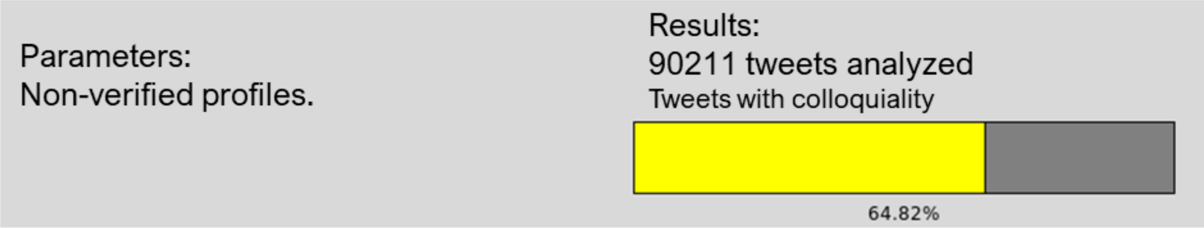
Profiles verified by Twitter are provided to users that it considers of public interest and mainly artists and celebrities, politicians, media and companies get this verification. The colloquiality rate for verified profiles is only 41.38% as seen in Figure 15. Even considering the discrepancy in sample size the difference of an additional 20% between verified and unverified profiles (Figure 16) shows a remarkable trend for verified profiles to avoid colloquiality.



Parameters:
Verified profiles.

Results:
2516 tweets analyzed
Tweets with colloquiality

Figure 16: Verified profiles and colloquiality ratio.



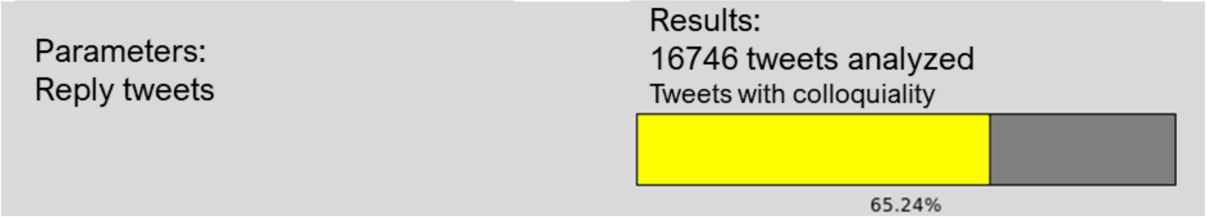
Parameters:
Non-verified profiles.

Results:
90211 tweets analyzed
Tweets with colloquiality

Figure 17: Not verified profiles and colloquiality ratio.

### *4.5 Tweets replies*

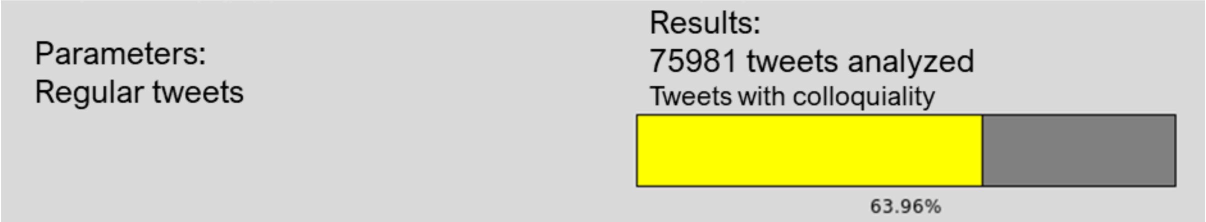
Tweets replies are tweets that respond directly to another tweet and are not directed to the author's followers. The rate of colloquiality is slightly higher, only 1.28%, for tweets replies when compared to tweets that do not respond. Showing a tendency to use colloquiality when one is not addressing all followers but only one person.



Parameters:
Reply tweets

Results:
16746 tweets analyzed
Tweets with colloquiality

Figure 18: Tweet reply and colloquiality ratio.



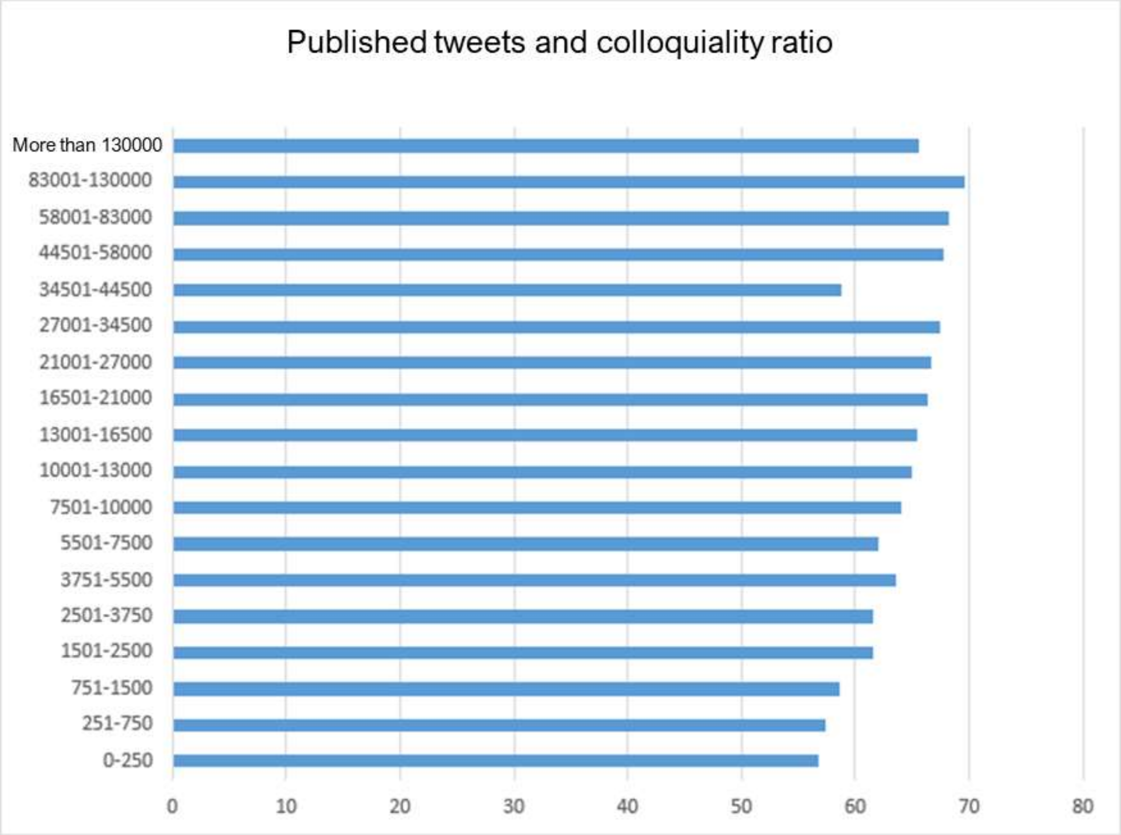
Parameters:
Regular tweets

Results:
75981 tweets analyzed
Tweets with colloquiality

Figure 19: Regular tweets and colloquiality.

### *4.6 Number of tweets published*

Graph 4 shows the relationship between the number of published tweets and the rate of tweets with colloquiality.



Graph 4: Published tweets and colloquiality ratio.

It can be noted that the more tweets the user publishes the greater his tendency to use colloquial terms. The range of published tweets was made so that each sample had approximately 5000 tweets as shown in Table 5 so that discrepancies in sample size would not interfere with the results.

Table 5: Published tweets and colloquiality ratio.

| Tweets | Sample size | Sample percentage (%) | Tweets with colloquiality percentage (%) |
|---|---|---|---|
| 0-250 | 5084 | 5,48 | 56,83 |
| 251-750 | 5008 | 5,40 | 57,39 |
| 751-1500 | 5382 | 5,80 | 58,58 |
| 1501-2500 | 5371 | 5,79 | 61,57 |
| 2501-3750 | 5172 | 5,58 | 61,48 |
| 3751-5500 | 5620 | 6,06 | 63,54 |
| 5501-7500 | 5221 | 5,63 | 62,02 |
| 7501-10000 | 5242 | 5,65 | 64,02 |
| 10001-13000 | 5042 | 5,44 | 64,99 |
| 13001-16500 | 5065 | 5,46 | 65,37 |
| 16501-21000 | 5024 | 5,42 | 66,38 |
| 21001-27000 | 5260 | 5,67 | 66,65 |

| Tweets | Sample size | Sample percentage (%) | Tweets with colloquiality percentage (%) |
|---|---|---|---|
| 27001-34500 | 5130 | 5,53 | 67,50 |
| 34501-44500 | 5174 | 5,58 | 58,82 |
| 44501-58000 | 4960 | 5,35 | 67,74 |
| 58001-83000 | 5291 | 5,71 | 68,13 |
| 83001-130000 | 4674 | 5,04 | 69,62 |
| More than 130000 | 5007 | 5,40 | 65,51 |
| Total | 92727 | 100,00 | 64,19 |

## 5. Conclusion

The analyzes carried out show that there is a relationship between the use of colloquialities in tweets and several factors inherent to the social network Twitter. For example colloquialities versus popularity, colloquialities versus verified accounts, colloquialities versus number of followers, colloquialities versus number of published tweets.

Regarding the relationship between popularity and replication of a tweet considering the occurrence of colloquialities, we can point out that we observed that tweets without colloquialities have a positive trend towards greater popularity. Meanwhile, tweets with colloquialities tend to be less popular. Tweets with few colloquial terms have a slight positive tendency to be marked as a favorite.

Regarding verified accounts, we can see that verified users (artists, sportsmen, politicians, and companies), publish fewer tweets with colloquiality (about 41%) than unverified users (about 65%). Showing that personalities, or companies, that have a public image to care for, are concerned with the correct spelling of their published tweets, avoiding the abundant use of colloquialities.

The more followers a user has, the more he tends to use colloquialities in his tweets. This trend occurs up to a range of 0 to 2000 followers. From 2000 followers onwards, the greater the number of followers of a user, users tend to use a few colloquial terms.

Regarding the number of tweets published per user, we can see that the greater the number of tweets published, the more the user tends to use colloquial structures. This shows us a lack of concern with formalities with the social network over time.

In summary, we realize that the use of colloquial terms in tweets is common. However, its use is more frequent in accounts with few followers, and from unverified users. Other studies can be derived from this paper, for example, the verification of colloquialities versus gender of the user; colloquialities versus specific hashtags (considering specific facts and events); colloquialities versus sentimental analyses; and studies dealing with colloquialities and languages (considering other languages besides Portuguese).

## 6. References

[1] Boyd, D.; Golder. S; Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In HICSS-43. IEEE: Kauai, HI, January 6. 2010.

[2] Davies, M.; Preto-Bay, A. A Frequency Dictionary of Portuguese (Routledge Frequency Dictionaries). 1. ed.: Taylor Francis Ltd. 2008.

[3] Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification using Distant Supervision. In CS224N Project Report, Stanford. 2009.

[4] Gouws, S.; Metzler, D.; Cai, C.; Hovy, E. Contextual Bearing on Linguistic Variation in Social Media. Proceedings of the Workshop on Language in Social Media (LSM 2011). 2011.

[5] Hagen, L.; Uzuner, O.; Harrison, T. M.; Katragaddda, S. E-petition popularity: Do linguistic and semantic factors matter?. Government Information Quarterly. 2016.

[6] Internet Live Stats, Twitter Usage Statistics. Retrieved from: <http://www.internetlivestats.com/twitter-statistics/>, accessed 30 January 2020.

[7] Liu, Kun-Lin and Li, Wu-Jun and Guo, Minyi, Emoticon Smoothed Language Models for Twitter Sentiment Analysis., AAAI, 2012.

[8] Nguyen, D.; Gravel, R.; Trieschnigg, D.; Meder, T. How Old Do You Think I Am? A Study of Language and Age in Twitter. ICWSM. 2013.

[9] Perez-Sabater, C.; The linguistics of social networking: A study of writing conventions on Facebook. Linguistik online, 56(6/12):81-93. 2012.

[10] Souza, L. P.; Deps, V. L. A linguagem utilizada nas redes sociais e sua interferência na escrita tradicional: um estudo com adolescentes brasileiros. Proceedings II Congresso Internacional TIC e Educação, Portugal. 2012.

[11] Statista, Statistics and facts about social media usage. Retrieved from: <https://www.statista.com/topics/1164/social-networks/>, accessed 30 January 2020.

[12] Statista, Statistics and facts about Twitter. Retrieved from: <https://www.statista.com/topics/737/twitter/>, accessed 30 January 2020.

[13] Statista. Social media usage worldwide. Retrieved from <https://www.statista.com/study/12393/social-networks-statista-dossier/>, accessed 30 January 2020.

[14] Twitter. Retrieved from: <https://twitter.com/>, accessed 30 January 2020.

[15] UNFPA. State of World Population 2019. New York: UNFPA Press; 2019. Retrieved from <https://www.unfpa.org/swop-2019>, accessed 30 January 2020.

[16] Yagui, M. M., Maia, L. F. M. P., Oliveira, J., & Vivacqua, A. S. Data mining of social manifestations in Twitter: analysis and aspects of the social movement “bela, recatada e do lar” (beautiful, demure and housewife). Journal of Computer Science, 17(1), 23-37. 2018.