Sentiment analysis of tweets about the Tokyo 2020 Olympic games

Edwin C. Montiel-Vázquez
Graduate Student
Monterrey Institute of Technology and Higher Education
Monterrey, Mexico
A01324633@itesm.mx

Yareth Lafarga-Osuna
Graduate Student
Monterrey Institute of Technology and Higher Education
Monterrey, Mexico
a00835326@tec.mx

***Abstract*—This research describes the use of sentiment analysis applied to Twitter users regarding the Olympic games of Tokyo 2020. We provide an examination of several tweets using Natural Language Processing and discrete mathematical methods, intending to obtain statistical results about the public opinion regarding a trending topic. The research uses a database of tweets published during the timeframe of the Olympic games, with each tweet's polarity cataloged to obtain information. We provide an approach that presents information about public opinion found on the Twittersphere to provide a data model, metrics, and explicative information for future use.**

***Index Terms*—Sentiment analysis, Natural language processing, Twitter, Olympic games, Trending topic**

I. Introduction

Sentiment Analysis and Natural Language processing has been used widely during the Information age to gather insight on a variety of topics [15]. Due to the advantages of high-speed communication provided by the use of social media by an increasing percentage of the population, it is now possible to gather general trends of public opinion in a variety of topics [13].

Twitter is one of the most popular Online Social Networks, this platform is qualified as an useful to tool to express personal opinions about any event in the form short sentences with less than 140 characters (micro-blog), it is easy to use and it has four ways of interactions: tweets(expressing personal opinion), likes (liking someone else content), retweets (reposting someone else content) and reply (expressing on someone else content) [14]. The importance of using the data from this social media is that the opinions/feelings expressed by users is in real-time.

Twitter is a social media site that has been previously used by organization, businesses and other actors in order to monitor the feelings of a population of people regarding subjects from brands and marketing to political movements [11], [13]. Nowadays, thanks to the advancement of technology regarding machine learning and natural language processing (NLP) it is possible to obtain real-time reactions of events happening all over the world, which is invaluable for a variety of applications like social and political analysis [11].

The Olympic games are considered the world's most prestigious sports event, with the modern era of the Olympics starting in 1896 there's a large sample of cases where they have lead to urban and social changes in the countries were they take place [5]. The importance of hosting the Olympics games lies in economic spill, the host country is in the lens of everyone in the world, and it could have a positive or negative impact.This type of event have a political and economical impact, regarding tourism, worldwide visibility, and cooperative relationships. Furthermore, the every 4 years tournament promotes cultural interchange and cooperation between the participant countries as well as inspiring future generations to follow a sport career and promoting sports events [7].

However, due to the Olympic games of 2020 having been postponed to 2021 due to the global pandemic caused by the Sars-Cov-2 virus. In the last year, many countries imposed total lockdown for their population to prevent the spread of coronavirus, sports events and practices were canceled as well [2]. People had to adapt their regular activities to being at home [2].

For health purposes, there were some adjustments and restrictions to protect the athletes and the population of the host country. Some of the measurements were no public in the tournaments except for the athletes and their teams, COVID test for athletes, and more [2]. It can be inferred that these Olympics were very different from the ones hosted in the past.

For that reason, the importance of analysing this information lies in the reaction and opinions of the general public and athletes to the event taking place. The impact of the Olympic games during the pandemic must be studied, as it will allow us to not only understand the explicit ramifications of the event, but also provide insight on other celebrations during this global crisis, as well as future ones.

Pattern classification is an area of machine learning that focuses on the use of certain characteristics unique to a series of instances for the prediction of a class [8]. We call these characteristics *patterns*. This method of classification has been previously used for a variety of purposes, from imaging to the medical field [1], [3], [9]. The approach taken by the most successful of these algorithms is in the form of contrast patterns [9]. Any pattern that is present in a significant amount of a class in a database can be considered a contrast pattern. These are useful, since classification algorithms can rely on the discriminative power to make accurate predictions [3].

One of the most important advantages for this approach of classification is in the fact that contrast patterns are useful for explaining the results to experts in real-world applications, that is, they can present explicative results [9], which can be useful for a variety of areas. In the case of sentiment analysis, these classifiers are useful since they can provide not only the polarity of the sentence, but also the reasons for this classification.

II. Method and Data

For this research, it was necessary to follow a proven methodology. We decided on using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, since it is used as a standard in business settings for decision making. This methodology is composed of six different phases: Business Understanding, Data understanding, Data Preparation, Modelling, Evaluation, and Deployment [6]. The development of each of the phases during our research will be presented and explained in the following sections and subsections of this paper.

A. Business understanding

There have been a lot of opinions regarding the postponed Tokyo 2020 Olympic games prior its inauguration. The Olympic committee took the decision of moving the games one year forward, due to that the World Health Organization declared a world pandemic in March 2020 for the cause of COVID-19 [16]. Since in most countries were a lockdown, to prevent the spread of the virus, sports events and practices were suspended [2].

Athletes and staff were forced to adapt their training at home with reduced spaces and resources. There are studies that evaluate the impact of this pandemic to mental health, it is proved that many people presented stress and anxiety [16]. This effected athletes, and their performance in sports could be reduced. However, as the vaccines released more opening events were considered to open such as the pandemic Tokyo 2020 Olympic Games.

Experts in the subject argues that it was unnecessary to gather so many athletes to compete, it was state that hosting the Olympics was a high risk of COVID-19 exposure [10]. Nevertheless, due to the importance and impact of this event the Olympic committee took the decision of having the tournament with measurements, such as no public in stadiums, to guaranty the safety of the athletes, staff and the host country population [2].

Hosting this event provided a perspective for future sports events in pandemic times. The investments of this type of events was uncertain for stakeholders, but with the organization of Tokyo2020 opened a door to continue looking for opportunities to move the economy of sports [10].

Analysis of opinions and feelings about past Olympics was performed to evaluate the social, economical and political impact [14]. The importance of analysing this information lies in the particular situation of these pandemic Olympic games. There were a lot of changes in the organization, since the general public was allowed to watch the competitions only by online distribution channels. Currently the economic impact of the host country is being evaluated, and with the study of analysing information from the general public in an Online Social Network give a perspective of the social and political impact.

B. Data understanding

In order to develop the sentiment analysis, it is necessary to obtain a database and define the structure and methodology that will be carried out. Considering that this research project is focused on the analysis of a trending topic on *Twitter*, it was required to obtain a series of tweets that matched the context of the Olympic games in Tokyo.

A database was procured through the website *Kaggle.com*. This database contained 160,000 tweets collected up until August 2021 that referenced the Olympic Games of Tokyo 2020. Aside from the text of the tweet, each data instance also contains metadata that will be useful for this research. The attributes of each instance are:

- **id**: An identifier for the database.
- **user_name**: The name of the Twitter user on the website.
- **user_location**: The location of the user.
- **user_description**: The user's personal description.
- **user_created**: The date when the user's profile was created.
- **user_followers**: The number of followers of the user.
- **user_friends**: The number of the user's friends.
- **user_favourites**: The number of the user's favorites.
- **user_verified**: Whether the user is verified or not.
- **date**: The date when the tweet was made.
- **text**: The text of the tweet.
- **hashtags**: The hashtags present in the tweet.
- **source**: The source of the tweet, meaning the device or app that was used to publish it.
- **favorites**: The number of users that added the tweet to their favorites.
- **retweets**: The number of users that retweeted the tweet.
- **is_retweet**: Whether the tweet is a retweet.

The data contained information pertaining to the nature of each of the tweets. This was important to consider since it could be used to explain various aspects of the public sentiment towards the Olympic games. Features such as available hashtags and the time period in which the tweet was made would be invaluable for this research. The reason for this is that it could provide insight on the reasons for the sentiment displayed towards on the tweet. For example, if many negative tweets were found during a day, it could be reasonably inferred that some element of the games was unpleasant during that period of time. Additionally, thanks to the use of contrast-pattern classifiers, it was possible to expect that this information could be present in patterns after the classification process process. This would be useful, since our NLP approach would provide not only the classification, but contrast patterns that could be analyzed.

While the data present on the database could be of great use, it is important to point out that information pertaining to the user or tweet that served as a unique identifier would not provide any value. If we were to take into account the identifying features of each user, we would just provide the classifier with unique values to the instance that would impact negatively the process of classification [17]. We desired to classify sentiment through the use of the tweets, as such, the features of each of tweet would be taken priority, with the user characteristics being considered as secondary.

C. Data processing

For the processing of the data, we would make use of Valence Aware Dictionary for Sentiment Reasoning (VADER). This model used for sentiment analysis for both polarities and strengths [4]. The sentiment score is calculated by an average of the intensity of all the words in a sentence or paragraph. Considering that the database originally did not contain any feature for sentiment analysis, we found it necessary to use previously available resources for the classification of each tweet with a polarity. In order to do this, the text needed to be processed. The processing of the text consisted in eliminating any possible character or symbol that did not conform to the English language, since the model was not compatible with languages other than English. To do this, we made use of different regex expressions, which ensured that the text did not contain any unknown characters or URLs, and that the emojis in each tweet were changed towards their equivalent words (e.g "happy_face" to "happy face"). No further processing for the text was required, since symbols outside the English alphabet that could present problems were included in this step [18]. The text of the tweet will be then passed through the VADER processor in order to obtain the target class: Sentiment.

Aside from obtaining the target class, modifications were applied to some of the available features. The location and the source of the tweets were encoded into unique categorical values. Additionally, any feature regarding dates were separated into its constituent components. As for hashtags, all of the unique values were recorded and categorized, since we did not think it was necessary to generate a feature for each one like in a bag-of-words model [17], we decided to only use the most common ones and represent them as unique characteristics with a binary value representing if the hashtag was present or not. It is also important to mention that the identifier, and the user's name and description were dropped from the database, since they did not contribute useful information. Finally, since the text could not be processed directly by a classifier, we opted to drop this feature as well. We desired to obtain a sentiment analysis with the use of meta data that could provide useful and explicative patterns. Therefore, the text would not be necessary for our purpose [8], [17].

While we considered using various models for text representation, such as word embeddings or bag-of-words [17], we opted to not do so, since they could not be used for obtaining explicate patterns. Most word embeddings are not in a format that can be intuitively understood by human interpreters. As such, their patterns would not be useful. Bag-of-words was the most promising one, since the use of certain words could lead to the identification of contrast patterns. However, due to the size of the necessary dictionary, we believed this to have diminishing returns in pattern-mining.

In order to provide a similar result to bag-of-words, we decided to use hashtags as our dictionary. We compiled all of the hashtags used by tweeter users in the database and compiled them into a dictionary. Only the hashtags that did not contain derivatives of "Tokyo2020 Olympic games" were kept. After this, we decided to maintain only those hashtags that were used more than 500 times in the database. The remaining 47 hashtags were then represented as binary features for each tweet. If the tweet presented these hashtags, they would be assigned a "1" in the feature.

D. Modelling

For the modelling of this task, we decided to focus on the use of a Contrast-Pattern classifier. Specifically, PBC4cip [9]. The reason for using this classification algorithm was that it has been proved to function in many circumstances for various fields [1], [3], [9]. PBC4cip was developed with the expressed purpose of being used in classification tasks with class imbalance. Therefore, we considered it useful for our research. In order to measure the performance of the classifier we would use Accuracy and the Area Under the ROC curve (AUC) [12]. Additionally, and as a control variable, we would make use of a K-nearest neighbor classifier, to measure the performance of a pattern-based classifier in contrast to a simpler classification algorithm.

This task would be evaluated using k-fold cross validation [8]. Our reasoning behind this decision was that using the Hold-out method, we would possibly make the training data biased. Since the mechanism in the algorithm hinges on the finding of patterns, we required this set of data to be as representative of the database. Therefore, it by necessity requires to be as large as possible, while not affecting the testing performance. 10-fold cross validation allowed us to circumvent this challenge [8], since it was possible for us to train on 9 out of 10 folds, while still maintaining good representation of the algorithm's performance, since all classification results would be averaged at the end of the process.

III. Results

The results of processing the data through the VADER sentiment analysis tool were interesting. First, we obtained the compound score of each tweet, which considered the probabilities of whether the text was positive, neutral, or negative and assigned them to a single numerical value between -1 and 1. In order to carry out this classification task, we decided on using three labels: Positve, neutral, and negative. The labels were assigned to equal portions of the score, this meant that the negative category would be assigned to those tweets with a value of less than -0.33 in the compound score. The neutral tweets would be those in between a value of -0.33 and 0.33.

Meanwhile, the positive tweets would be those with a value of more than 0.33.

Taking the categorization into account, the tweets were analyzed. The results of this analysis show that there was high class imbalance in the tweets. However, instead of seeing an imbalance in favor of negative or positive tweets, we saw an over-representation of neutral tweets. The category behind the neutral tweets in representation was the positive tweets. Finally, negative tweets were not very present in the database. The results from this analysis can be seen in Figure 1.

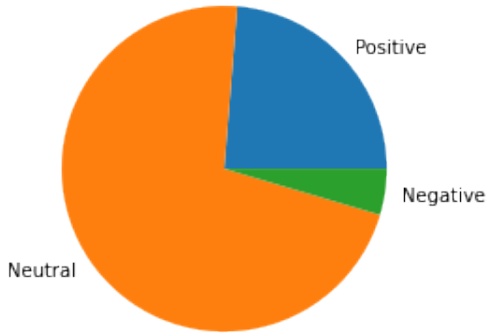

Fig. 1. Distribution of tweet polarity

In order to gain insight on the nature of the tweets, we decided to obtain the Word Cloud representation of the text found in each of the categories. The results are presented in Figures 2, 3, and 4.

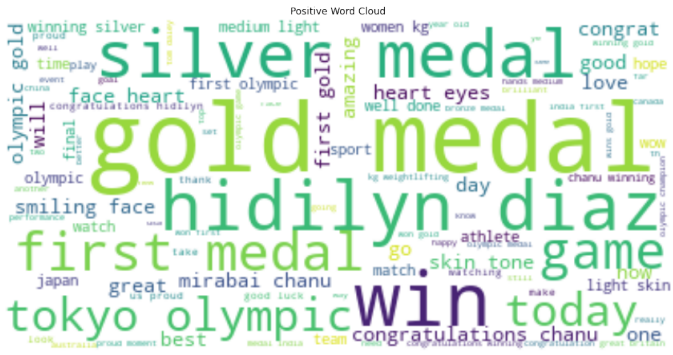

Fig. 2. Positive word cloud

In relation to the classification results, we obtained moderate performance using both classifiers. While the results from using PBC4cip were indeed better than those with the KNN classifier, they were not satisfactory to ensure that this was the best method for the analysis of the tweets. We present a comparison between the results of the classification algorithms in Figure 5.

However, it is important to point out that PBC4cip also obtains a list of patterns for each of the classes. While in this case we did not obtain any ”contrast” patterns, meaning

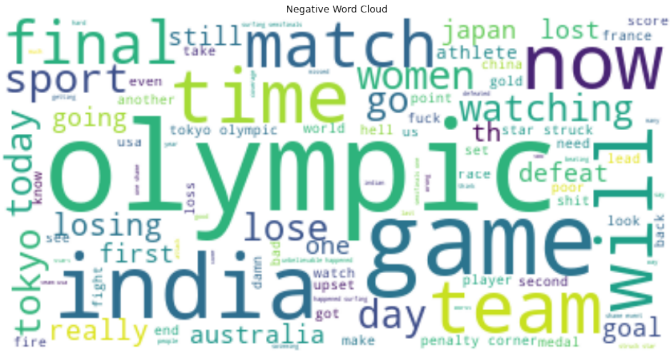

Fig. 3. Negative word cloud

Fig. 4. Neutral word cloud

patterns with more than 25% representation in their respective class, we did obtain a series of regular patterns that could later be analyzed in detail in order to gain more knowledge regarding the topic. Unfortunately, this was beyond the scope of this research.

IV. Discussion

When it comes to the analysis of the sentiment distribution, we see that the event was generally successful if we extrapolate the social media sentiment towards the general population. However, while the positive tweets are almost a quarter of the tweets about the Olympics, we see that largely the response from the population was neutral.

The Word Cloud representation shows us that most positive tweets are related to the results of the games, with consideration towards the medals won by the athletes and the countries. Additionally, we observe that emojis were used more in positive tweets. Neutral tweets represent descriptive words, which was to be expected, although some are words that could be considered positive. Meanwhile, we see negative tweets presenting some neutral words, similar to those in the neutral category. However, the negative tweets also present some clear negative words related to the results of the games, such as ”lose” or ”losing”, as well as explicit language.

In general, thanks to the analysis we see that there were not many mentions of the COVID-19 pandemic in neither of the categories. This was surprising, and it is what compels us to

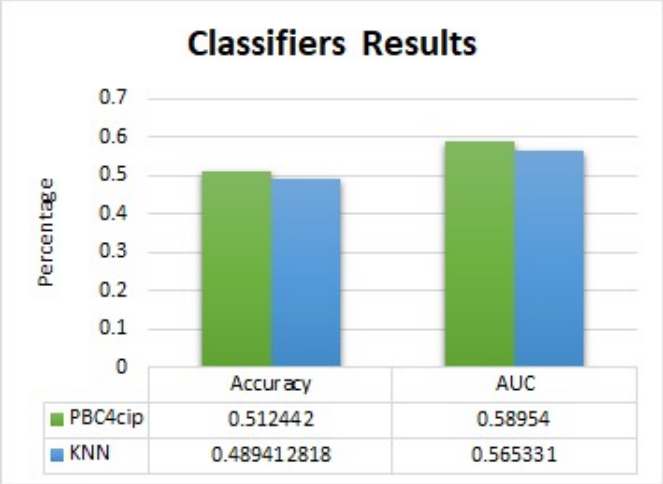

Fig. 5. Classification results

declare that the event was not unsuccessful. The fact that the view of the twitter-sphere was not largely negative towards the event even in such circumstances, makes us infer that the response was positive in general, although most tweets were classified as neutral.

In regards to classification, we propose two classifiers for possible use in future prediction of sentiment. The results were not optimal using neither of these classifiers. However, the approach taken with PBC4cip in comparison to classical machine learning approaches, in this case KNN, shows promise for future use. We attribute the results to the data representation used in this research, it is possible that the features obtained did not provide enough information to create a model that could discriminate the tweets with high accuracy. However, this can be resolved by using different features or possibly using the bag-of-words text representation for the tweets. Feature selection techniques could also be applied to improve the results. Nevertheless, PBC4cip demonstrated a higher accuracy and AUC than classical machine learning classification. Therefore, we propose to use this classifier with different features in future research.

The results show that using several sentiment analysis tools is a feasible method for gaining information regarding the public sentiment towards the Olympic event. We present a method for obtaining this sentiment, analyzing it, and tools for future prediction. The latter being the only one shown to require further adjustment. Nevertheless, since the results obtained showed a performance that was better than chance, we believe that the approach is solid, but data must be curated in regards to available features and possibly use.

V. Conclusion

We applied sentiment analysis tools based on machine learning to gather information from a public event. In this case, we obtained the general polarity of users of the Twitter social media website towards the Tokyo 2020 Olympic games. The results of this research, show that there was a general sentiment of neutrality and, in a lesser part, positive feelings towards the Olympic games. We see that the users that showed negativity towards the event in any form were in the minority. So much so, that the negative tweets were severely under-represented in our database. As for providing a method for predicting the polarity of future tweets, we see that the approach taken using patter-classifier shows more promise than a KNN model. This means that analysis of future similar events could be approached using this type of classification algorithms, albeit with different features or applying data refining techniques.

Future work could include trying to improve modelling results, as well as apply the sentiment analysis tools to different events that took place during the COVID-19 pandemic in different stages. This could help us gain insight on the progress of public opinion towards events with large crowds. Another possible approach could be the application of other libraries to see if the results hold.

References

[1] Faisal Aburub and Wa'El Hadi. A New Associative Classification Algorithm for Predicting Groundwater Locations. *Journal of Information and Knowledge Management*, 17(4):1–26, 2018.

[2] Marco Cardinale. Preparing athletes and staff for the first'pandemic'olympic games. *The Journal of Sports Medicine and Physical Fitness*, 2021.

[3] Xiangtao Chen, Yajing Gao, and Siqi Ren. A new contrast pattern-based classification for imbalanced data. *ACM International Conference Proceeding Series*, (June), 2018.

[4] Shihab Elbagir and Jing Yang. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 122, page 16, 2019.

[5] Stephen Essex and Brian Chalkley. Olympic games: Catalyst of urban change. *Leisure Studies*, 17(3):187–206, 1998.

[6] Ana Carmen Estrada-real and Francisco J Cantu-ortiz. A Data Analytics Approach for University Competitiveness : The QS World University Rankings. (1):1–33.

[7] Andrei P Kirilenko and Svetlana O Stepchenkova. Sochi 2014 olympics on twitter: Perspectives of hosts and guests. *Tourism Management*, 63:54–65, 2017.

[8] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*, volume 47. 2005.

[9] Octavio Loyola-González, Miguel Angel Medina-Pérez, José Fco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, Raúl Monroy, and Milton García-Borroto. PBC4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowledge-Based Systems*, 115:100–109, 2017.

[10] Jan Andre Lee Ludvigsen and Daniel Parnell. Redesigning the games? the 2020 olympic games, playbooks and new sports event risk management tools. *Managing Sport and Leisure*, pages 1–13, 2021.

[11] David W. Nickerson and Todd Rogers. Political campaigns and big data. *Journal of Economic Perspectives*, 28(2):51–74, 2014.

[12] Foster Provost and Tom Fawceet L. *Data Science for Business: What You Need to Know about*. O'Reilly Media, Inc., Sevastopol, California, second edition edition, 2013.

[13] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7649 LNCS(PART 1):508–524, 2012.

[14] Saurabh Sharma and Vishal Gupta. Rio olympics 2016 on twitter: A descriptive analysis. In *Computational Methods and Data Engineering*, pages 151–162. Springer, 2021.

[15] Ankit Srivastava, Vijendra Singh, and Gurdeep Singh Drall. Sentiment analysis of twitter data: A hybrid approach. *International Journal of Healthcare Information Systems and Informatics*, 14(2):1–16, 2019.

[16] Marta Szczypińska, Aleksandra Samełko, and Monika Guszkowska. What predicts the mood of athletes involved in preparations for tokyo 2020/2021 olympic games during the covid–19 pandemic? the role of sense of coherence, hope for success and coping strategies. *Journal of Sports Science and Medicine*, 20(3):421–430, 2021.

[17] Soqmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana. *Practical Natural Language Processing*. O'Reilly Media, Inc., first edit edition, 2020.

[18] Terrence E. White and Manjeet Rege. Sentiment Analysis on Google Cloud Platform. *Issues in Information Systems*, 13(2):112–122, 2012.