Emoji Usage Analysis between English Speakers and Spanish Speakers on World Cup 2022

Yu-Hsin Huang (yh3666@columbia.edu), Jittisa Kraprayoon (jjk2239@columbia.edu)

Supervisor: Dr. John R. Kender

Columbia University

Abstract

Emojis are used to convey common emotions and symbols, and are considered to transcend language barriers. However, it is important to consider that differing cultures influence differing usage of emojis. In this study, we aim to identify cultural differences in emoji usage between English and Spanish speakers. We first construct a dataset of English and Spanish Instagram comments related to the 2022 World Cup. In our preliminary analysis, we use emoji counts and compute TF-IDF scores. To perform a more semantic analysis, we convert the Instagram comments to sentence embeddings using LASER (Language-Agnostic SEntence Representations) and emoji2vec. We employ both hierarchical clustering and k-means clustering on the resultant sentence embeddings and use t-SNE to visualize the clusters. As is expected, emoji usage for English and Spanish is extremely similar. Emojis such as 😂, 🔥, and 👏 are commonly used in both languages. These emojis serve common purposes and emotions for both languages, supporting the global nature of emojis. However, cultural differences are also found. For example, the more frequent usage of emojis like 🤣, 💀, and 😝 convey a mocking or playful tone distinct to the English dataset. Additionally, there is a tendency to repeat emojis in diverse heart colors (Ex. 💙🤍💙) among Spanish speakers which is indicative of cultural ties to the national flag. Another finding is that English speakers tend to use the clapping hand emoji (👏) and the fire emoji (🔥) in combination. These findings offer more insight into the cultural differences in emoji usage, specifically between the English and Spanish language on social media.

1. Introduction

In recent years, emojis have emerged as powerful tools for expressing emotions, sentiments, and cultural nuances in a visual format. The 2022 FIFA World Cup, a global spectacle that unites millions of fans worldwide, provides a unique context to investigate the intricate interplay between emoji usage and cultural differences.

1.1 Instagram for data extraction

We decided to extract our data from Instagram due to its large and global user base of over 2.4 billion. Additionally, Instagram comments are known to contain emojis: By 2015, "nearly half of instagram" content contained emojis (Dimson, 2015). Instead of extracting all content, we decided to focus on comments as that was where users expressed opinions and sentiments in regards to a post.

There are many advantages to Instagram comments. For example, because each post has multiple comments, this ensures that we have more text and emoji data to work with. Moreover, the existence of an official FIFA account enables us to focus on the 2022 World Cup event while drawing from a global but mostly consistent fanbase of users. Another advantage is that we ensure that even when drawing data from comments across different posts, the content and format will be relatively the same as it is posted from the same FIFA account. This is important as each post may invoke a different comment reaction from users.

1.2 The 2022 World Cup

The FIFA World Cup, which takes place every 4 years, is considered to be the largest football tournament in the world. The 2022 FIFA World Cup took place in Qatar from November 20 to December 18 and featured 32 teams.

The 2022 World Cup comment section presents an optimal arena for exploring cultural differences due to its global reach and the diverse fanbase it attracts. As football stands as a unifying force, the tournament transcends geographical and cultural boundaries, bringing together individuals from various corners of the world. The comments section becomes a digital forum where fans express their emotions, opinions, and affiliations. For example, we can explore how different cultures employ emojis to celebrate victories, lament defeats, and convey a

spectrum of emotions related to the universal language of football. Cultural differences can manifest not only in the selection of emojis but also in their interpretation, as can be inferred with the accompanying text in each comment.

1.3 English versus Spanish

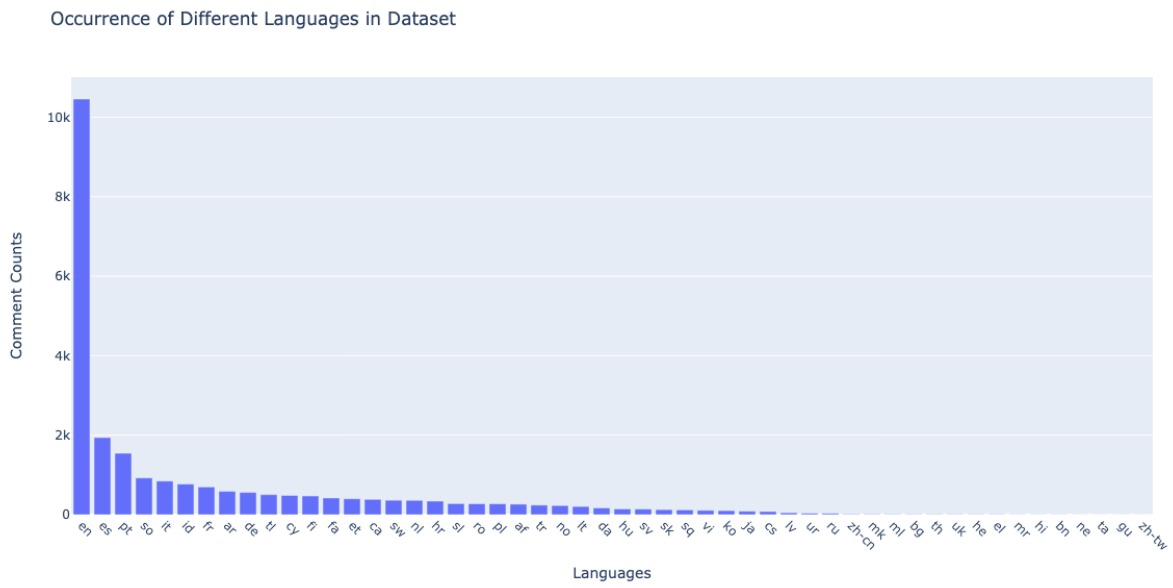

Figure 1. Comment counts for all languages in raw data

We chose to explore cultural differences between English and Spanish due to the large number of comments existing in these languages. Figure 1 shows comment counts in multiple languages after our data scraping. 'es' and 'en' denote English and Spanish languages respectively. The number of English comments exceeds counts in all other languages at 10465 comments. However, from the data that we acquired, Spanish places second at 1922 comments (Note that further data preprocessing will significantly reduce the number of comments we actually end up working with).

It also makes sense to explore English and Spanish comments in the context of the World Cup. English is a global language so it is advantageous to use it due to the global nature of the World Cup. Spanish is a language spoken in many South American countries, a continent which is

known to have a rich footballing culture. Countries that have historically dominated the World Cup also include Brazil, Argentina, and Uruguay.

1.4 Related Work

A motivating research paper from our lab studied emoji usage in response to news videos with the aim of finding cultural differences (Zhang). We employed the same method of data collection by scraping comments from posts. Additionally, we were inspired to use emoji2vec embeddings as a representation for the emojis. Previous research has also shown there is specific emoji usage distinct to certain cultures. Individuals from Finland and Pakistan create and use specific emojis that align with their respective cultural norms (Sadiq & Shahida, 2019). In another paper, K-means clustering is applied on emoji probability distributions to relate national development indicators to emoji usage (Ljubešić & Fišer, 2016). Using emoji counts to construct probability distributions for each country, researchers found that there exist distinct differences in emoji usage between "first", "second", "third", and "fourth world" countries. Most relevantly, for "first world" countries (in this definition, includes countries in North America and Australia among others), individuals show a lack of emotion through their emoji usage. Conversely, for "second world" countries (in this definition, includes most of South America as well as other countries), individuals use emojis with more emotional clarity. This research demonstrates that emoji counts may provide deeper insights into differences between countries. Additionally, in our research we similarly employ K-means clustering. In an additional research paper, Word2Vec is trained on tweets with emojis and the emojis are then clustered into different emotions (Mayank & Pal, 2016). In our work, we were also inspired to perform clustering to find whether different emojis were used to convey similar emotions in English and Spanish. While research has been done to study cultural differences in emoji usage and clustering is a common method to derive further insights from emojis, our research introduces a focused study between English and Spanish within a specific context: the 2022 World Cup. Additionally, we also experiment with a new embedding that captures both text and emoji meanings.

2. Methods

2.1 Data Acquisition

In this research study, our data acquisition involved the extraction of comments from the FIFA official Instagram account. This approach was primarily facilitated through the utilization of Apify, a web scraping tool. Due to the inaccessibility of certain comments from private accounts and Instagram's policy, where only 50 random comments per post could be viewed anonymously, we were able to extract approximately 40 comments per post. The data collection period extended from November 20th to December 18th, aligning with the duration of the World Cup event. This yielded our raw dataset, consisting of 993 posts and 40,785 comments.

Following the data acquisition phase, data cleaning processes were performed to ensure the quality and relevance of the dataset. Initially, language categorization was implemented to segregate the collected raw data into two primary languages—English and Spanish (see figure 2). Any data in languages other than these, or lacking any textual content, were excluded from the dataset. Furthermore, given the research focus on emoji analysis, comments devoid of emojis were also excluded. Subsequently, we extracted comments not generated during the FIFA World Cup 2022 Event, refining the dataset to emphasize the event-specific context.

Lastly, to address potential biases and ensure a balanced dataset, we normalized our data by randomly selected 10 comments per day per language. As a result, a total of 560 Instagram comments were included in the final dataset, with 280 comments each in English and Spanish.

	text	language
0	@xidiwldnxjqpspqpskxls morroco is a siblingcountry aswell. 20% of them live in Belgium😂.	en
1	Son🤣	en
2	We are ready 💛💙❤️	en
3	Can't wait for this shitshow to be over and for next world cup. This one is such a waste 🗑️ 😁. #nextplease #notoqatar2022 #boicotworldcup	en
4	@bf_steve Who do you think I cheer for 😂	en
5	TERANGA boys 🇸🇳❤️	en
6	You are the best lady😻😂🙏👏♥⚽🤩fifa.	en
7	@fatenalshoubaki respecting cultures goes both ways, i won't invite a muslim man to my house and force him to eat pork because "it's in my culture to eat pork", both hosts and guests should be able to respect eachother 👍	en
8	@itisjuanderful AAJAA POR NO PERMITIR GAYS EN LA COPA MUNDIAL YO LO VEO EXCELENTE👏🔥	en
9	@tiago.rendas2 bro I do know football but yesterday match that wasn't an offside lol 🤣🤣🤣what cause the camera showed his leg over the line but have you seen the last defender in the box 🤣🤣🤣go and learn football then come back chat to me	en
10	Orgullosamente Ecuatorianos 😍	es
11	Vamos con todo 🇪🇨♥⚽	es
12	@carlos_arte24 no se porque but I love the jersey 🥶	es
13	Felicidades Ecuador! 🙌	es
14	@alvarodnl Si no vemos el mundial, Qatar pierde plata q ganó dejando a personas morir 😐	es
15	👏👏vamos con todo, strong	es
16	Ecuador! 🇪🇨🇪🇨🇪🇨🇪🇨❤❤	es
17	Corrupción 💰💰💰💰	es
18	@ikercasillasrules No veo donde sale neuer en el video 😮	es
19	❤️❤️❤️❤️❤️❤️ Ecuatorianos 😍😍😍😍😍	es

Figure 2. Dataframe Example

2.2 Initial Exploration

As our data contained both text and emojis, we first wanted to know the ratio between the words and emojis used in comments. Our hypothesis was that one language may be more likely to use a greater number of emojis per word. Words were separated using spaces and emojis were counted by checking whether the token was also contained in emoji.EMOJI_DATA (a set of known emojis from Python's emoji library). Additionally, we counted the total number of occurrences of each emoji across both language datasets. This frequency count is used as a basis for further exploration of the data.

2.3 TFIDF

To assess the significance and distinctiveness of emojis in our dataset, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) metric. Adapting the conventional definition by replacing "word" with "emoji," the TF-IDF calculation proceeds as follows:

1. Term Frequency (TF):

$$\mathrm{TF}(emoji, comment) = \frac{\text{Number of times the emoji appears in the comment}}{\text{Total number of emojis in the comment}}$$

Term Frequency (TF) is established by computing the ratio of the count of a specific emoji in a comment to the total number of emojis in that particular comment. This calculation measured the emoji's frequency within the context of an individual comment. Consequently, each emoji is assigned a TF score within each comment, reflecting its relative significance at the comment-level.

2. Inverse Document Frequency (IDF):

$$\mathrm{IDF}(emoji, corpus) = \log\left(\frac{\text{Total number of comments in the corpus}}{\text{Number of comments containing the emoji}}\right)$$

Inverse Document Frequency (IDF) is computed by applying the logarithm to the ratio of the total number of comments in the corpus to the number of comments containing the specific emoji. This calculation provides an IDF score for each emoji, reflecting the uniqueness of the emoji across the entire corpus.

3. TF-IDF Score:

$$\text{TF-IDF}(emoji, comment, corpus) = \text{TF}(emoji, comment) \times \text{IDF}(emoji, corpus)$$

The TF-IDF score for an emoji in a specific comment is derived by multiplying its TF and IDF scores. TF represents the frequency of the emoji within individual comments, while IDF assesses its uniqueness across the entire corpus. The resulting TF-IDF score reflects how important an emoji is within a comment and its distinctiveness across the dataset. Through the computation of TF-IDF scores for each emoji in every comment, we aimed to explore more emoji patterns and emoji usage differences within our data.

2.4 Emoji Repetition

As mentioned in "The Emoji Code: How Smiley Faces, Love Hearts, and Thumbs Up Are Changing the Way We Communicate," which notes that "Emojis, for instance, are often repeated, adding emphasis by visual repetition" (Evans & Vyvyan, 2017). To gain more insights into emoji usage and explore how cultural differences impact such usage, we delved into "Repeated emoji," referring to instances where the same emoji is used consecutively multiple times within a comment. For instance, the sequence 😀😀😀 exemplifies repeated emoji usage, while patterns involving a mix of emojis such as 😀🥲😀 do not fall within this classification.

Our analysis involved examining two key aspects of repeated emoji usage. Firstly, we investigated the Maximum Repetition Count of each emoji, representing the occurrence of an emoji within a single comment. This metric reflects the highest frequency with which a specific emoji is consecutively repeated within a comment.

Additionally, we explored the Emoji Repetition Frequency, indicating the prevalence of repeated emoji usage within each dataset. This provided insights into how frequently emojis are iteratively

employed, contributing to an understanding of the prevalence of repeated emoji usage in our data.

2.5 Embeddings

To perform a more semantic-based analysis of our data, we decided to convert our comment data to word embeddings. This will enable us to perform clustering to potentially find more meaningful patterns in the data.

First, LASER (Language-Agnostic SEntence Representations) was chosen to convert our comments to sentence embeddings. This embedding model was chosen due to its cross-lingual transferability; it is able to represent sentences from different languages in the same space (Artetxe & Schwenk, 2019). LASER was applied to the text-only component of comments from both languages. Each comment would then be represented as a 1024-length vector.

To handle emojis, we used Emoji2vec to convert the emoji-only components of each comment to emoji embeddings. Emoji2vec produces a 300-length vector for each emoji occurrence. For comments with multiple emojis, we chose to take the average of the emoji embeddings to represent the emoji component of the comment. This was done so that the length of the embeddings were consistent across all comments. Additionally, we wanted to maintain an appropriate ratio between the length of the text embeddings and the length of the emoji embeddings.

The resulting text and emoji embeddings were concatenated for each comment. As the emojis from both English and Spanish comments come from the same set, and the LASER embeddings are language-transferable, our concatenated embeddings would represent the entire comment (both text and emoji) and enable comparisons between English and Spanish.

2.6 Clustering

2.6.1 Hierarchical Clustering

In addition to employing K-Means Clustering, we also leveraged hierarchical clustering to discern inherent relationships within our dataset. Hierarchical clustering, an unsupervised learning method, is utilized to group similar items based on their characteristics, specifically embeddings in our case, thereby forming a hierarchical structure of clusters. Unlike other clustering methods, hierarchical clustering organizes data in a tree-like structure, known as a dendrogram, which offers a visual representation of the relationships among different entities in the dataset (see figures 3 and 4).

Hierarchical clustering iteratively merges or splits clusters based on the similarity or dissimilarity of data points. To determine the optimal linkage method and threshold for clustering, we drew insights from a comparative study on hierarchical clustering methods. The resource suggests that Ward's method tends to perform well with equal sample sizes but may not be as effective with unequal sample sizes from bivariate data. The comparison also highlights that complete linkage is most similar to Ward's method in specific scenarios, and the choice of the best method depends on the specific characteristics of the dataset, including cluster sizes (Ferreira & Hitchcock, 2009). Additionally, a study by Blashfield (1976) comparing four hierarchical clustering methods (single linkage, complete linkage, average linkage, and Ward's method) found that the accuracy in the recovery of original population clusters varied, indicating that the effectiveness of each method may depend on specific dataset characteristics (Blashfield, 1976).

In our experimentation, we assessed various linkage methods, such as single, weighted, average, complete, centroid, median, and Ward. Since the Ward method proved most effective for our dataset, producing meaningful clusters without excessive fragmentation or collapsing all data into a single cluster, we refined our hierarchical clustering function by setting the distance threshold to 3 and adopting the Ward method.

The Ward method is a widely used linkage criterion in hierarchical clustering. It minimizes the variance within clusters when determining which clusters to merge. This is achieved by evaluating the sum of squared differences within clusters, aiming to create cohesive and compact groups. According to the SciPy v1.11.4 Manual, the Ward linkage criterion is expressed as follows:

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

where:

- u is the newly joined cluster consisting of clusters s and t,
- v is an unused cluster in the forest,
- $T = |v| + |s| + |t|$, and
- $|\cdot|$ denotes the cardinality of its argument.

This formula represents the distance between the newly formed cluster u and the unused cluster v, with $(d(v,s))$, $(d(v,t))$, and $(d(s,t))$ being the distances between clusters (v) and (s), (v) and (t), and (s) and (t) respectively.

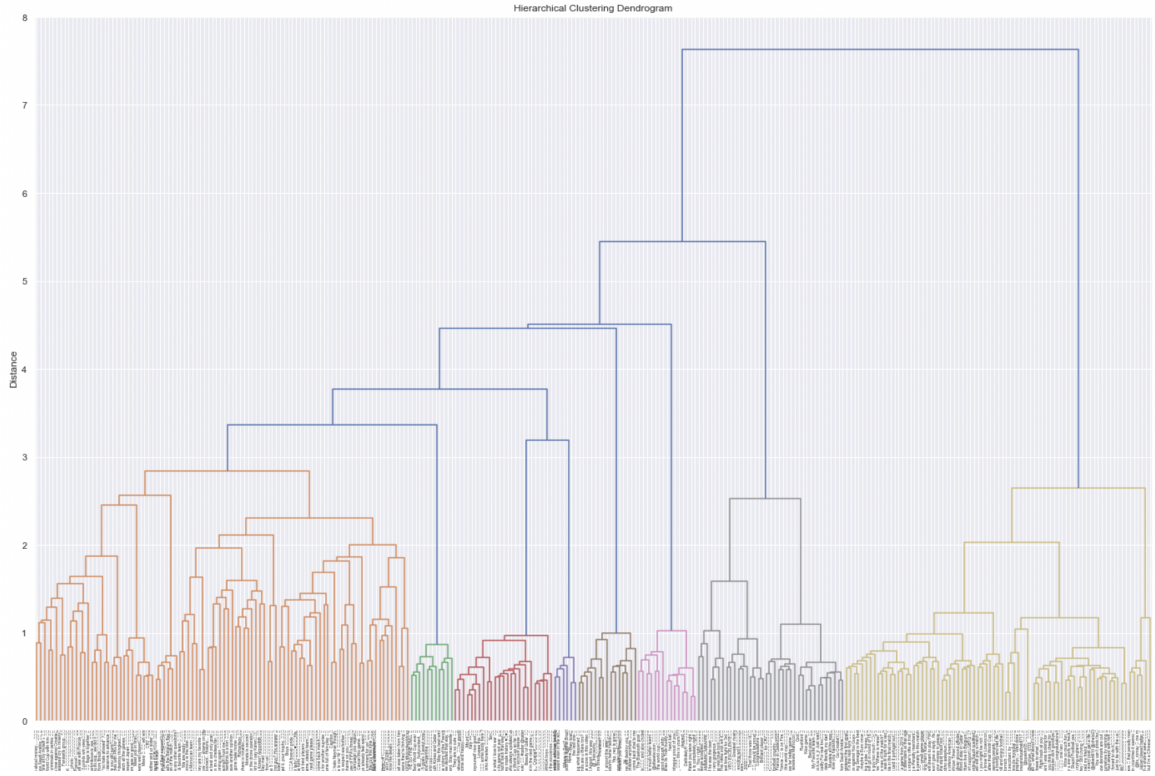

Figure 3. Dendrogram(English)

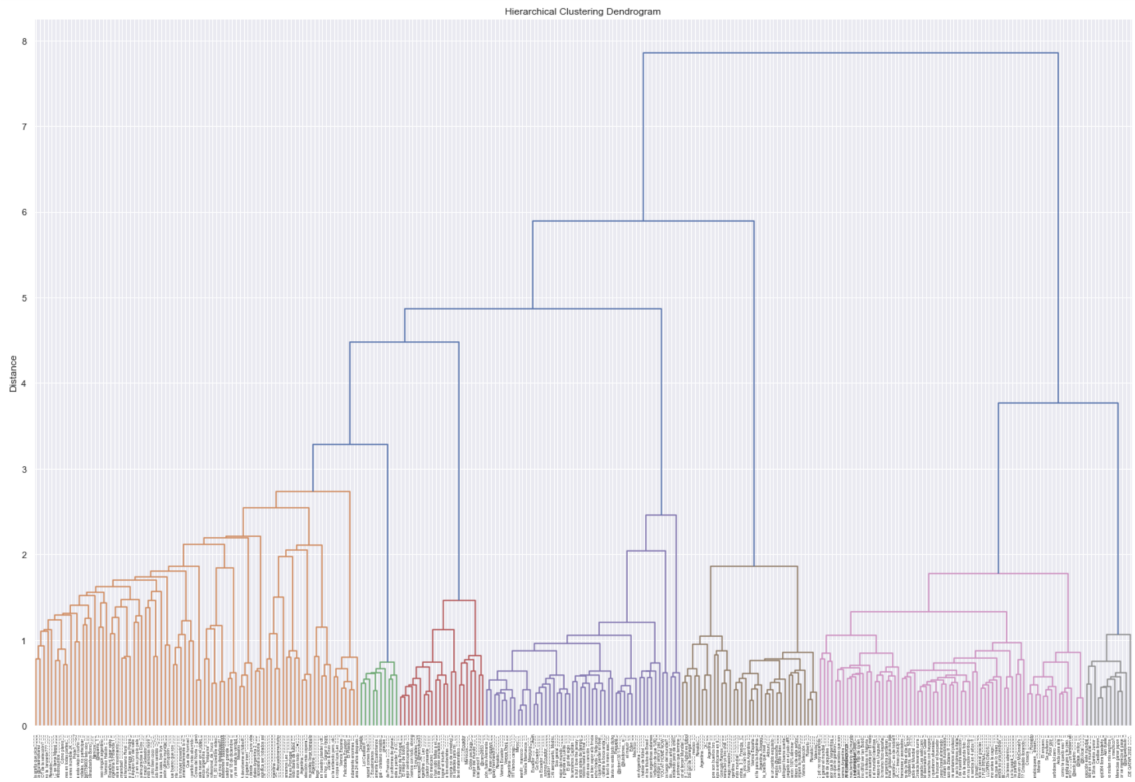

Figure 4. Dendrogram (Spanish)

2.6.2 K-Means Clustering

In addition to employing hierarchical clustering, we also applied the K-means clustering algorithm. The K-means clustering algorithm is a distance-based algorithm that repeatedly recomputes K centroids until a stopping condition is reached. First, K centroids are randomly assigned. Each data point is assigned a cluster based on its nearest centroid. New centroids are then computed as the mean of the data points of each cluster and the process is repeated. K-means aims to minimize the within-cluster sum of squared distances (Hartigan & Wong, 1979)

The K-means algorithm was chosen to cluster our sentence embeddings due to its ability to handle high-dimensional data. To choose the number of optimal clusters, K, we use the "Elbow Method" where K is plotted against the distortion score. The distortion score denotes the within-cluster sum of squared errors and clustering is done iteratively with values ranging from K=2 to K=30. The optimal value of clusters is then approximated by the "elbow" or inflection point in the curve. The resultant value of K for the English dataset was 11 (see figure 5). For consistency, the Spanish dataset was also clustered into 11 clusters for K-means.

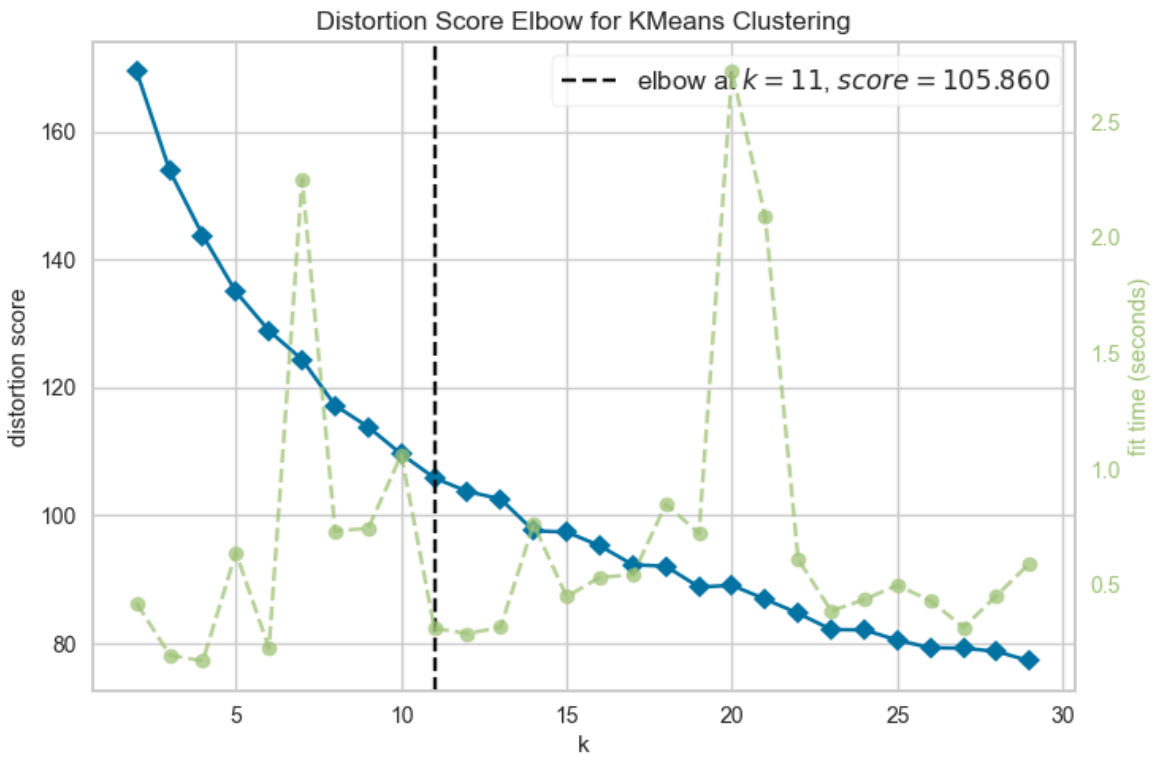

Figure 5. Elbow method using distortion score for the English dataset

2.6.3 Adjusted Rand Score

Lastly, we utilized the Adjusted Rand Score (ARS) to assess the similarity between clusters generated from different embedding strategies: text embeddings, emoji embeddings, and combined text and emoji embeddings. The Adjusted Rand Score is a modification of the Rand Index, specifically designed for evaluating clustering algorithms. It takes into account the expected similarity between random clusterings and adjusts the Rand Index accordingly. The adjusted Rand index ensures a value close to 0.0 for random labeling, and reaches exactly 1.0 when the clusterings are identical. The lower bound of the adjusted Rand index is -0.5 for particularly discordant clusterings. This metric serves as a tool for gauging the effectiveness of the clustering methods in capturing patterns and relationships within the data.

3. Results

3.1 Initial Exploration

The ratio of emojis to words in English was 0.244, while the ratio of emojis to words for Spanish was 0.285. In both languages, words were used more than emojis as expected as emojis are often used to emphasize what is being said in text. Spanish contained slightly more emojis per word than English, however the ratios are similar.

After data processing, we also analyzed the frequency distribution of the emojis (see figure 6 and 7). The 'count' is a summation of emoji occurrences over all comments. The resulting graphs show that the most frequently used emojis in both English and Spanish are quite similar. Both languages share a common 1st and 2nd place with the tears of joy emoji (😂) and fire emoji (🔥). The 3rd and 4th place emojis for both languages are also the clapping (👏) and black heart emoji (❤). However, their places are switched. The heart eyes emoji (😍) is the 6th most used emoji for both English and Spanish comments. The tears of joy emoji (😂) is used to express humor or lightheartedness. This result shows that both English and Spanish speakers approach the World Cup event with humor. On the other hand, the applauding emoji (👏), fire emoji (🔥), black heart emoji (❤), and heart eyes emoji (😍) can be used to acknowledge great performance from a player or team. This shows that both English and Spanish speakers use emojis to emphasize positive sentiments. This frequency analysis demonstrates that both English and Spanish speakers use similar emojis to express their sentiments during the 2022 World Cup. A notable difference is that English speakers use the tilted laughing emoji (🤣) significantly more than Spanish speakers (5th versus 17th most commonly used emoji).

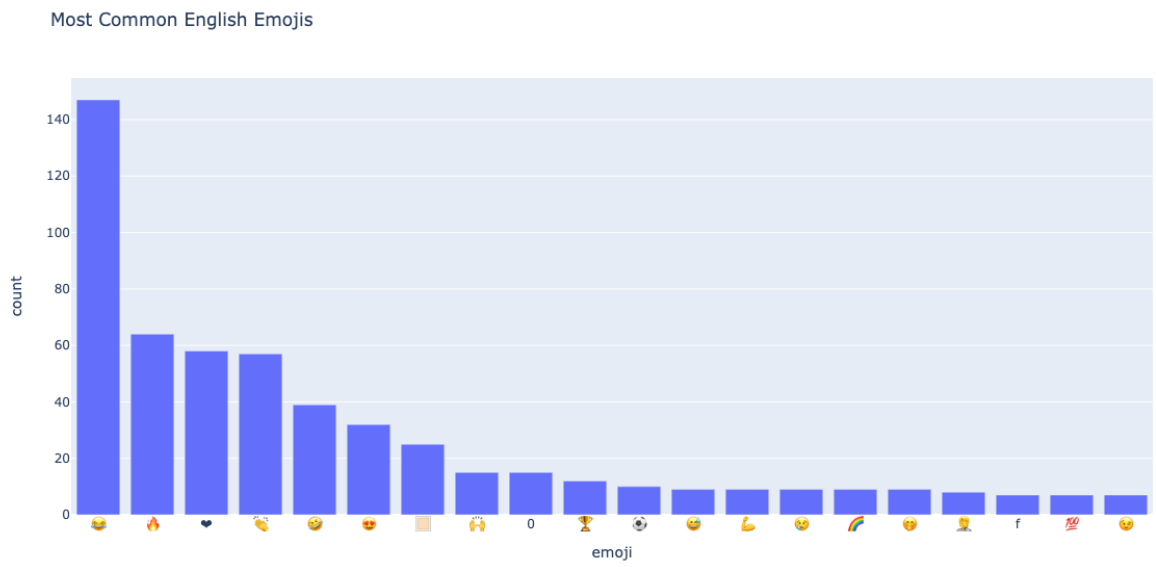

Figure 6. Frequency counts of emojis (English)

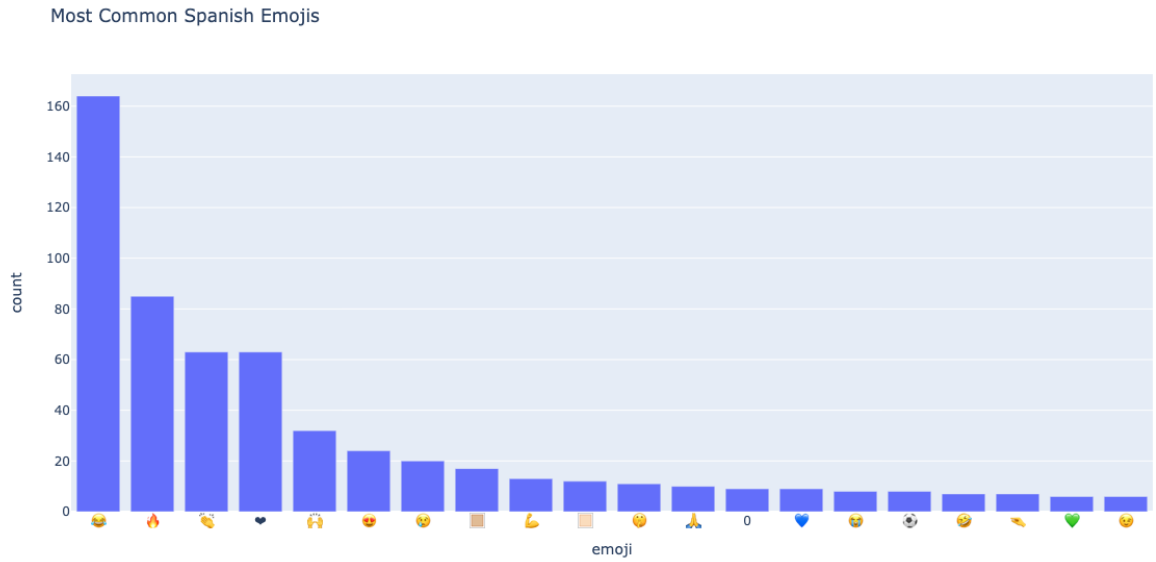

Figure 7. Frequency counts of emojis (Spanish)

Additionally, we found the set of common emojis occurring in both languages and the set of emojis only occurring in either English or Spanish. The set of common emojis had size 62. The set of emojis found exclusively in English comments had size 60 while for Spanish was 41. This implies that English speakers may use a more diverse set of unique emojis in Instagram comments. Figure 8 shows the log difference between counts of emojis in English subtracted by

counts of emojis in Spanish (where logarithmic count difference > |1|). English comments are more likely to contain the rainbow emoji (🌈) while Spanish comments are more likely to contain the pinching emoji (🤏). The rainbow emoji is used to represent LGBTQ rights. One controversy during the 2022 World Cup relates to LGBTQ rights in Qatar, which is the host nation (Turak, 2022). The results show that English users comment about the LGBTQ controversies at the 2022 World Cup more than Spanish users do.

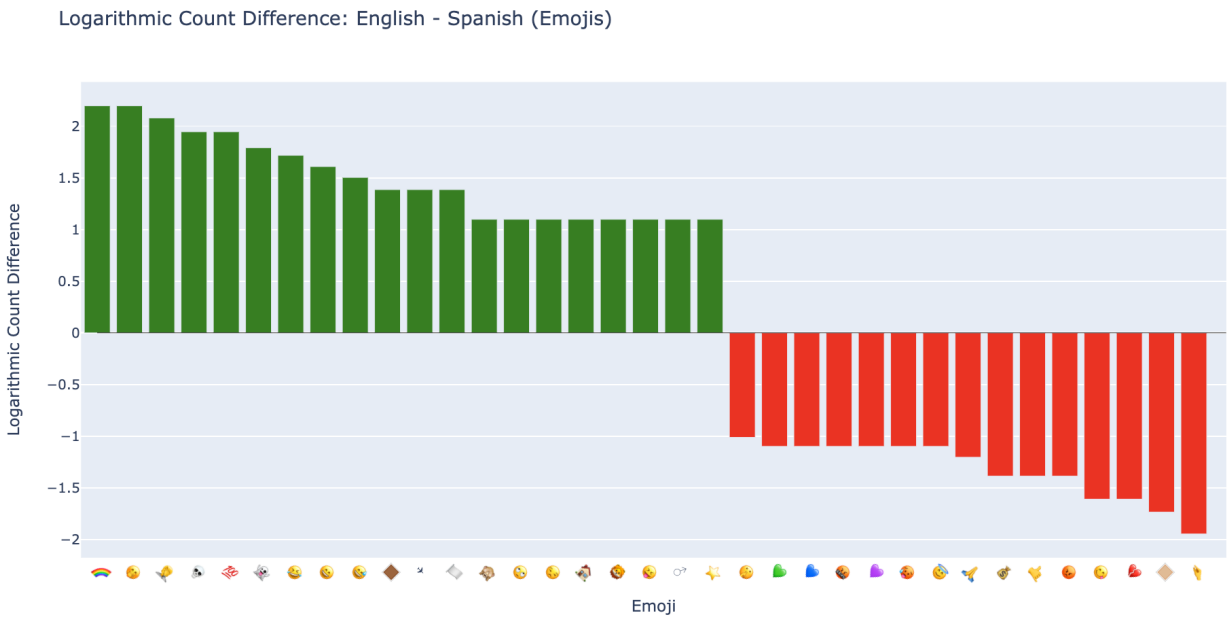

Figure 8. Logarithmic emoji count difference between English and Spanish

3.2 TFIDF

Upon computation of TF-IDF scores for each emoji within every comment, we found that a majority of TF-IDF scores equate to zero. This is primarily associated with the prevalence of comments containing only a single type of emoji. Consequently, TF scores for many emojis across a subsequent portion of comments resulted in zero values. In light of this pattern, we identified and recorded some emoji-comment combinations with the highest TF-IDF scores from both the English and Spanish language dataframes. This approach ensures that our analysis focuses on the most distinctive instances of emoji usage within the dataset.

	🔥	❤	👁️	👎	》	👆	😇	🤘	🤙	😄	...	👨	🕯	🥇	✍	💚	😓	💝	💛	🧙	😢
0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	0.0	0.314003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.656679	0.0	0.0
3	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
275	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
276	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
277	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
278	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
279	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

280 rows × 155 columns

Figure 9. TF-IDF (English)

	🔥	❤	👁️	👎	》	👆	😇	🤘	🤙	😄	...	👨	🕯	🥇	✍	💚	😓	💝	💛	🧙	😢
0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
1	0.0	0.856093	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.0	0.285364	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
275	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
276	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
277	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
278	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
279	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	2.146128	0.0	0.0	0.0

280 rows × 155 columns

Figure 10. TF-IDF (Spanish)

Instagram Comment Sample

Rematch 😡😡😡😡🇲🇦🇫🇷
Why I don’t see women anywhere on TV in Qatar? 😠😠😠😠
BGM is from Malayalam movie 🥰🥰
@sofia_chaverri welcome to #visitturkey 🌹🌹

Table 1. Top 4 Emoji-Comment Combinations in English

Instagram Comment Sample
El mejor , el máximo representante de🇵🇹 y este entrenador aferrado a dejrlo en la banca sr santos! El@quiere jugar no solo ver desde la banca, … No se vale que le hagan eso 😡😡😡 y en su propio país 😡
@enzosbr que ódio 🥺🥺🥺
Que no era que iba a ganar ?? Cristaldo llorando😹😹
Ya está viejo el cucho 😕😕
El mayor 🤩🤩

Table 2. Top 5 Emoji-Comment Combinations in Spanish

We observed that all emoji-comment combinations with high TF-IDF scores involve the repetition of uncommon emojis, irrespective of language, aligning with the definition of TF-IDF. According to TF-IDF definitions, emoji repetition contributes to high TF scores, while the uniqueness of emojis leads to high IDF scores.

A noteworthy finding in our study was the recurrence of angry faces (😡) in both English and Spanish results, demonstrating relatively high TF-IDF scores. This observation resonates with research suggesting that emojis, as stimuli, elicit strong emotional responses, with angry emojis

being rated highest in emotionality (Fischer & Herbert, 2021). Despite the infrequent use of angry faces, users consistently employ them multiple times within a single comment, highlighting their distinctive role in expressing intense emotions.

3.3 Emoji Repetition

In our analysis of the Maximum Repetition Count of each emoji, we observed that a majority of emojis are repeated between 2 to 5 times within a single comment. This range indicates a common pattern in the frequency of emoji repetition across both English and Spanish speakers, to emphasize what they want to express or highlight their emotions towards a discourse.

To examine potential distinctions in emoji usage between English and Spanish speakers, we computed the Maximum Repetition Count of each emoji in both languages and calculated the logarithmic difference, denoted by English minus Spanish (see figure 11). A notable finding suggested that Spanish speakers exhibit a tendency to repeat emojis in different colors of heart more frequently, whereas English speakers utilize only the red heart emoji when commenting on social media. This result may be attributed to cultural preferences, in which Spanish speakers often symbolize their support for sports teams through a color-coded representation of countries. This color-coding is associated with the national flag colors. For example, the repetition of green hearts represents Brazil, and the repetition of blue hearts may signify Argentina, where light blue holds prominence in their national flag.

This cultural finding suggests that emoji usage among Spanish speakers goes beyond a simple act of repetition. The intentional repetition of specific emoji colors enables individuals to not only express their emotions but also visually convey their cultural affiliations and emotional connections linked to specific countries.

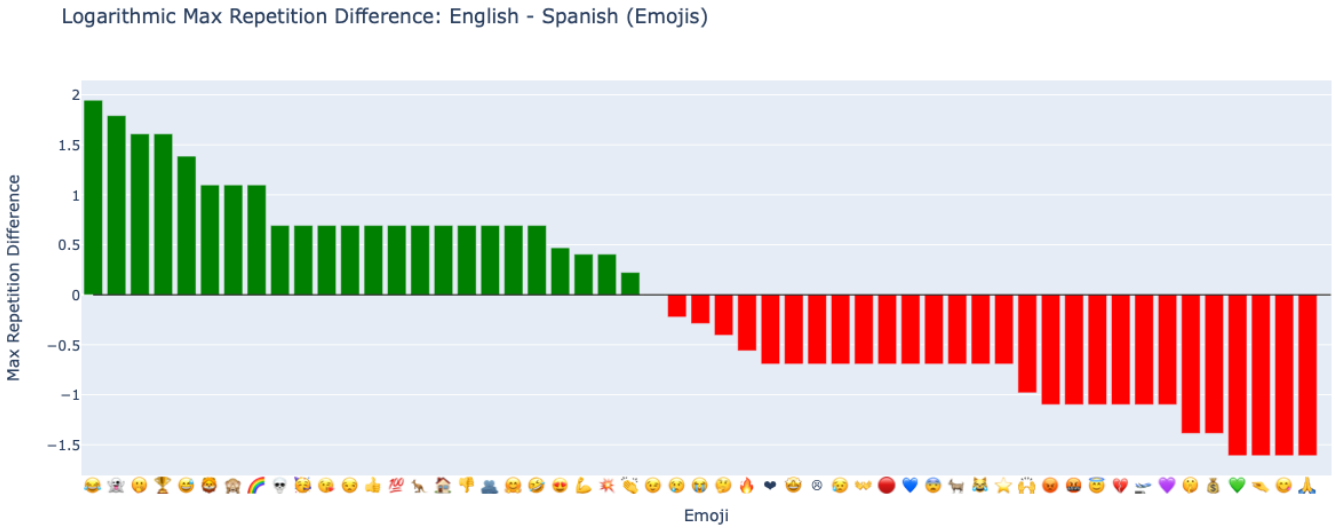

Figure 11. Logarithmic Maximum Repetition Difference of Emojis (English-Spanish)

Lastly, Emoji Repetition Frequency reveals that the emojis 😂, 👏, 🔥, and 😍 are consistently repeated frequently in comments across both English and Spanish languages. This finding mirrors our earlier observation in the initial exploration analysis, where these emojis emerged as widely used among both groups of English and Spanish speakers.

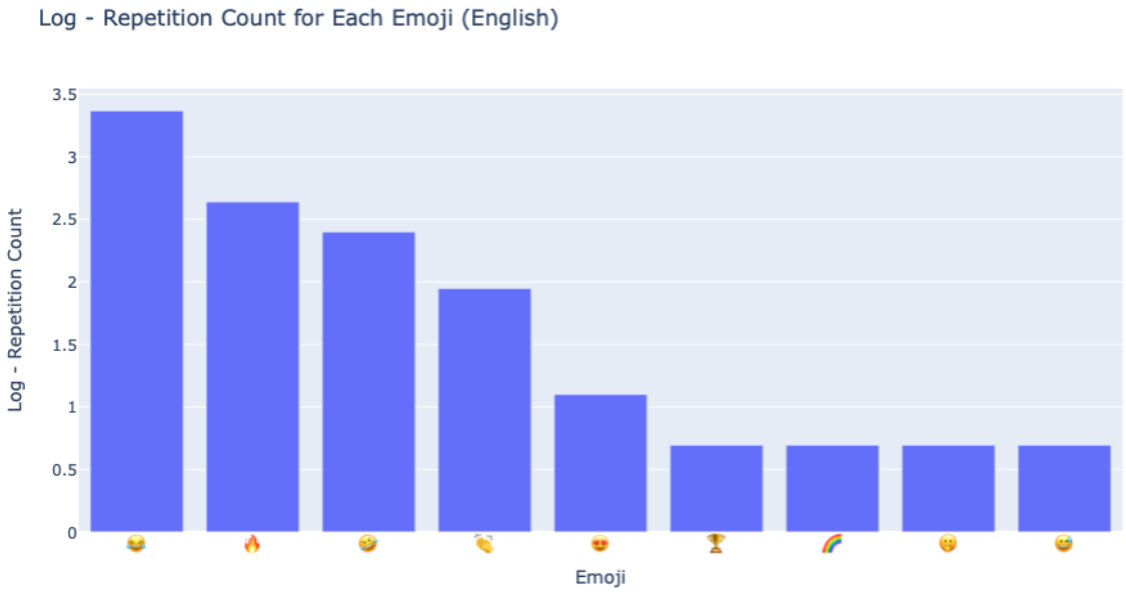

Figure 12. Logarithmic Repetition Count For Each Emoji (English)

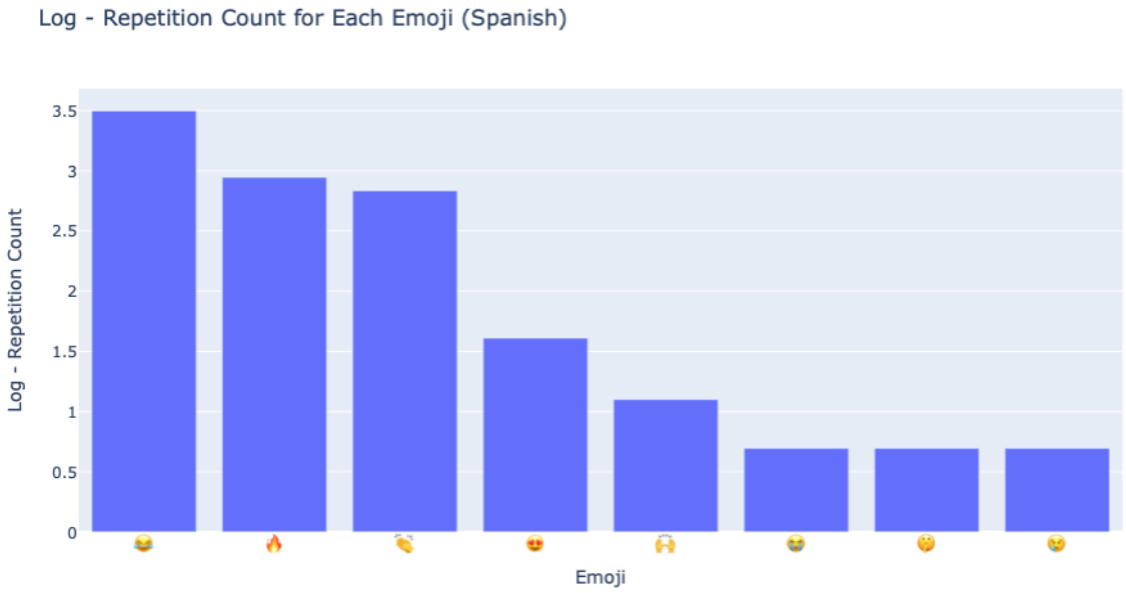

Figure 13. Logarithmic Repetition Count For Each Emoji (Spanish)

3.4 Clustering

3.4.1 Hierarchical Clustering

After applying hierarchical clustering to text, emoji, and combined text and emoji embeddings, the Adjusted Rand Index Scores were computed to quantify the similarity between the three types of embeddings. In the English dataset, the Adjusted Rand Index Score between text clustering and emoji clustering was -0.0185, while the score between text clustering and both clustering methods was -0.0265. Both scores being close to 0 suggest a lack of meaningful structure in the clusters, indicating the randomness of the clusterings. However, the Adjusted Rand Index Score between emoji clustering and both emoji and text clustering was notably higher at 0.8479, indicating a strong correspondence between emoji-centric clusters and those produced by both text and emoji embeddings. Similarly, in the Spanish dataset, the Adjusted Rand Index Score between text clustering and emoji clustering was -0.0112, between text clustering and both clustering was -0.0145, and between emoji clustering and both clustering was

significantly high at 0.8659. These findings suggested that the clusters are mainly emoji-based, with a significant agreement between emoji clusters and the combined text and emoji clusters.

Additionally, we utilized t-SNE scatterplots to visualize the distribution of clusters in both the English and Spanish datasets. Each color in the plots represents a distinct cluster, with the English dataset exhibiting 8 clusters and the Spanish dataset having 7 clusters. To better interpret our results, we labeled each cluster with its most dominant emojis, and found a consistent alignment between the labels of hierarchical clusters and the observations gained from the "most common emojis" analysis in the initial exploratory method. We also observed that both datasets included an "other" cluster, encompassing comments with less common emoji patterns.

One remarkable difference observed from the English clusters and Spanish clusters was the absence of clusters labeled with 💀 and 🤣 in the Spanish dataset. This discrepancy indicates a more frequent usage of these two emojis in English comments. Further investigation into the textual content of comments featuring these emojis, we observed that they share a common mocking or playful tone, highlighting the unique way of expressing sentiments with emojis in the English language.

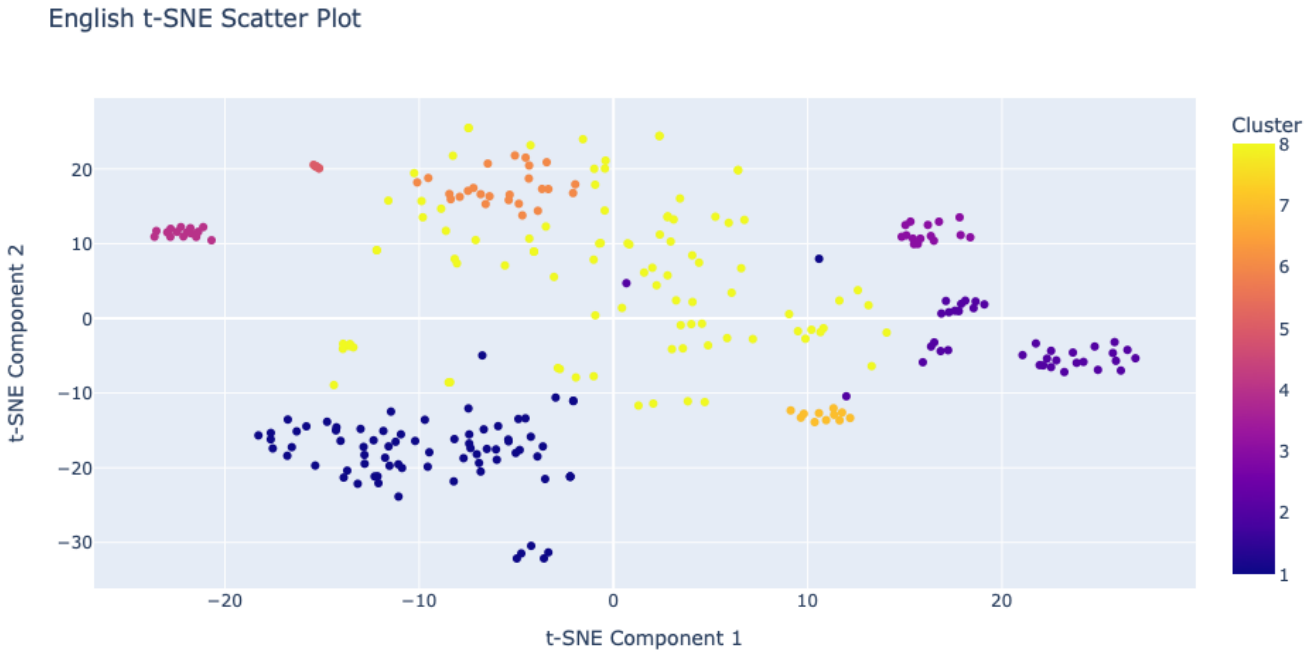

Figure 14. Hierarchical Clustering t-SNE Scatter Plot (English)

Figure 15. Hierarchical Clustering t-SNE Scatter Plot (Spanish)

In light of the Adjusted Rand Index Score results indicating the strong connection of emojis on hierarchical clusters, we generated additional t-SNE scatterplots for both the English and Spanish datasets. In this visualization, we assigned labels to each data point based on the first emoji that appeared in the corresponding comment. Our observations revealed that similar emojis tended to cluster together, providing evidence for the emoji-based clusters as indicated by the Adjusted Rand Index Score. Furthermore, the prevalence of comments with labels 💀 and 🤣 in English compared to Spanish became is also shown in this visualization, reinforcing our earlier conclusion from scatterplots colored by clusters.

The emoji-labeled scatterplots also revealed a distinct pattern in English, where the clapping hand emoji (👏) and the fire emoji (🔥) were frequently combined together. In contrast, the Spanish scatterplots showed a common pairing of the clapping hand emoji (👏) with the raising hands emoji (🙌).

Drawing insights from "Emoji as Digital Gestures," which explores individual and aggregate examples of emoji usage by English speakers, the fire emoji (🔥) is often employed as a positive force marker, associated with idioms like 'on fire' or 'lit' (Gawne & McCulloch, 2019). The term 'lit' is commonly used in English slang, where it signifies something cool or exciting. Consequently, the frequent combination of the fire emoji with the clapping hand emoji among English speakers suggests an expression of excitement or enthusiasm, influenced by the slang "It's lit". However, since Spanish lacks a comparable slang connecting fire with sentiments of excitement, the fire emoji is not commonly used in combination with the clapping emoji to cheer for a team among Spanish speakers.

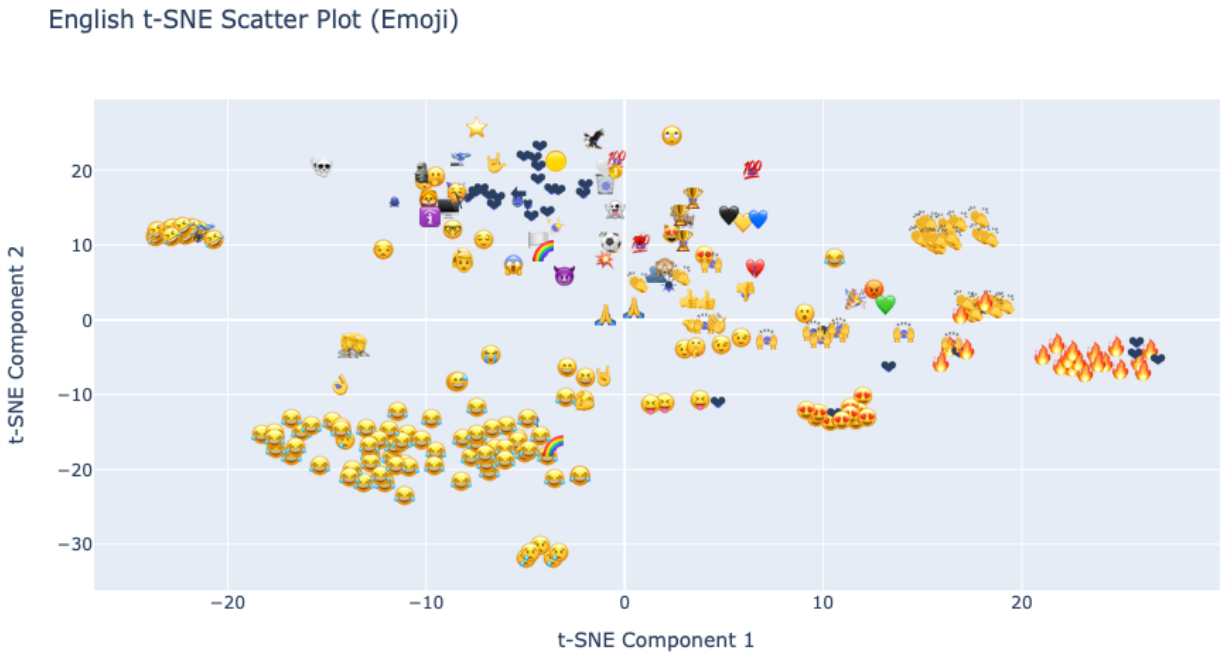

Figure 16. Hierarchical Clustering t-SNE Emoji Scatter Plot (English)

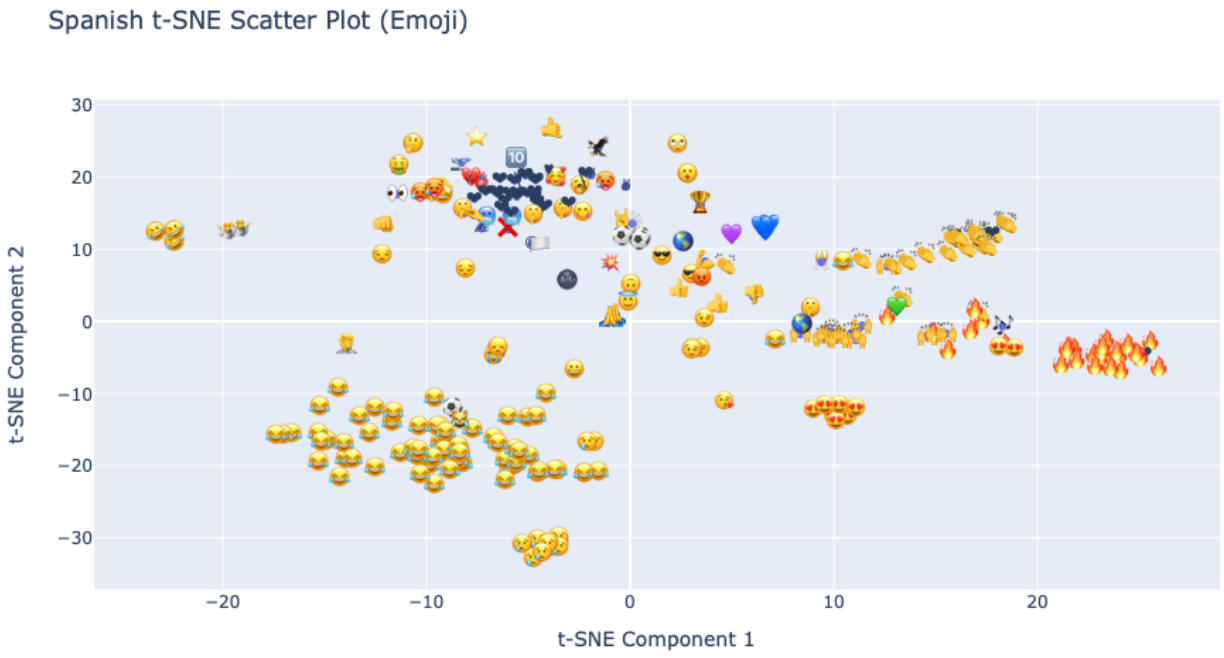

Figure 17. Hierarchical Clustering t-SNE Emoji Scatter Plot (Spanish)

3.4.2 K-Means

Like hierarchical clustering, K-means clustering was applied to text, emoji, and text and emoji embeddings. The Adjusted Rand Index Scores were then calculated. For the English data, the Adjusted Rand Index Score between text clustering and emoji clustering was 0.0008. The score between text clustering and both clustering was 0.0169. However, the score between emoji clustering and both clustering is much higher at 0.7079. Applying the Adjusted Rand Index Score with the Spanish dataset shows a similar trend. The Adjusted Rand Index Score for text clustering and emoji clustering was 0.0208. The score for text clustering and both clustering was 0.0216. Lastly, the Adjusted Rand Index Score for emoji clustering and both clustering was 0.8522. Likewise to the hierarchical clustering results for different embeddings, the resulting sets of clusters from K-means show that the text embeddings have small influence on the concatenated embedding results. These clusters are not similar to each other. However, the emoji embeddings significantly influence the concatenated embedding cluster results, producing high Adjusted Rand Index Scores of 0.7079 and 0.8522.

To further understand the common patterns in emoji usage, we look at the top 5 clusters. The top 5 clusters are chosen based on a high silhouette score which represents the tightness of a cluster. The silhouette score can range from -1 to 1; 1 means the clusters are clearly distinguished, 0 denotes indifference, and -1 means the clusters are assigned incorrectly. Additionally, as we aim to identify common usages, we add a condition that the clusters must contain more than 5 data samples. In tables 3 and 4, the first 3 comments are chosen as samples for the top 5 clusters in English and Spanish. The samples show that the top clusters are emoji based, which is consistent with the Adjusted Index Scores.

Comment Samples	**Cluster Size**	**Average Silhouette Score**
['Wtf Mbappe wasn’t even the man of the match 🤦', 'This is an insult\nWe lost because of their support🤦', '@blueblink5254 dude also, I am embarrassed FOR YOU! 🤦']	6	0.46

['We love you hazards🔥🔥🔥', 'The France 🇫🇷 is on Fire 🔥', 'dats literally my dad 🔥', 'Best player 🔥🔥', '@pure_sh1thousery we are winer❤️❤️🔥🇮🇷🇮🇷🇮🇷🇮🇷']	24	0.44
["@tiago.rendas2 bro I do know football but yesterday match that wasn't an offside lol 🤣🤣🤣what cause the camera showed his leg over the line but have you seen the last defender in the box 🤣🤣🤣go and learn football then come back chat to me", '@_francechy_ 🤣🤣🤣', '@5400.seconds are u there bro? 🤣']	15	0.44
['@xidiwldnxjqpspqpskxls morroco is a sibling country as well. 20% of them live in Belgium😂.', '@bf_steve Who do you think I cheer for 😂', "@ryjpopp yeah says the people who can't afford a stay in Qatar 😂"]	77	0.35
['Netherlands👏👏👏 and also cricket', 'Definitely!!! 👏👏👏👏👏', '@pr1ncesa_mar1 never! Korea is much better team! 👏']	24	0.30

Table 3. Top 5 K-means Clusters (English)

Comment Samples	**Cluster Size**	**Average Silhouette Score**
['¡El mago maestro🔥🔥🤍👏🏻⚡', 'Te amamos @yosoy8a 🇲🇽🔥', 'Vamos España 🔥']	28	0.46
['Ecuador! 🇪🇨🇪🇨🇪🇨🇪🇨❤️❤️', 'Cambien la bandera, somos Ecuador 🇪🇨☝🏻', '@jctorr130 1-1🤫🤫🤫🤫']	49	0.42
['❤️❤️❤️❤️❤️❤️ Ecuatorianos 😍😍😍😍😍', '@btito_.avak mejor portero del mundo, procede a perder 7 a 0😍😍😍', 'Eden hazard comes back😍😍😍']	11	0.41
['@magggitomcruz ya estamos en la copa, se llama USA 🇺🇸 lmao p3nd3j0 😂😂', 'Osea el uno tapa penal el otro hace gol y este es el mejor del partido es lo de ayer Ferrán hizo los 2 goles y grande jugadas y el mejor fue Gavi 😂😂', '@daje_mitico 😂😂']	67	0.40

['👏👏🙌🙌Ecuador', 'Felicidades a Ecuador por ganarle a los esclavistas 👏🏽👏🏽', '15 mundiales 👏👏👏']	25	0.31

Table 4. Top 5 K-means Clusters (Spanish)

To further visualize the clusters, we tried two methods. First, we used Principal Component Analysis to initially reduce the dimensions to 50, and then t-SNE to reduce to 2. The second method to visualize the clusters was to only use t-SNE and directly reduce to 2 components. As both yielded similar performance, the second method was chosen for simplicity. The resultant graphs for both the English and Spanish datasets (see figures 18 and 19) show that some distinct and tight clusters were able to be formed from K-means clustering. This means that there are distinct patterns in how both English and Spanish use emojis. The more scattered data points signify infrequent or unique usage of emojis.

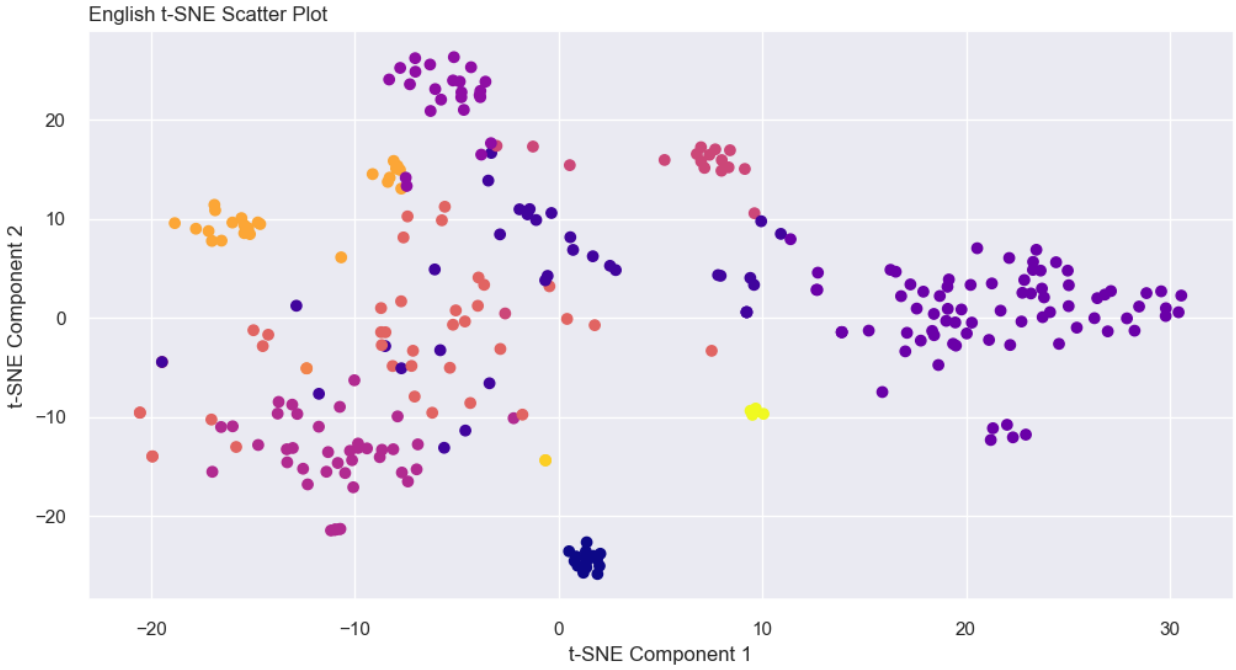

Figure 18. K-means Clustering t-SNE Scatter Plot (English)

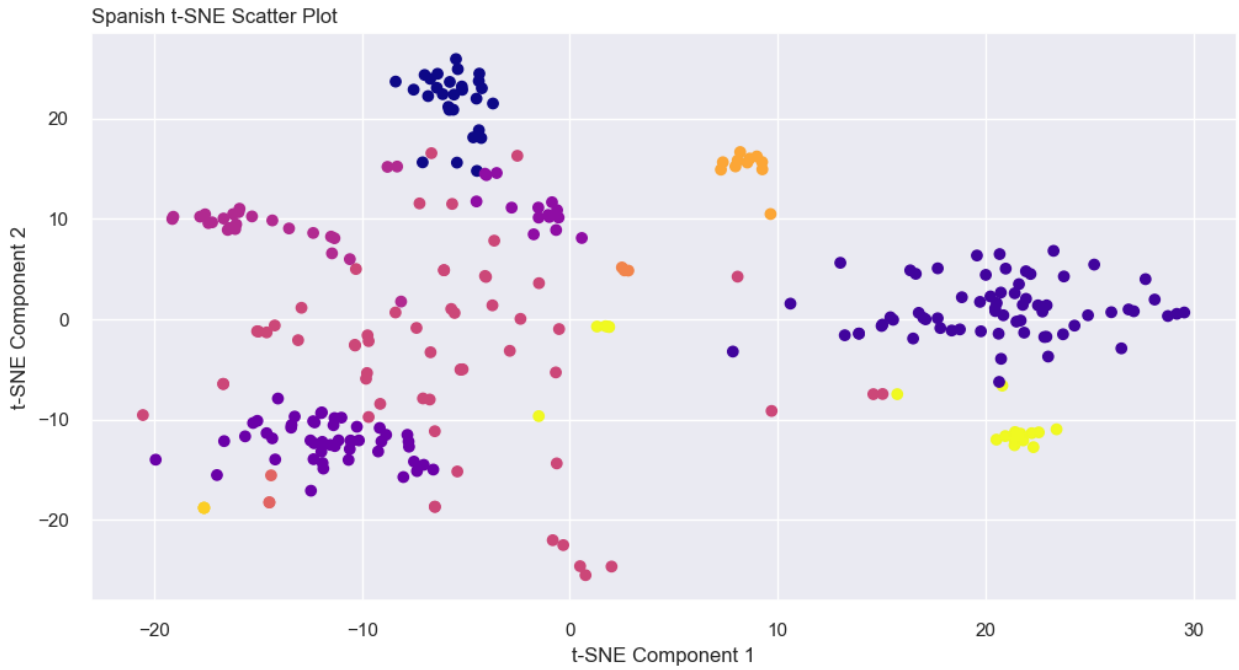

Figure 19. K-means Clustering t-SNE Scatter Plot (Spanish)

Additionally, we concatenated English and Spanish datasets to form a combined dataset, with a label to distinguish between the English and Spanish instances. t-SNE was then applied to this combined dataset to compare the English and Spanish embeddings within a common space. The graph shows that some similar clusters do form between the English and Spanish datasets (see figure 20). This aligns with the results in "3.1. Initial Exploration," where the frequency counts of emojis from both languages are similar, particularly the most frequently used emojis. From 3.1, we had inferred that English and Spanish speakers use emojis in very similar ways. In the combined t-SNE scatter plot, we see distinct clusters that are formed by both English and Spanish instances with significant overlap.

While most clusters of data points in the combined t-SNE plot show a balance of red and yellow, there also exist unbalanced clusters (see figure 20). For instance, Group 1 in the graph is mostly made up of red data points, representing English speakers. Group 2 mostly consists of yellow data points, representing Spanish speakers. To further interpret this difference, we again employed the emoji scatter plot: We labeled each data point (Instagram comment) with its most

dominant emoji (see figures 21 and 22). The resultant graphs show that Group 1 consisted of comments with the tilted laughing emoji (🤣). English speakers use this emoji much more frequently than Spanish speakers. This supports the results from hierarchical clustering. The emoji graphs also show that Group 2 consists of comments with the crying face emoji (😢), showing that Spanish speakers express sadness much more frequently than English speakers. Another distinct difference is the lack of usage of the squinting face with tongue emoji (😝) in the Spanish dataset. To interpret this observation, we can look at the comments themselves:

> ["ok will see with France 😝😘", "serbian players are shit and so are the populist criminals in serbia 😝🥶", "sadly all good things must come to an end but he will be taking this World Cup with him 😝😂"].

The 😝 emoji is used to express excitement or playfulness and hilarity. Similar to the differences found with hierarchical clustering where English use 💀 and 🤣 to convey mockery or playfulness, this same tone is also exemplified by English usage of 😝.

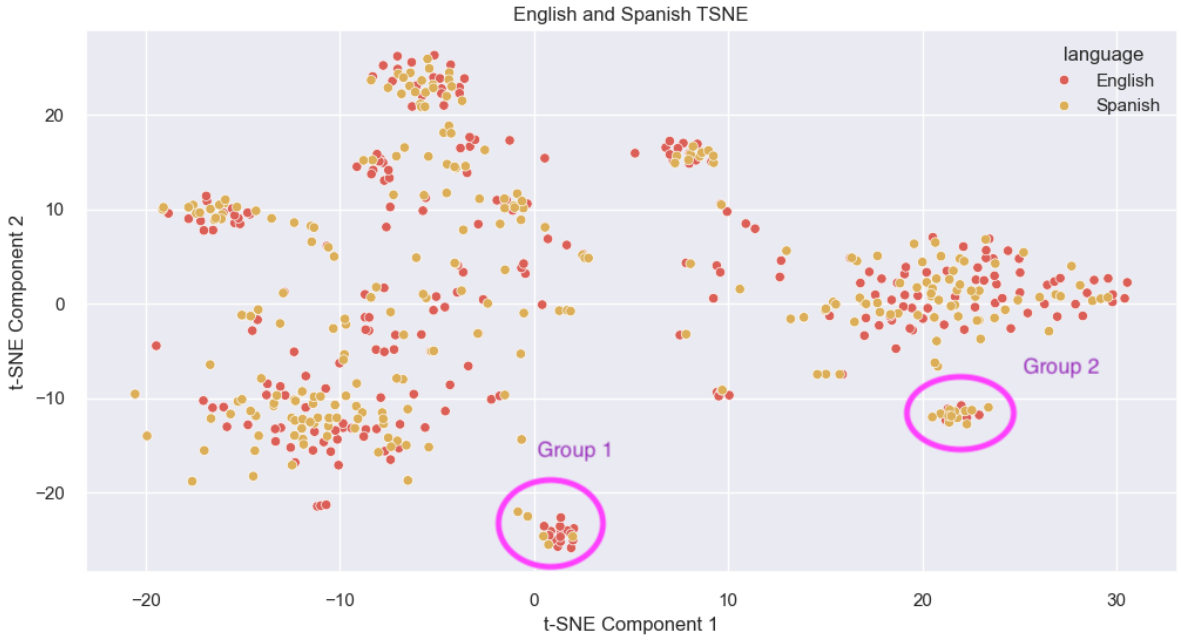

Figure 20. Annotated K-means Clustering t-SNE Scatter Plot (English and Spanish)

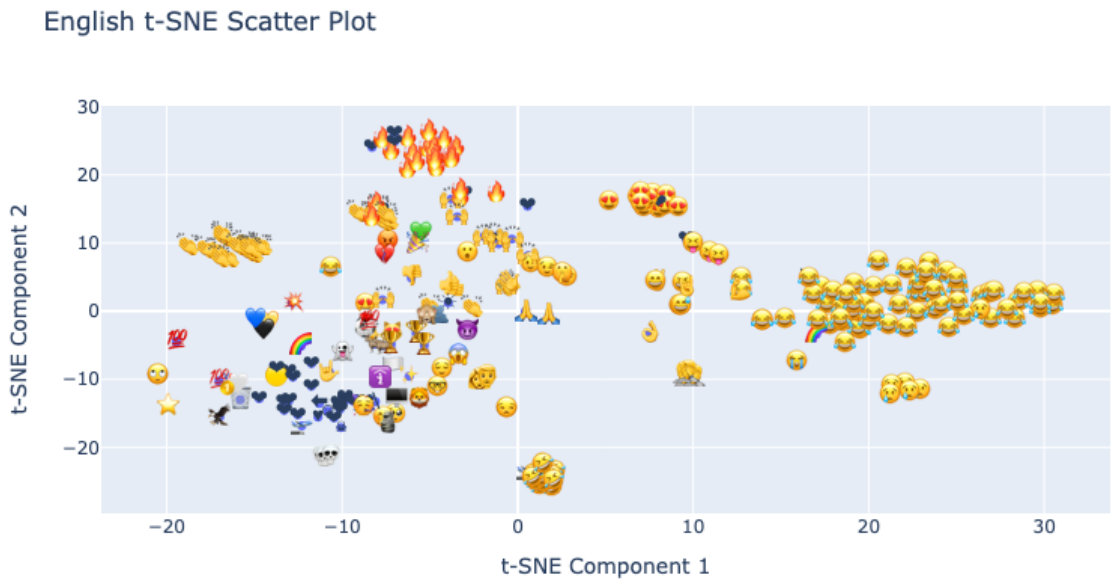

Figure 21. K-means Clustering t-SNE Emoji Scatter Plot (English)

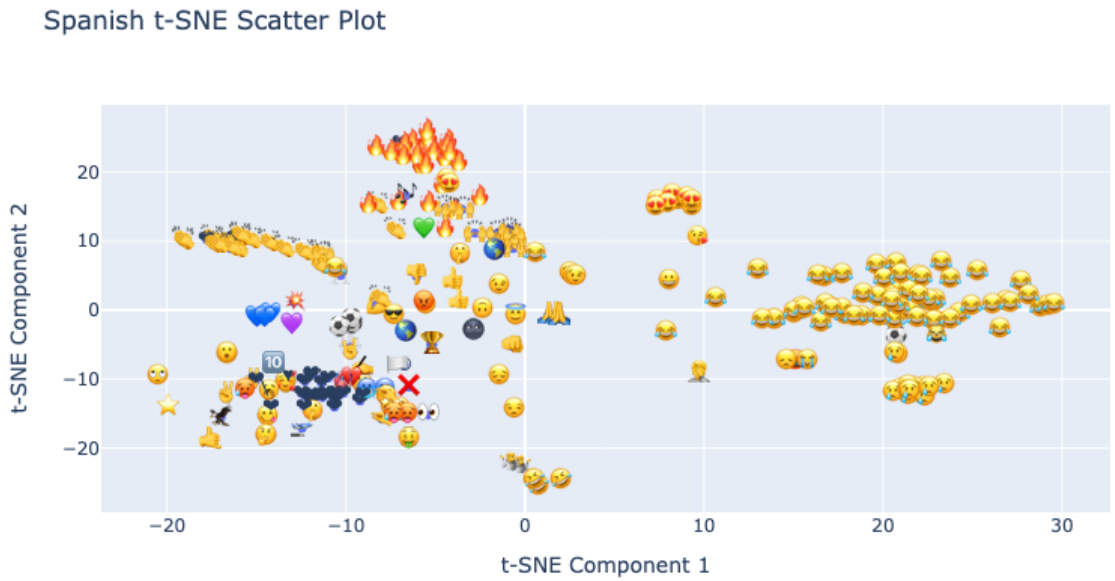

Figure 22. K-means Clustering t-SNE Emoji Scatter Plot (Spanish)

4. Conclusions and Future Work

Analyzing emoji usage during the 2022 FIFA World Cup on Instagram among English and Spanish speakers revealed both shared and distinctive cultural expressions. Commonly used emojis like 😂 and 🔥 were present in both language groups, however, emoji usage differences were also found through several methods utilizing Natural Language Processing and Machine

Learning techniques. The prevalence of the tilted laughing emoji (🤣), the skull emoji (💀), and the squinting face with tongue emoji (😝) conveying a mocking or playful tone among English speakers highlighted the unique way of expression in English. On the other hand, a tendency to repeat emojis in diverse heart colors among Spanish speakers indicated cultural ties to the national flag. Furthermore, the combination of the clapping hand emoji (👏) and the fire emoji (🔥) found in English comments implied the influence of slang in digital communication. Our analysis also showed that Spanish comments express sadness more than English comments, due to the occurrences of the crying emoji (😢).

While the above findings can be generalized to other sporting events, there are also other observations specific to the 2022 World Cup. For example, the prevalence of the rainbow emoji (🌈) to support LGBTQ rights. And it is interesting to find that this emoji only occurs in the English dataset, pointing to how English-speaking fans may be more proactive in social movements during the World Cup. These results offered a deeper understanding of cultural differences in online communication during the World Cup discourse on social media.

For future work, incorporating sentiment analysis with emoji usage analysis could provide a more detailed understanding of emotional expressions on social media. Additionally, exploring how sentiments evolve throughout matches through temporal sentiment analysis techniques and expanding the study to include additional languages spoken by participating countries would offer a more comprehensive insight into cultural differences in emoji usage.

5. References

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610. https://doi.org/10.1162/tacl_a_00288

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. The Psychological Bulletin, 83, 377–388.

Evans, Vyvyan. The emoji code: How smiley faces, love hearts and thumbs up are changing the way we communicate. Michael O'Mara Books, 2017.

Ferreira, Laura, and David B. Hitchcock. "A comparison of hierarchical methods for clustering functional data." Communications in Statistics-Simulation and Computation 38.9 (2009): 1925-1949.

Fischer, B., & Herbert, C. (2021). Emoji as affective symbols: Affective judgments of emoji, emoticons, and human faces varying in emotional content. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.645173

Gawne, L., & McCulloch, G. (2019). Emoji as digital gestures. Language@Internet, 17, article 2. (urn:nbn:de:0009-7-48882)

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Ljubešić, N., & Fišer, D. (2016). A Global Analysis of Emoji Usage. In P. Cook, S. Evert, R. Schäfer, & E. Stemle (Eds.), *Proceedings of the 10th Web as Corpus Workshop* (pp. 82-89). Association for Computational Linguistics. https://aclanthology.org/W16-2610. doi:10.18653/v1/W16-2610.

Mayank, D., Padmanabhan, K., & Pal, K. (2016). Multi-sentiment Modeling with Scalable Systematic Labeled Data Generation via Word2Vec Clustering. In *2016 IEEE 16th*

International Conference on Data Mining Workshops (ICDMW) (pp. 952-959). Barcelona, Spain. doi: 10.1109/ICDMW.2016.0139.

Sadiq, M., & Shahida. (2019). Learning Pakistani Culture through The Namaz Emoji. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-8). Sukkur, Pakistan. doi: 10.1109/ICOMET.2019.8673479.

Turak, N. (2022, November 21). *"We are very frustrated": World Cup teams in Qatar ax pro-LGBTQ armbands after FIFA threat*. CNBC. https://www.cnbc.com/2022/11/21/qatar-world-cup-2022-teams-ax-pro-lgbtq-armbands-after-fifa-threats.html

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Zhang, A. *Analysis of Emoji Use in Response to News Videos*. Columbia University. http://www.cs.columbia.edu/~jrk/NSFgrants/videoaffinity/Interim/21x_Angela.pdf

6. Appendix

Github: **https://github.com/yuhsin-huang/EmojiResearch-WorldCup2022**