Sentiments and Opinions From Twitter About Peruvian Touristic Places Using Correspondence Analysis

Luis Cajachahua
Universidad Nacional de Ingeniería
lcajachahua@gmail.com

Indira Burga
Universidad de Ingeniería y Tecnología
indira.burga@gmail.com

Abstract

Tourism in Perú has become very important, since there is a growing number of tourists arriving each year. This paper focus in understand what do speaking-english tourists have in consideration when they visit Perú. We obtained all the tweets published in english during the year 2016, filtered by touristic places visited. In total, more than 192 thousand tweets were collected. We performed different analysis to describe the data, including correspondence analysis, a statistical technique which is normally applied to categorical data. The goal was to understand the sentiments and opinions expressed in those tweets.

Keywords: Twitter, Tourism, Sentiment Analysis, Text Mining, Correspondence Analysis, Perú.

1 Introduction

Tourism is an important source of economic growth in Perú. In 2015, 3.5 million of tourists visited Perú, leaving more than USD 4,151 million that correspond to the 3.75% of the gross domestic product (Promperú, 2016). For this reason, all the government tourism bureaus, specially Promperú, are working hard to keep those numbers growing on Figure 1.

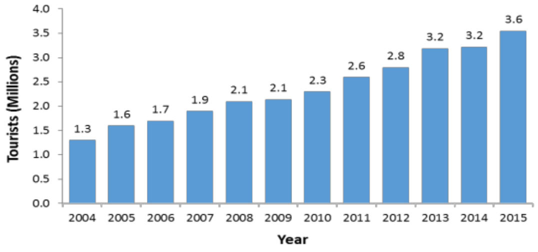


Figure 1: Tourist arrivals to Perú in the last 10 years (Promperú, 2016).

In this context, every initiative oriented to understand the tourist preferences and sentiments is very valuable, because it allows to discover and develop the main attributes of touristic places in Perú, considering not only the place itself, but all the ecosystem: hotels, touristic agencies, handicraft stores, transportation, public services, etc.

For that reason, we downloaded 192,525 tweets, published during 2016, considering only english language. After a careful data preparation and feature extraction, we applied some sentiment analysis techniques to tag each tweet using the following eight sentiments: anger, fear, sadness, disgust, surprise, anticipation, trust and joy, proposed by Plutchik (Mohammad and Turney, 2010). Using correspondence analysis, we discovered some associations between touristic places visited and sentiments expressed in the extracted tweets.

In addition, we used other text mining techniques to determine some general concepts mentioned on tweets and identify the association between the places visited and these concepts.

2 Background

Text mining has become a useful tool to deal with the data deluge. With few and simple tools, we can now extract valuable insights from a big amount of unstructured data available today in the internet. For example, one of the first studies of sentiment analysis in Twitter was developed in 2011 by Agarwal et al. They built a formula to calculate the polarity of a comment, if the tweet is positive or negative. They developed a tree kernel approach to improve the classification (Agarwal, et al., 2011).

In the same year, Dodds et al analyzed a huge list of tweets, around 4.6 billion, posted by over 63 million individual users and extracted in a 33-month span (Dodds, et al., 2011). The main goal was to develop a dictionary optimized to measure the level of happiness of a given text, e.g. another tweet. They named their solution as "The Hedometer"

One of the first studies about tourism using tweets was developed by Antoniadis. This study was focused in verifying the correlation between Twitter performance and tourism competitive index for 38 destination management organizations (DMO's). They found that Twitter use was in accordance with countries' tourism performance (Antoniadis, et al., 2014).

Oku et al, analyzed 4.5 million tweets to find the location where a tourist posted a photo. They used only geotagging information from the pictures, leaving text data out of the analysis (Oku, et al., 2015).

In addition, Bassolas performed an interesting analysis about the 20 most interesting touristic places around the world, including Machu Picchu. They tried to measure the "attractiveness" of each place, considering two metrics: a) Radius, defined as the average distance between the places of residence and the touristic site and b) Coverage, the area covered by the users' places of residence computed as the number of distinct zones (or countries) of residence (Bassolas, et al., 2016).

All these studies were considered to develop a different approach, focused on finding the emotions expressed by the tourists in the visited places.

2.1 Objectives

This research was focused in five goals:

- Understand the sentiment expressed in the tweet (positive or negative) about the place.
- Tag each tweet using the sentiments that the user is expressing in the tweet
- Identify the main qualitative attributes mentioned in the tweets, related to the touristic place.
- Analyze the possible relationships between the places and the identified attributes.
- Explore the general activity of the tourists which use twitter (frequent, eventual).

Having those objectives in mind, we determined the limits of our efforts, given the nature of the information source and the tools used.

2.2 Limitations

There are various limitations to be considered:

- Being Twitter an internet social network, all the users must be registered and logged to use it. But not every person or tourist in the world has a Twitter account. For this reason, results are not generalizable/representative for the tourists' population.
- We acquired all the tweets directly from Twitter. We didn't used streaming or scraping methods to download the tweets. But, we used many filters to select tweets referred to touristic places. In consequence, we could have selected a non-representative sample of tweets.
- We considered only original posted tweets. We are not considering retweets. But, there are many platforms that post news or copies of tweets, e.g. IFTTT. Therefore, we could find the exact same tweet published by many different users.
- Some Tweets (10%) include the location of the user but this information is not always real which makes classification by origin difficult. Besides, the georeferenced information is not representative.

3 Methodology

Having reviewed the references and settled the scope of the study, we selected some tools and techniques available to analyze the information.

3.1 Scope

The population considered for this study was formed by 192,525 tweets downloaded directly from Twitter. The distribution of tweets is shown in the Figure 2.

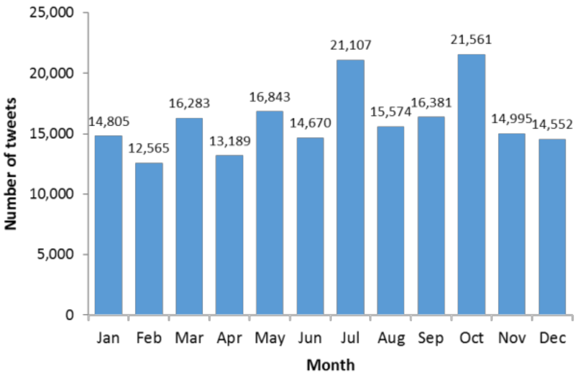


Figure 2: Tweets per month during 2016.

3.2 Text Mining

Text Mining is the methodological process of information extraction, where researchers deal with collections of documents, using specialized analysis methods (Balbi, et al., 2012).

The complete process we have followed is shown in Figure 3. We collected tweets from 12 touristic places from Perú, that were suggested by Promperú, using words related to those places. We used Python with GNIP PowerTrack® API to retrieve and download the information.

The sintax used to obtain tweets was: *(lang:en) (machupicchu OR picchu OR sacsayhuaman OR camino inca OR cusco OR cuzco OR plaza Armas cusco OR plaza armas cuzco OR ollataytambo OR salinas maras OR qorikancha OR coricancha OR ccoricancha OR pisac OR aguas calientes OR moray OR valle sagrado OR nasca lineas OR nazca lineas OR paracas OR islas ballestas OR manglares tumbes OR puerto pizarro OR parque reserva lima OR larcomar OR plaza armas lima OR parque kennedy OR catedral lima OR pachacamac OR monasterio santa catalina OR cañon colca OR colca OR uros OR taquile OR sillustani OR cathedral cusco OR cathedral cuzco OR main square lima OR nazca lines OR nasca lines OR ballestas island OR main square cusco OR square kennedy OR santa catalina monastery OR colca canyon OR sacred valley OR inca trail OR inKa trail OR cordillera blanca OR huaraz OR llanganuco OR amazon river peru OR mancora OR vichayito OR madre dios OR manu park OR iquitos OR tarapoto OR kuelap OR moche OR titicaca OR lake 69 OR laguna 69 OR señor sipan OR lord sipan OR pullana OR misti volcan OR misti volcano OR yanahuara OR huascaran OR churup OR chavin OR mochica OR pacaya samiria OR tucume OR cumbemayo OR baños inca OR wari OR punta sal OR amotape OR catacaos OR bahuaja sonene OR cotahuasi OR tingo maria)*

During initial pre-processing of the datasets, all tweets where labelled with the cities and touristic places, by matching words related to the places. For example, if the text contained the words "Sacsayhuaman" and "Chan-Chan", we put 1 in the columns "Cusco" and "La Libertad". It means, the tourist is talking about attractions in two different regions. And 0 in all other columns, corresponding to other Peruvian regions.

Not every tweet is clear and correctly written, we had to perform a cleaning process, considering the next steps:

a) Drop Retweets, Hashtags and mentions.
b) Drop strange symbols
c) Drop webpages
d) Drop punctuation signs
e) Drop numbers
f) Drop tabs or multiple spaces

We frequently found spelling mistakes in the tweets. For that reason, we used a spelling corrector. E.g. we replaced 'speling' by 'spelling' with an algorithm in Python (Dean and Bill 2007), for the Stemming we used SAS Text Miner and NLTK (Bird, et al., 2009). Finally, we removed the most common *stopwords* such as 'a', 'the', 'an', etc. from the datasets.

After the cleaning process, we tried to understand the dataset, analyzing the comments, using different R packages (qdap, koRpus, tm). SAS Text Miner was helpful for Topic Modeling and Term Relationship diagrams. Finally, R was used for Sentiment Classification (library syuzhet) and Correspondence Analysis (function corresp).

We have pre-defined some "concepts", using the Text Profile Node of SAS Text Mining (SAS Institute Inc. 2013). It performs a Hierarchical Bayesian Model which predict the concepts associated to each touristic place. We have grouped some terms in concepts, after lemmatization, associated to positive and negative characteristics of the attractions visited.

In our case, we also identified, all these words, considering **verbs**, to represent the activities that tourists do in the places and **adjectives**, to represent the opinions or value perceived by the tourist.

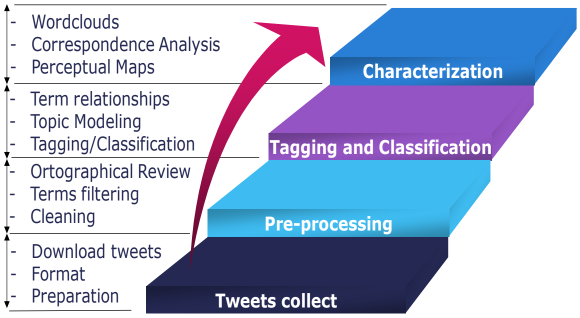


Figure 3: Text mining process performed.

3.3 Sentiment Analysis

The first sentiment analysis performed with text data, were published around 2002. As Agarwal stated, most of the studies were focused in text classification, positive or negative categories (Agarwal, et al., 2011).

One of the first teams performing sentiment analysis with Twitter information, was leaded by O'Connor. They also considered only the classification of 1 billion Twitter messages, but looking the changes over time (O'Connor, et al., 2010).

Mohammad developed a historic review of Sentiment Analysis, but considering Emotions (Mohammad, et al., 2010). They consider the Plutchik classification of emotions (Figure 4). Plutchik stated that emotions, like in color theory, can produce others, when they are combined. This allows to calculate scores for a given text, trying to classify it in any of those combinations.

In fact, some tools include a process which use these concepts to classify texts, for example the Library Syuzhet in R. This library calls some others, to have a great performance when scoring sentiments and calculating polarity. With this library, we can use the function get_sentiment with two arguments (text and method, the default is 'syuzhet') in order to calculate the score for the text in each one of the eight main sentiment categories listed in Figure 4. The greatest score indicates the more prevalence of the sentiment. The default method uses a sentiment lexicon, developed in the Nebraska Literary Lab under the direction of Matthew L. Jockers (Jockers, 2015).

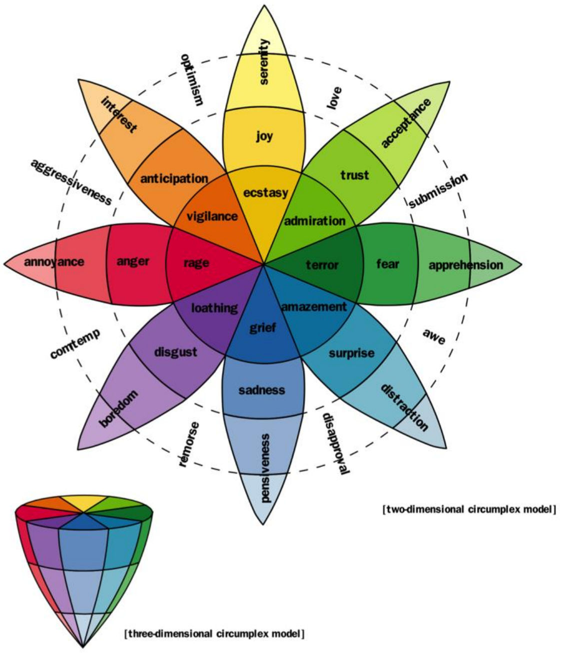


Figure 4: Plutchik wheel of emotions (Plutchik, 1980)

3.4 Correspondence Analysis

Maybe, the most representative researcher in Correspondence Analysis made by Benzécri. He has not invented this technique, but consolidated the methods involved (Benzécri, 1973). He was working since early 60's in Categorical Data Analysis and Mathematic Linguistics in French.

Correspondence Analysis was created as a qualitative statistical dimension reduction technique. Is very useful for everyone to understand the numeric associations between two categorical factors. After performing a matrix decomposition of the analyzed crosstab, you can represent the original frequencies in a bi-dimensional space.

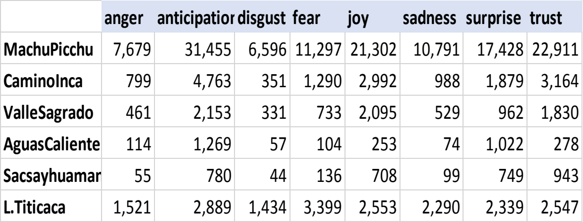
As stated in the book of Venables and Ripley, after applying the method Singular Value Decomposition (SVD) is easy to calculate the row and column Scores.

Suppose we have an r x c table N of counts. Correspondence analysis seeks 'scores' f and g for the rows and columns which are maximally correlated. Clearly the maximum correlation is one, attained by constant scores, so we seek the largest non-trivial solution. Let R and C be matrices of the group indicators of the rows and columns, so $R^TC = N$. Consider the singular value decomposition (SVD) of their correlation matrix:

$$X_{ij} = \frac{n_{ij}/n-(n_{i.}/n)(n_{.j}/n)}{\sqrt{(n_{i.}/n)(n_{.j}/n)}} = \frac{n_{ij}-nr_ic_j}{n\sqrt{r_ic_j}} \quad (1)$$

Where $r_i = n_{i.}/n$ and $c_j = n_{.j}/n$ are the proportions in each row and column. Let D_r and D_c be the diagonal matrices of r and c. Correspondence Analysis corresponds to selecting the first singular value and the left and right singular vectors of X_{ij} and rescaling by $D_r^{-1/2}$ and $D_c^{-1/2}$, respectively (Venables and Ripley, 2002).

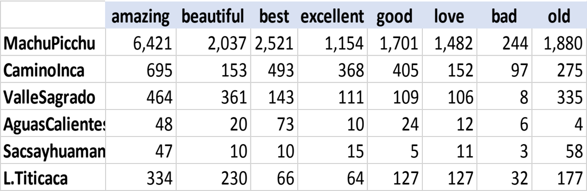
In our case, see Table 1 and 2, the first input matrix is:



	anger	anticipatior	disgust	fear	joy	sadness	surprise	trust
MachuPicchu	7,679	31,455	6,596	11,297	21,302	10,791	17,428	22,911
CaminoInca	799	4,763	351	1,290	2,992	988	1,879	3,164
ValleSagrado	461	2,153	331	733	2,095	529	962	1,830
AguasCaliente	114	1,269	57	104	253	74	1,022	278
Sacsayhuamar	55	780	44	136	708	99	749	943
L.Titicaca	1,521	2,889	1,434	3,399	2,553	2,290	2,339	2,547

Table 1: Distribution of tweets by touristic places and sentiments expressed.

The other matrix is:



	amazing	beautiful	best	excellent	good	love	bad	old
MachuPicchu	6,421	2,037	2,521	1,154	1,701	1,482	244	1,880
CaminoInca	695	153	493	368	405	152	97	275
ValleSagrado	464	361	143	111	109	106	8	335
AguasCalientes	48	20	73	10	24	12	6	4
Sacsayhuaman	47	10	10	15	5	11	3	58
L.Titicaca	334	230	66	64	127	127	32	177

Table 2: Distribution of tweets by touristic places and concepts.

3.5 Correspondence Analysis Plot

After applying SVD, in each matrix, we get the row and columns scores. We can use them to generate the Correspondence Plot. It could be done putting in a scatterplot the scores calculated for each matrix.

Some researchers have used this kind of charts in text analysis before, but having different approaches. For example, Balbi et al developed two studies (Balbi, et al.,2012; Balbi, et al.,2013), drawing plots to model language. In addition, they analyzed relationships between terms, to understand the meaning of the texts, shown Figure 5.

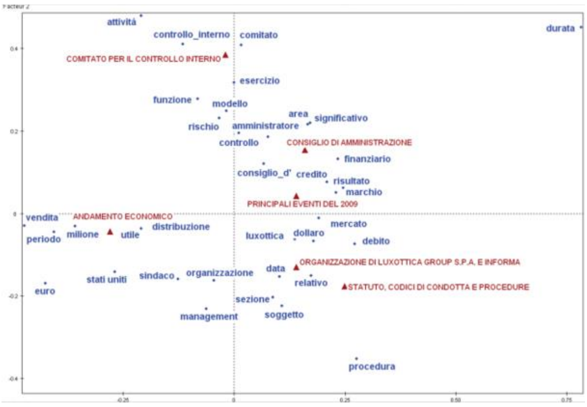


Figure 5: Correspondence analysis first factorial plane (Balbi, et al.,2012).

Interestingly they used first all the verbs and adjectives of the analyzed texts to find those words called "Concepts", which express key characteristics for them.

4 Results

Having downloaded more than 500 thousand tweets, pre-filtering touristic places and content, we needed to filter the data more carefully, because it contained a lot of tweets talking about other subjects. In consequence, we discarded:

a) Retweets

b) Tweets without text, only pictures

c) Tweets in other languages.

After this process, we got 192,525 tweets, but published by only 70,161 unique users. Even twitter is a popular social network, users have different patterns of activity. We considered "Active users" those who published at least 12 tweets in all the span analyzed, from January to December 2016. At least one tweet per month, in average. We discovered that only 2.5% of the total were active users shown in Figure 6.

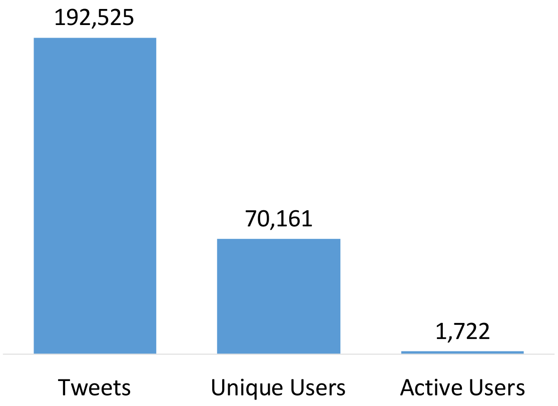
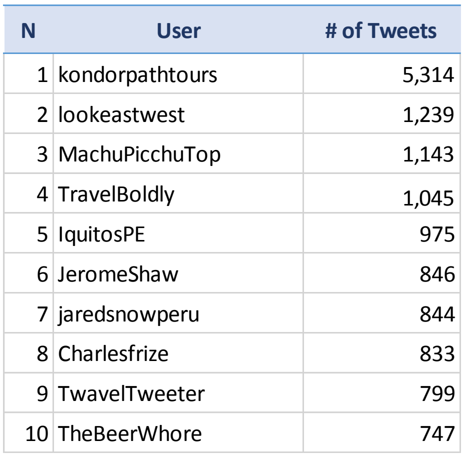


Figure 6: Number of tweets and users.

After that, we identified the most active users, in order to discover specialized users, bots (robots or automatic processes that publish tweets) and other kind of strange or interesting patterns. We found in the top some touristic companies, users who love to travel, accounts for travelers, etc. Shown in Table 3.

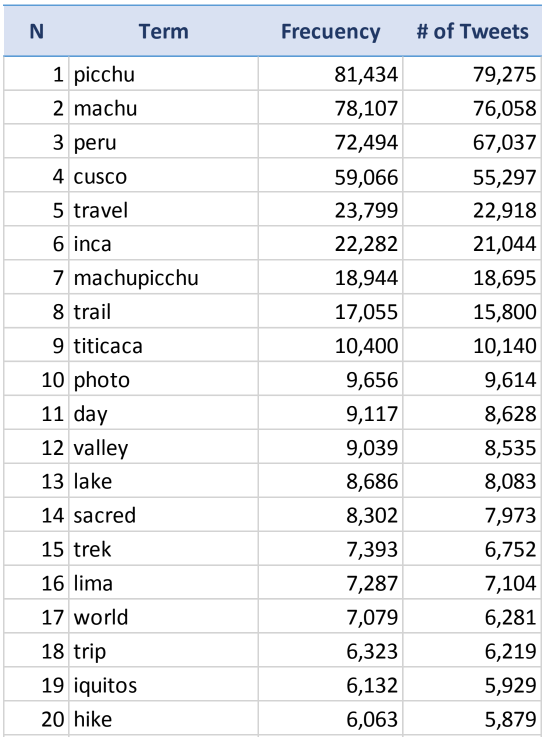


N	User	# of Tweets
1	kondorpathtours	5,314
2	lookeastwest	1,239
3	MachuPicchuTop	1,143
4	TravelBoldly	1,045
5	IquitosPE	975
6	JeromeShaw	846
7	jaredsnowperu	844
8	Charlesfrize	833
9	TwavelTweeter	799
10	TheBeerWhore	747

Table 3: Top ten of Twitter users.

Kondor Path Tours (@kondorpathtours) is a travel agency, very active in Twitter, publishing five times tweets more than the next one in the list. L. J. Blake (@lookeastwest) is a travel blogger, who travels around the world and post photos, places reviews and advices to other travelers.

After performing the twitter cleaning process, explained in section 3.2, we obtained the most frequent words in the corpus. In English tweets, Machu Picchu is most popular than Perú or any other Peruvian place or word. And it appears on the first places in the list again as “Machupicchu” in Table 4.



N	Term	Frecuency	# of Tweets
1	picchu	81,434	79,275
2	machu	78,107	76,058
3	peru	72,494	67,037
4	cusco	59,066	55,297
5	travel	23,799	22,918
6	inca	22,282	21,044
7	machupicchu	18,944	18,695
8	trail	17,055	15,800
9	titicaca	10,400	10,140
10	photo	9,656	9,614
11	day	9,117	8,628
12	valley	9,039	8,535
13	lake	8,686	8,083
14	sacred	8,302	7,973
15	trek	7,393	6,752
16	lima	7,287	7,104
17	world	7,079	6,281
18	trip	6,323	6,219
19	iquitos	6,132	5,929
20	hike	6,063	5,879

Table 4: Top twenty words in tweets.

If we use wordclouds, the visualization is not very helpful, because the most frequent words hide all the other ones in Figure 7.

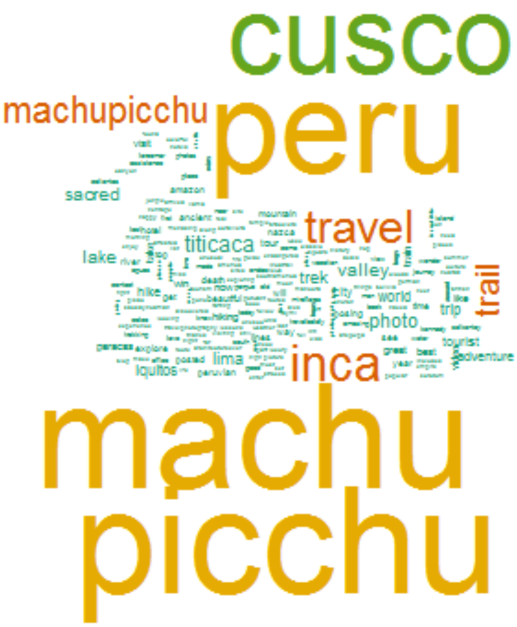


Figure 7: Wordcloud of 192,525 tweets.

We filtered some words to see the picture most clear, in our case we decided filter machu picchu, machupicchu, peru, cusco, travel, inca and trail in Figure 8.

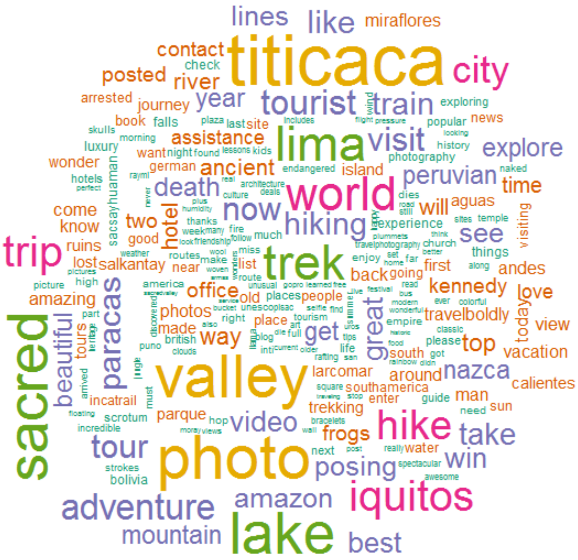


Figure 8: Wordcloud of tweets (words filtered).

Attitude Valuation: After cleaning processes, we used syuzhet package, to classify each tweet in Positive or Negative, depending on the polarity expressed. This method is not 100% accurate, but it gives a general idea. Other limitation is that tweets have only 140 characters, a limited way to express complex feelings (Jockers, 2015).

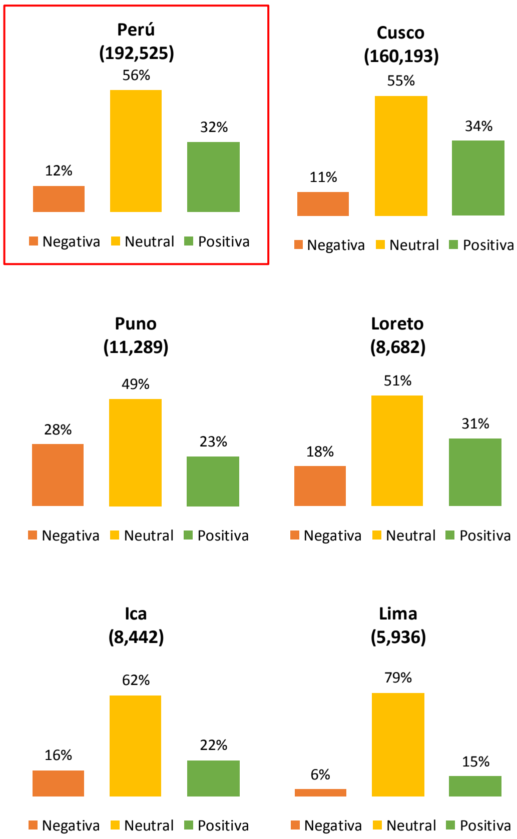


Figure 9: Positive, negative or neutral tweets.

We analyzed the opinions of all tweets related to touristic places in Perú, and after that, each region was considered in the analysis separately, with the count of positive, negative and neutral tags. According to the general results, Cusco is the region with the highest percentage of positive tweets and Lima is leading the group of neutral opinions shown in Figure 9.

We could see that, in general, there are many tweets classified as neutral. It is common in text classification, due to the nature of social media. The most frequent case is a user talking about news or things that are happening. For example: "I'm in Cusco", or "The truth about Paracas Skulls". In our context, trying to understand the attitude of the tourist, we could simplify the classification, from three levels to two. Only negative opinions or non-negative opinions, grouping positive and neutral ones.

Puno has the highest rate of negative tweets, because in early October, ten thousand of giant toads were found dead in the Titicaca Lake. According to Journalists, pollution in the Coata River would be blame for the deaths. It triggered a lot of claims and campaigns in social networks. Some protesting activists took around 100 dead frogs to the central square in the regional capital, Puno. A BBC article describing the disaster was shared a lot of times in Twitter (BBC News. 2016).

Touristic Places: Considering the most visited places mentioned by Promperú, we looked for negative and non-negative opinions in Table 5.

N	Place	# of Tweets	% Non Negatives
1	Machu Picchu	95,392	87.5%
2	Caminos del Inca	16,538	94.2%
3	Valle Sagrado	8,588	92.1%
4	Aguas Calientes	2,381	95.4%
5	Sacsayhuaman	1,672	86.0%
Cusco (all the region)		**160,193**	**89.1%**
1	Lago Titicaca	10,436	69.9%
Puno (all the region)		**11,289**	**71.8%**
1	Ciudad de Iquitos	5,950	79.8%
2	Río Amazonas	3,267	72.9%
Loreto (all the region)		**8,682**	**81.7%**
1	Reserva de Paracas	4,510	85.3%
2	Líneas de Nazca	3,785	82.0%
Ica (all the region)		**8,442**	**83.9%**
1	Parque Kennedy	2,927	94.5%
2	Larcomar	2,014	94.5%
Lima (all the region)		**5,936**	**94.5%**

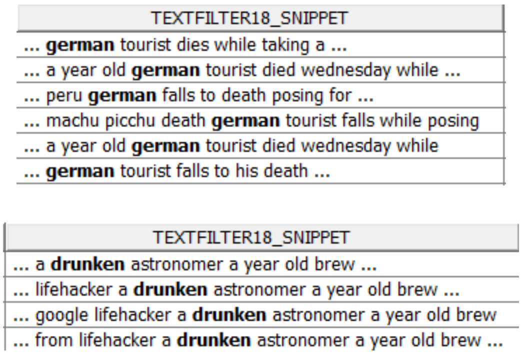
Table 5: Percentage of non-negative tweets.

The place with more negative tweets is Titicaca Lake. The places with better performance are Valle Sagrado, Aguas Calientes, and Caminos del Inca, in Cusco. Machu Picchu has a good performance too, but some incidents affected the reputation of them in Twitter.

Parque Kennedy and Larcomar, in Lima, have good image, too. The district of Miraflores is very touristic. They realized the great potential of tourism in Perú and many good hotels and restaurants are in the zone.

Concept Identification: Using Hierarchical Bayesian Methods, with the option 'Text Rules' in SAS Enterprise Miner, we identified the words which most appeared in negative tweets. It considers rules based on words co-occurrence to perform a simple classification (SAS Institute Inc. 2013).

We focused the analysis in each touristic place analyzed, in order to find specific words associated to the place. Some examples are shown in Figure 10.



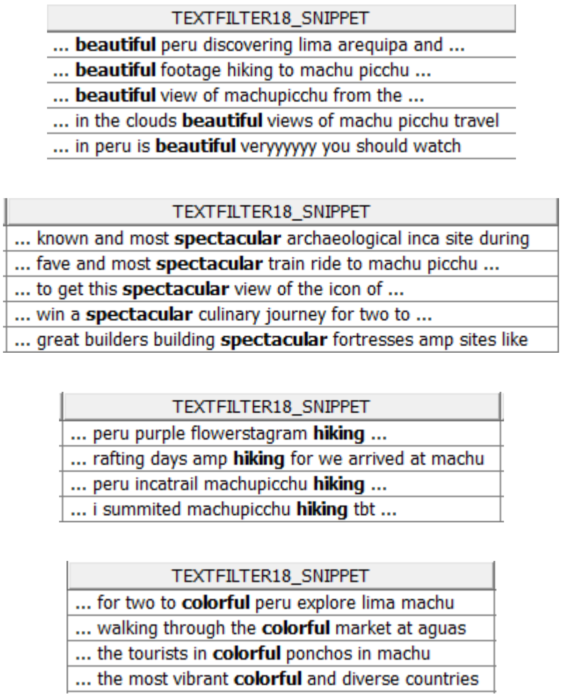
TEXTFILTER18_SNIPPET
... **german** tourist dies while taking a ...
... a year old **german** tourist died wednesday while ...
... peru **german** falls to death posing for ...
... machu picchu death **german** tourist falls while posing
... a year old **german** tourist died wednesday while
... **german** tourist falls to his death ...

TEXTFILTER18_SNIPPET
... a **drunken** astronomer a year old brew ...
... lifehacker a **drunken** astronomer a year old brew ...
... google lifehacker a **drunken** astronomer a year old brew
... from lifehacker a **drunken** astronomer a year old brew ...

Figure 10: Words found in negative tweets.

Those words are meaningful, because reflect some incidents occurred in Machu Picchu. A German tourist died in October, while trying take a picture of himself. He fell off a cliff in the mountains. In addition, we found tweets about "andean astronomers" or "shamans" who perform a ritual known as "Ayahuasca", which consists in drinking an Amazonian plant mixture, that is capable of inducing altered states of consciousness, usually lasting between 4 to 8 hours after ingestion. They could be related to bad or strange experiences, at least.

In the other direction, we also selected other words that appeared in non-negative tweets. Most them are related to describe good experiences with the landscapes, trekking routes and people in Figure 11.



TEXTFILTER18_SNIPPET
... **beautiful** peru discovering lima arequipa and ...
... **beautiful** footage hiking to machu picchu ...
... **beautiful** view of machupicchu from the ...
... in the clouds **beautiful** views of machu picchu travel
... in peru is **beautiful** veryyyyyy you should watch

TEXTFILTER18_SNIPPET
... known and most **spectacular** archaeological inca site during
... fave and most **spectacular** train ride to machu picchu ...
... to get this **spectacular** view of the icon of ...
... win a **spectacular** culinary journey for two to ...
... great builders building **spectacular** fortresses amp sites like

TEXTFILTER18_SNIPPET
... peru purple flowerstagram **hiking** ...
... rafting days amp **hiking** for we arrived at machu
... peru incatrail machupicchu **hiking** ...
... i summited machupicchu **hiking** tbt ...

TEXTFILTER18_SNIPPET
... for two to **colorful** peru explore lima machu
... walking through the **colorful** market at aguas
... the tourists in **colorful** ponchos in machu
... the most vibrant **colorful** and diverse countries

Figure 11: Words found in non-negative tweets.

In general, tourists enjoy 'beautiful' natural landscapes, 'spectacular' ancient buildings, ruins, food and services, 'colorful' handicrafts and traditional clothes. They liked to go hiking across the valley, too. All those words were collected to do the next task, which is grouping most relevant expressions in concepts, using co-ocurrence and synonyms.

Relationships between terms: To have a better understanding of the words and concepts discovered in the previous step, we performed a Link Analysis, between terms A and B, using a single metric, called Strength (SAS Institute Inc. 2013).

$$Strength = Ln(1/Prob_k) \quad (2)$$

Where:

$$Prob_k = \sum_{r=k}^{r=n} Prob(r) \quad (3)$$

$$Prob(r) \sim Binom(n, p) \quad (4)$$

n: Number of documents that contain term B

k: Number of documents containing A and B

p: k/n

Then, we can build some graphs, to represent the strongest relationships:

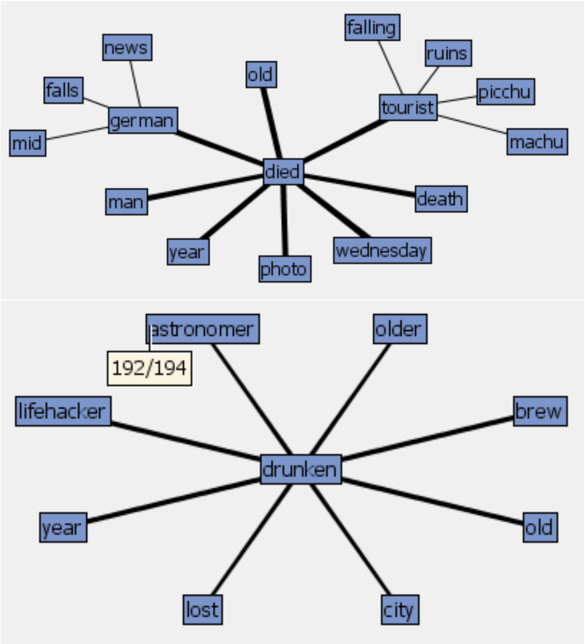


Figure 12: Relation between negative terms.

In the first graph in the Figure 12, we can see many words related to the German tourist incident. In the second one, all the words talking about the negative experiences, the thickness of the lines represents the degree of relationship. We can build the same graphs for Non-Negative terms to find things that tourists enjoy or like to do, shown in Figure 13.

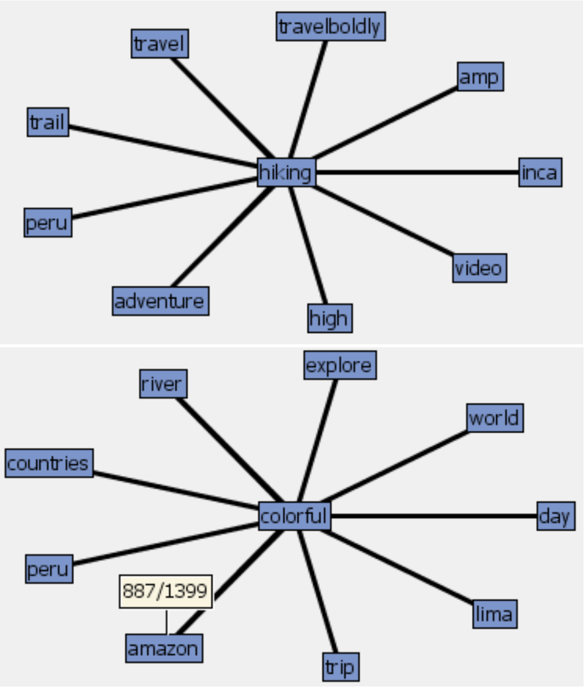
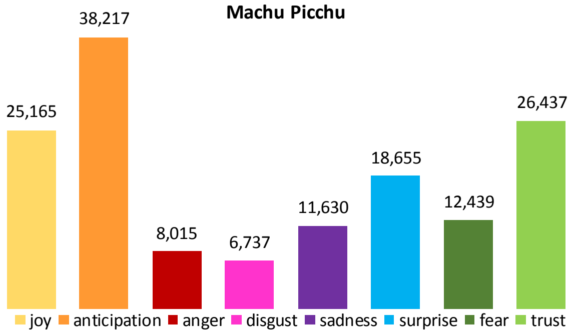


Figure 13: Relation between non-negative terms.

Emotions discovery: The next step is dealing with sentiments. In order to classify all the tweets, we use the syuzhet package (Jockers, 2015) again to identify the eight main emotions. After that, we have plot the emotions distribution for each place visited. Some of them are shown in Figure 14.



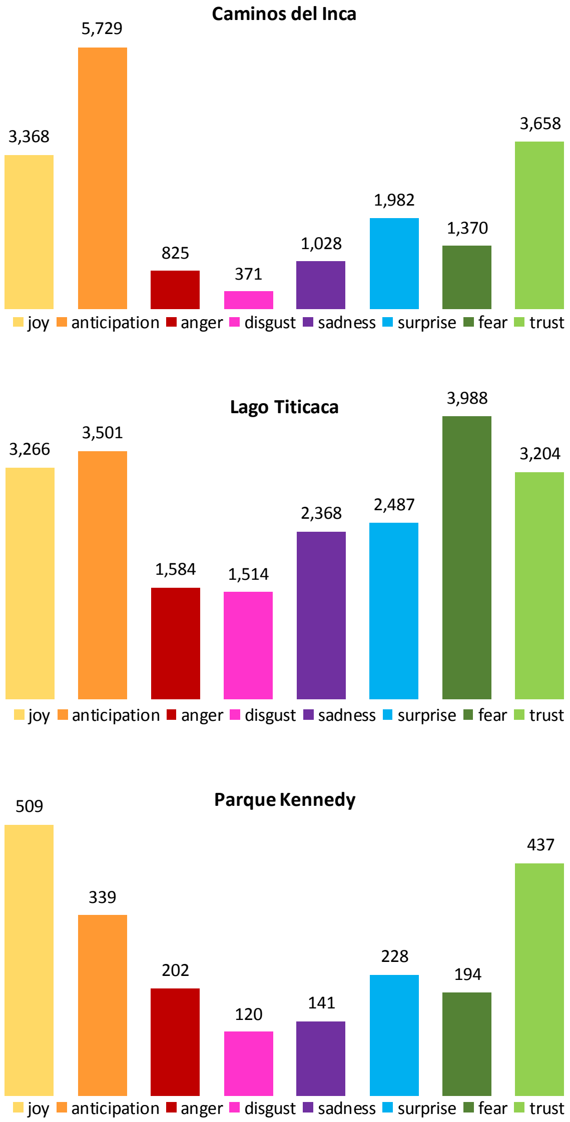


Figure 14: Tweets classified by emotions.

We can notice the differences between places most accepted, like Caminos del Inca or Parque Kennedy. The percentage of tweets that contains bad feelings like anger, disgust, sadness or fear is really low, against the other positive ones. In contrast, Lago Titicaca has fewer differences between the sentiments expressed in the tweets. The most frequent is fear.

Correspondence Analysis: We now have the key elements to build the crosstabs, perform the correspondence analysis and draw the plots. First, we can observe the distribution of sentiments per place visited:

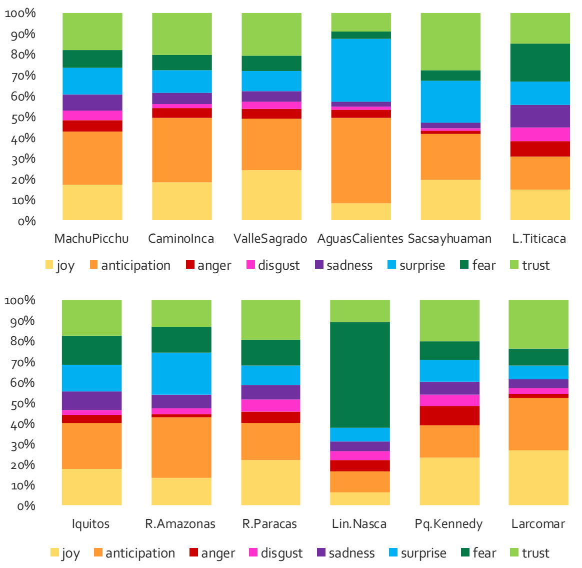


Figure 15: Tweets by emotions and place**.**

We have divided the visualization in two parts, shown in Figure 15, because is hard to represent twelve places in one shot. Having the complete picture, we can see the predominant emotions by place. For example, Aguas Calientes has a lot of tweets expressing surprise and anticipation. Lago Titicaca is associated with fear and sadness expressions. Lineas de Nasca has many words related to fear, like “misterious” or “unknown”. Other similar reactions are caused by accidents occurring to aircrafts with an alarming frequency (StudioKnow, 2016). In fact, posts like the mentioned, were sparsed on Twitter, accompanied by other alarming messages.

The next step is to reduce dimensionality, calculating the row and column scores, from the tables shown in Table 1 and 2. With the scores, we built the correspondence plots. Like the bar plots in Figure 14, we divided the representation in two blocks, to keep the results visible. We can see the sentiments associated to each place in the plots.

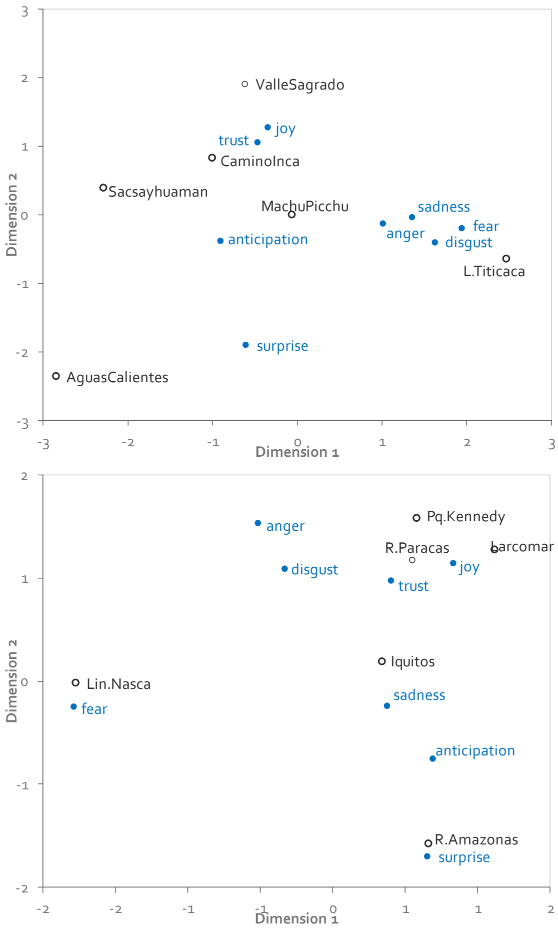


Figure 16: Correspondence plots by emotion and place.

The first plot in figure 16 is very clear about Lago Titicaca. All the negative emotions (‘disgust’, ‘fear’, ‘sad’ and ‘anger’) are close to this place, which is interpreted as a relationship. The words ‘trust’ and ‘joy’ are more related to places such Valle Sagrado and Camino Inca. We have explained the situation with pollution and death of animals in Puno, the relationship could be confirmed. In the case of Cusco, there local government and authorities are aware about the importance of providing the best experiences to the visitors, so they and the touristic business (operators, hotels, restaurants, guides, etc.) work together to bring good services and hospitality.

In the second plot in figure 16, we can see the proximity between Lineas de Nasca and the sentiment ‘fear’, Rio Amazonas with ‘surprise’ and Paracas, Parque Kennedy and Larcomar associated with the sentiments ‘joy’ and ‘trust’. In the case of Nasca, many tourists commented about the insecurity to take a flight in old and out-of-dated airplanes.

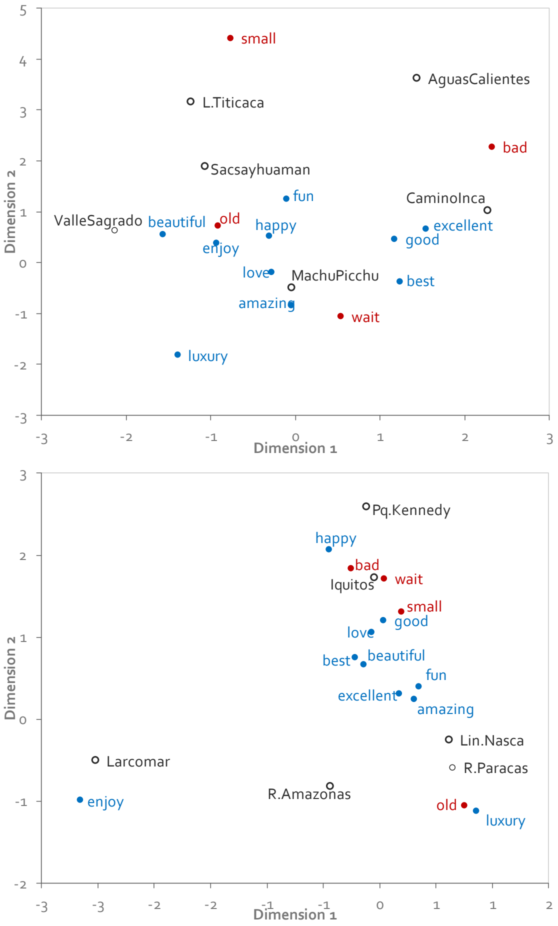


Figure 17: Correspondence plots by concept and place.

Considering the concepts, we have selected some Non-Negative ones: Amazing, Best, Beautiful, Enjoy, Excellent, Fun, Good, Happy, Love and Luxury. The other, Negative Concepts selected were: Bad, Old, Small, Wait. All of them were found as common and frequently appeared in the tweets, see Figure 17. They are used to describe or express valuations about their experiences in the visited places. They resume other common expressions, to avoid synonym issues in the interpretation of texts, and were grouped using linking formula (1).

Geolocation: In addition, we represented all the geo-tagged tweets in a map. Only 17 thousand tweets had Longitude and Latitude coordinates. Notice that coordinates correspond to the user location in the moment when the tweet was published. For that reason, 92% of the tweets were posted from any place in Perú, shown in Figure 18.

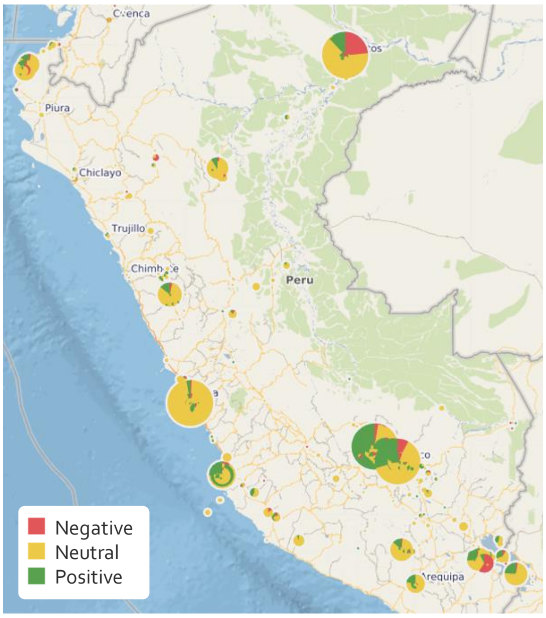


Figure 18: Map of geotagged tweets

5 Conclusions

Considering the insights found in this study, we can conclude:

- The number of users tweeting about touristic places is interesting, but only 2.5% of them are active users, posting 12 or more tweets per year.
- The most commented place in Perú, by far, is Machu Picchu. It is mentioned more than words like Cusco, or even Perú. It became a symbol, since it was declared a New Wonder around the world.
- The site most valued by English-speaking tweeters is Aguas Calientes, with more than 95% of comments classified as positive. Lago Titicaca is the worst rated, since it does not reach 70% of positive comments.
- We can notice a great influence of negative incidents in Twitter, due to the viral effect.
- This study was a good way to explore the opinions about Perú, but we have to keep in mind some limitations, like representability or social network popularity in our country.
- Twitter activity is the highest in July, during the northern hemisphere countries summer vacation time. And in October, when some incidents occurred in the most visited place, Cusco.

Acknowledgments

We would like to express our special thanks to Market research team in Promperú, who gave us the opportunity to develop this project.

References

Promperú. 2016. *PENTUR: Plan Estratégico Nacional de Turismo*. Ministerio de Turismo y Comercio Exterior. Lima, Perú.

Mohammad, S. and Turney, P. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an Emotion Lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text,* California, LA.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. 2011. Sentiment Analysis of Twitter Data. In *Proceedings of the workshop on languages in social media,* pages 30-38. Association for Computational Linguistics. Portland, OR.

Dodds, P., Harris, K., Kloumann, I., Bliss, C. and Danforth, C. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE 6(12): e26752. doi:10.1371/journal.pone.0026752

Antoniadis, K., Vrana, V. and Zafiropoulos, K. 2014. Promoting European Countries' Destination Image Though Twitter. In *European Journal of Tourism, Hospitality and Recreation,* 5(1):85-103. Leiria, Portugal.

Oku, K., Hattori, F. and Kawagoe, K. 2015. Tweet-mapping method for tourist spots based on now-tweets and spot-photos. In *19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems* 60(2015):1318-1327. Shiga, Japan.

Bassolas, A., Lenormand, M., Tugores, A. Gonçalves, B. and Ramasco, J. 2016. Touristic site attractiveness seen through Twitter. In *EPJ Data Science* 2016:5-12. doi:10.1140/epjds/s13688-016-0073-5

O'Connor, B., Balasubramanyan, R., Routledge, B. and Smith, N. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media,* Washington, DC.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research and experience,* 1(3):3-33.

Jockers, M. 2015. Syuzhet: Extract Sentiment and Plot Arcs from Text. CRAN Network. https://github.com/mjockers/syuzhet

Benzécri J.-P. 1973. *L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'Analyse des Correspondances,* Dunod, Paris, France.

Venables, W. and Ripley, B. 2002. *Modern Applied Statistics with S. Fourth edition.* Springer, London, UK.

SAS Institute Inc. 2013. SAS® Text Miner 13.1 Reference Help. Cary, NC: SAS Institute Inc.

Balbi, S., Stawinoga, A. and Triunfo N. 2012. Text Mining tools for extracting knowledge from Firms Annual Reports. In book: *JADT 2012: 11es Journées internationales d'Analyse statistique des Données Textuelles*. Edit. Dister A., Longrée D., Burnelle G., pages 67-79.

Balbi, S. and Stawinoga, A. 2013. Mining the ambiguity: correspondence and network analysis for discovering word sense. In book: *SIS 2016 Scientific Meeting,* pages 1-7.

BBC News. 2016. Peru investigates death of 10,000 Titicaca water frogs. Published in Oct, 18.

StudioKnow. 2016. Is it Safe to Fly Over the Nazca Lines? Published in May,11.

Bird, Steven, Edward Loper and Ewan Klein 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Dean and Bill 2007. How to Write a Spelling Corrector http://norvig.com/spell-correct.html