Querying on Google Sheets

Designing a Sentiments Analysis Alternative for rating tweets regarding the Ecuadorian 2021 Presidential Campaigns

Sariah López-Fierro¹, Carlos Chiriboga², and Rubén Pacheco³

 ¹ Universidad T. F. Santa María, Chile sariah.lopez@sansano.usm.cl
² Soluciones Wandarina S. A., Ecuador carlos@wandarina.com http://www.wandarina.com
³ Universidad de E. Espíritu Santo, Ecuador rpachecov@uees.edu.ec

Abstract. This document contains an approach for the implementation of a sentiment analysis alternative after using Google Sheets for the rating of tweets retrieved with respect to the Ecuadorian Presidential 2021 Campaign.

Keywords: Google Clouds \cdot Google Sheets \cdot Querying \cdot Data Analysis \cdot Sentiment Analysis.

1 Introduction

Since 2004, when "Web 2.0" was loosely expressed and defined at a Silicon Valley Conference [1], software technologies worked on applications to improve the interaction and collaboration between the users, via the web [2].

During those years, Google not only managed to become the most popular search engine within the web [4], but it also evolved after introducing solutions such as Google News (2002), Gmail (2004), Google Maps (2005), among others; and expanding their business after buying popular web platforms such as Blogger (2003), Youtube (2006), Upstartle (2006), among others. Thus becoming in the "symbol of online innovation" [2].

In 2006, when Google Sheets was recently launched, some publications compared it to Microsoft's office Excel. While its innovative online collaboration was highlighted, limitations in presentation, features, and formats were also noted [2][5].

About 15 years after its release, Google Sheets is not only a collaborative sheet. Along the mathematical functions common in other spreadsheets, it also allows users to "program" or write custom functions through the Script Editor, which it is a JavaScript platform; and to "query" similar to SQL environments, which makes it capable of processing petabytes of data [3].

Being Google Sheets part of Google Cloud, thus being backed up by the power of High Performance Computers, this article aims to describe an alternative platform for analysing data available for any user, while describing its implementation to process and interpret sentiment analysis from tweets retrieved during the Ecuadorian 2021 Campaigns.

2 Methodology: Querying on Google Sheets

2.1 Retrieving Tweets

To retrieve the tweets, we used Martin Hawksey's TAGS¹. An available online script that allows you to easily set up a Twitter account and, through the Twitter API, collect the tweets. The only change that we made to his work was the addition of more fields in the "Archive" sheet that would allow us to obtain more details for a deeper analysis.

M. Hawksey's fields are: id str; from user, text, created at, time, geo coordinates, user lang, in reply to user id str, in reply to screen name, from user id str, in reply to status id str, source, profile image url, user followers count, user friends count, user location, status url, entities str.

Our fields were: id str, text, status url, retweet count, favorite count, created at, user created at, in reply to screen name, user geo enabled, user location, place, in reply to user id str, in reply to status id str, user id str, user name, from user, user profile image url, entities str, lang, user description, user followers count, user friends count, user favourites count, user statuses count, source, extended tweet.

2.2 Organizing Data

Google Sheets limits the use of its cells per document up to 5'000.000. Therefore, we were unable to work on the original TAGs file alone. Hence we established that each time we reached approximately the 100,000 row (100.000 tweets with details of the aforementioned fields), we would copy the data into a new spreadsheet.

Thus our step to organize the data involved to separate the tweets up to 100.000, in different files (from now on we will refer to them as *1-Archive*), also put them in different folders, consequently being able to work with them independently and then combine the results into a final step. Theoretically using the Divide-and-Conquer technique approach.

2.3 Cleaning Data

Our data was a mix of original tweets, retweets, and duplicate tweets that were separated into different files. As a result, in each one, we started by removing the duplicates through the *unique* command and then calling them to a new file (from now on we will refer to these new files as *2-Filter*).

Even though Google's "Sheets data connector for BigQuery" [3] would allow us to deal with large datasets "at once"; another way around to stick only in Google Sheets, was to use *importrange* command. Since this command works always with

¹ https://tags.hawksey.info/get-tags/

calls of 10.000 rows, we used it multiple times to import the 100.000+ rows from each *1-Archive* to fill the *2-Filter*.

Then, we separated the original tweets on a new sheet of the same document. From this new sheet, we used the *query* command.

2.4 Designing a Sentiment Analysis Alternative on Google Sheets

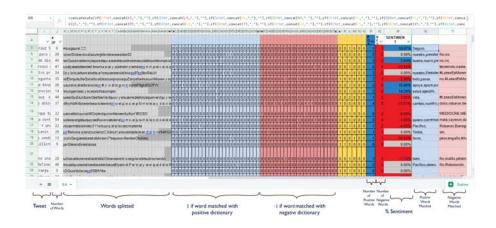


Fig. 1. Google Sheets - Screenshot of the Distribution of cells for analyzing tweets

Our approach for analyzing tweets in spreadsheets was to compare each word from each tweet against a dictionary of positive words and another of negative words. Add a positive value to each positive word found, or a negative value to each matching negative word, then calculate them. Besides, in order to get more accuracy in the rating, we decided to vary the value of the words from the dictionaries.

The final result determined the sentiment of each tweet. The higher the percentage of positive words, assures a positive tweet; and likewise, the lower the percentage of negative words identified a negative sentiment.

Besides, when a tweet had a result of "0", we marked it as "Neutral" and it meant that a same number of positive and negative words were found in the tweet. Or that none of its words matched with any vocabulary included in the dictionaries.

Setting up the dictionaries We created a new file (from now on we will refer to it as *Dictionaries*) with two sheets. One sheet contained all the positive words, and the other the negative words. We only used one column from each sheet and filled them with all the respective words. As these dictionaries were made in Spanish, and we were about to match exact words, our dictionaries sought to include possible conjugations and variations with accent and without accent.

Matching Words In a new spreadsheet (from now on we will refer to it as *3-SA*), we located the original tweets. The procedure described in this section was implemented in cells that were part of the same text row to be analyzed. To make clearer our explanation, figure 1 shows the distribution of the cells within the file.

First, we started by counting the words included in each tweet.

Second, we split the words of the tweets in different cells. In figure 1, these cells are grey colored.

Third, we added new cells for comparing each word of the tweet (up to this step, the words from the tweets were already spread across different cells) with the positive dictionary and each with the negative. Thus, for example in a tweet with 70 words, there will be at least 70 new more cells for comparing each word with the positive dictionary, and 70 new more cells for comparing them to the negative dictionary. Each time that a word matched with the positive or negative words, the number "1" was set in the respective cell, in figure 1 these cells are colored with blue for the positive and red for the negative columns.

2.5 Results

As we mentioned earlier, positivity and negativity were initially calculated as a percentage, increasing the probability of assertiveness when these values were further from 0.

These ratings may allow us to have a general glance of the sentiment shared through Twitter, regarding the Ecuadorian 2021 Presidential Campaigns. Hence, as a matter of example, and in order to attempt to get a more meaningful result. After standardizing the locations shared by the users, we were also able to obtain the sentiment through places, see figure 2 which displays the results from February 1st to the 7th. In this figure it is possible to observe that in Pichincha, for instance, there are more positive tweets towards Lasso (374) than against him (305). While, there are more negative content rated towards Arauz (437), than positive tweets (336).

	. 8	4	1	. R.	. *			F . A	. 9.	2		н.	. Alt .	35	AL	18.	12 4 41	AM.	- 44	40	48	- 超	48	45	AT	- 44	41	48
				7643	6578	1417	3381	1833	1368	281	641	4119	271	338	40	121	878	1384	1459	201	894	4135	412	873	1	32	3	6
1		1	1 il and 1	100								ARAUZ					10000						LASSO					
ŧ	Fesh T	Región		Past 7	Negat **	Moto *	Neutr 7	A+ *	A 7	A+. *	A 7	Teta *	sA+ T	sA- T	sAt T	14 7	Tota 7 7	L* . T	10 T	Les Tra	4	T	sL+ **	4-7	LL# T	LLe T	Ma	LMa
۹.,	1-7807		Counter-Asvey	170	110	31	01	92	-49	. 11	22	144	10	. +	Q			79	41	11	49	180	- 22		0	0	. 0	
	1-7404	Sera	Equator/Bolvar			2	2	3	2	0	2	1.2	2		ĝ		2		0		. 0	5	3	- 0		0		
8	5-7abr	Sens	Equador Cañar	3	. ,	0	1	3	. 0	0	0	3	0	. 0	0		0	- 0			1	2	0	0	0	0	. 0	
Ħ.,	1-7407	Sere	Exuador Caroni			- 0	1			0	- 4	2	0		¢	1	0		0		1	2	.0	0.		0		
4	1-Tabr	Sens	Equador Chimborade	27	54		7	6		- 2	2	- 20	0	2			- 4	-13	6	- 2	4	24	. 7			0	- 0	
4	1-7400	Sera	Envelor/Company		24	. 3	0	- 2	20	2	- 0	24	0.					. 1	2	2	0			0		¢.		
4	1-Tabr	Costs	Equator 8_0rp	- 40	. 90	- 4	28	21	. 6	2	. 11	40	6	0.	Ô	() (7	10	20	2	17	. 55	.7	3	. 0		- 0	
4	1-7807	Costa	Equador Esmeraldas	28	. 10	1		18		. 0	4	27	4		0	1. 1	6	14	5	. 0	. 5	24	. 6	8	8	0	0	
8	1-7807	insular.	Ecuador, Galagages	. 7	. 0		5	2	. 0	0	3	5	. 8	0	0		1		0		- 2	3	0	0	0	.0	0	
2	1-7801	Costa	Equador Quayes	727	050	129	252	272	278	45	122	730	28	55		2	130	340	206	45	217	915	- 95	40	-1	4	1	
8	1-7abr	Same	Equador, Imbabura	20					6	2		21			1		2	7		2	3	20	1	1		0	- 2	
	1-7404	Sera	Equador Litua	50	31	14	17	29	15			60	12	0		1	54		11			34		2	0	0	0	
4	1-Tabr	Costs	Equador Los_Ros	23		2	4	16	4		3	28	2	. 0	0	1	2	2	7		1	41		1	0	۴.	.0	
1	1-Talpr	Costa	Envedor Manabi	114	88	25	70	47	- 29		12	82	11	7	0	1	22	52	#2		50	100	.20	4		2	- ÷	
4	1-Talor	Oriente	Equador Morona, Samage			- 3		1		1		3	\$. 0	0	- 4	0	- 2	- 2		1		۵.	0		0	- 0	
4	1-7507	Oriante	Equador Napo	1		- 2		4	1	0	1	. 6	0	. 0	0		0	2	2		Ó	- 4	1	0	0		0	
4.1	1-Tater	Oriente	Enuador, Oraliana	1	4	2	1		2	1			0		0	1	1	1	2		0	4	0	1	0	0	0	
4	1-Tater	Onerte	Envelor/Pastata	1	1	1	20	1	1	0	1	- 3	0	- 0	0	1	0	1	1	- 0	2		0	0	- 0	0	- 0	
	1-Tabr	Sens	Equador Pichingha	923	250	210	623	338	427		128	1043	42	71		4	173	374	208	- 04	337	1100	69	40		. 6-		
81	1-7404	Ceala	Bouetor,Sanse_Elena	7	. 10	2	4	. 4		1	-0		2	. 0		1 4	3	. 1	6	. 1		.12	10	1		1	. 0	
4	5-Tabr	Costs	Equador Santo, Domingo, Taachia	18 10	15	- 2			10	2	4	-25	7	2	0		3	3		- 2	. 5	92	0	2	0	1		
	1-Talpr	Oriente	Exuator Supursia		2					0	- 4		1		¢	1		3	£.		1		.0	1		0		
	9-7abr	Sera	Equador Tungurahua	23	24		22	12	54	- 3	.7	38	4	3	0			7	4	3	13	27		0	0	0		
6	1-Tale	Oriente	EnvelorZamora_Chinohoa	4	4	. 1		3		1	1			. 0	0	6	2	2	2	1	0		0		- 5	0		
1	1.7400	Pair	ARCA.	54		- 23	42	10	47	= 10	24	117	7	12			28	37	43	+0	- 3.5	112				0		

Fig. 2. Rating tweets according to location

2.6 Limitations in the analysis

These are some of the limitations that we were able to find in our approach.

- To keep the document as small as possible (the fewer cells the better), we limited the analysis up to the 70th word of a tweet. If a tweet had more words than that, we ignored them.
- Ironies expressed with text or emojis were not well rated.
- If there was a tweet that replied an user, but mentioning a candidate, it was marked the sentiment towards the candidate.

3 Conclusion and Future Work

Analyzing data is mostly related to technologies that involved advanced technical support and constant rent in available public clouds. Our work included a different alternative that may respond to unattended parties who are more familiar with Google Suites.

Furthermore false-positives results allowed us to notice the limitations of measuring the sentiment with our approach. Thus for increasing the accuracy of analysing the tweets we have considered to measure the sentiment not only by words, but also in some cases by "phrases"; to determine sentiment according to context, to attempt to discriminate trolls from real accounts.

Acknowledgement

This research has been supported and funded by the Technological Research Department of the company Soluciones Wandarina S. A. of Ecuador.

References

- 1. Yu, C., and Du, H. (2007). Welcome to the World of Web 2.0. The CPA Journal, 77(5), 6.
- 2. Rienzo, T., and Han, B. (2009). Teaching Tip: Microsoft or Google Web 2.0 Tools for Course Management. Journal of Information Systems Education, 20(2), 123.
- 3. Gundrum, D. (2019, January 15). Connecting BigQuery and Google sheets to help with hefty data analysis. Retrieved March 15, 2021, from https://cloud.google.com/blog/products/g-suite/connecting-bigquery-and-google-sheets-to-help-with-hefty-data-analysis
- 4. Evans, M. P. (2007). Analysing Google rankings through search engine optimization data. Internet research.
- 5. Firth, M., and Mesureur, G. (2010). Innovative uses for Google Docs in a university. Jalt call journal, 6(1), 3-16