Special Olympics case study

ISYE 6501

Course project

Summer 2020

Outline

1. **Introduction**
2. **Keeping athletes safe: a modeling approach**
    - 2.1 Data collection
    - 2.2 Data preparation
    - 2.3 Preliminary modeling
    - 2.4 More in-depth modeling
        - 2.4.1 Predicting cardiovascular issues
        - 2.4.2 Predicting mental health issues
        - 2.4.3 Predicting seizures
    - 2.5 Closing statement
3. **Ensuring a great fan experience: sentiment analysis approach**
    - 3.1 Data collection and processing
    - 3.2 Feature extraction and modeling
    - 3.3 Closing statement
4. **Resources**

1. Introduction

This course project is inspired by how SAS Analytics was used in the Special Olympics world games in UAE in 2019[1] to ensure the safety of all the participating athletes while keeping the fans attending the games engaged. The report will explain my personal analytics approach to address the athletes' safety and the fans' engagement.

2. Keeping athletes safe: a modeling approach

The Special Olympics are nothing like the Summer or Winter Olympics. While all of the Olympic games occur every 4 years, the athletes participating in the Special Olympics have greater medical needs which require close health monitoring by the Olympic committee.

2.1 Data collection

To be able to monitor each participating athlete's health condition(s), the committee needs to define what metrics to use and how to gather such data.

Each Special Olympics athlete is different and has unique health needs. Upon registration for the Olympic games, each participant has to include their full medical history. We need to track all athletes that have pre-existing medical conditions closely while also monitoring all of the participating athletes for new or unexpected medical issues. A good way to monitor the athletes is through a wearable device that can collect biometric data. Given technological advancement in the consumer electronics domain, smart watches have become very popular for tracking physical activity and common health indicators. The following chart[2] shows some of the applications of wearable devices

We will use a smartwatch to track each athlete's biometric data. The following data can be recorded and used for our analysis

- Heart rate
- Blood pressure
- Body temperature
- Glucose rate
- Fall detection (binary variable)
- Sleep patterns
- Stress level
- Location

This data will be recorded periodically around the clock and at configurable time intervals. We can set up the devices to poll and record the data at short intervals such that we have sufficient resolution for useful analysis with each metric without producing excessive quantities of data.

2.2 Data preparation

There are many challenges to using a wearable device to track this sort of data. We have limited battery power and the smartwatch will need to be recharged or swapped at least once a day. We also need to interact with an integrated app or a third-party app through a Wi-Fi or cellular connection, which could be a problem if the athlete is somewhere where there is no network availability. Ideally, it should be possible to cache data locally on the smartwatch until connectivity becomes available again. There is also the possibility of sensors failing or otherwise providing degraded data. These challenges and others could result in missing or inaccurate data, and we need to handle those cases appropriately.

All of the recorded data will be transferred to a data warehouse server where the analysis and modeling will be conducted. We will begin our analysis with some data exploration. We will most likely have missing data, so imputation methods will be needed to estimate an appropriate value for these missing readings. We will use the mean to estimate missing data.

Dealing with outliers could be challenging in our case, because a major increase in any of the biometrics we are collecting could indicate that an athlete is having a real health issue. In this case, we can't carelessly remove outliers—we need to exercise caution and investigate outliers as potentially accurate readings. Assuming data is being processed in real time, it may be necessary to respond to outliers by contacting the athlete or a personal assistant (e.g. coach).

2.3 Preliminary modeling

Now that we have the data ready, we can begin our modeling approach. Preliminary modeling is needed because every athlete has particular health conditions and requirements. We can build a CUSUM model to detect any change in the key biometric data we collect through the watch, and then present it in a dashboard.

Any time we detect a change in any of the biometrics collected which is greater than a certain threshold for each metric, health professionals will receive an alert and can reach out to the athlete to make sure everything is ok. Consider the heart rate variable: if an athlete is believed to be in a resting state, yet their heart rate exceeds 100 BPM, the system could notify the health professionals. This initial analytics modeling approach could allow the special Olympics

committee to send out an alert to the athlete. And, if there is no response within certain time period, the committee could send first aid responders.

2.4 More in-depth modeling

Sometimes, an athlete might be experiencing a more serious condition, so waiting for them to respond may not be the best approach. Having greater modeling depth is vital if we want to respond to serious health conditions like seizures, cardiovascular issues, and asthma attacks in timely manner.

2.4.1 Predicting cardiovascular issues

In order to predict whether an athlete is suffering from a cardiovascular issue, the following modeling approach will be used:

- **Given:**
  - Heart rate
  - Blood pressure variability
  - Location
- **Use:**
  - SVM
- **To:**
  - Detect potential cardiovascular issues

Athletes with high blood pressure and/or fast heart rate may be experiencing some sort of cardiovascular issue. Using a classification model like **support vector machines**, we can classify athletes that have a fast heart hate (exceeds 100 BPM) as high risk. We can also use high blood pressure (or a combination of fast heart rate and high blood pressure) to classify at-risk athletes. We can then use the athlete's location to send medical help.

2.4.2 Predicting mental health issues

To predict if an athlete is suffering from a mental health issue, such as depression or anxiety, we can use the following predictive modeling approach:

- **Given:**
  - Sleeping patterns
  - Stress levels
  - Location
- **Use:**
  - K-means
- **To:**
  - Locate athletes who may need mental health assistance

Using these metrics, we can apply a K-means clustering algorithm to identify athletes who might need immediate psychological help. We can focus on two types of clusters:

- **Cluster 1: Athletes with high stress levels and low sleeping rate**
- **Cluster 2: Athletes low stress levels and good sleeping patterns**

We can then strategically position health professionals near cluster 1 to expedite care. We can also initiate outreach programs to target these athletes, to show support and to provide all options available in case they want to seek professional help.

2.4.3 Predicting seizures

To detect if an athlete is having a seizure, we can use the following model:

- **Given:**
  - Glucose levels
  - Stress levels
  - Body temperature
  - Fall detection
- **Use:**
  - Linear regression
- **To:**
  - Locate athletes who are at risk of having a seizure

Classification modeling here is the most appropriate since we are only classifying these data points as having or not having a seizure. Abnormal glucose levels and/or high fever can sometimes cause seizures. If the fall detection factor is high, our classification model will alert the on-site health professionals that this athlete is most likely experiencing a seizure. All the analysis would be done in real time, which could help provide the necessary medical attention to the athlete as fast as possible.

2.5 Closing statement

Applying the appropriate analytic techniques to the biometric data collected by athletes' wearable devices can help each athlete remain safe by enabling fast detection of potential issues and quick responses from health professionals. The selection of a modeling approach could make the difference between life and death for an athlete, and could ensure that each athlete is in good care as their health metrics are observed in real time around the clock.

3. Ensuring a great fan experience: sentiment analysis approach

3.1 Data collection and processing

The committee has access to data from ticketing and travel accommodations, which can be leveraged to identify and predict fan interests. Additionally, the rapid emergence of various social media outlets like Facebook, Twitter, Reddit, Instagram, etc. provides a wealth of information about the public opinion of various aspects of any athletic competition. The Special Olympics committee can mine this data for patterns and knowledge that can be utilized to make data-driven recommendations to fans attending the games or to viewers around the globe, to ensure a great fan experience. Using powerful text analytics software to comb through billions of social media posts and impressions can help reveal what fans like and dislike about the games.

Using Twitter and Facebook API's, we can mine Facebook posts and tweets. Tweets are short messages, but they can be full of shorthand, slang, and other jargon that may be difficult to parse. Facebook posts may present similar challenges, and can also be longer. Once we have the dataset comprised of Tweets and Facebook posts, we can preprocess it, extract relevant features and keywords from it, and then classify it into positive and negative classes. We can use the data from these two classes to build a data-driven analytics framework which could help the organizers improve the way they communicate with the fans and viewers during the event. This would enable more personalized messages to viewers, providing appropriate event recommendations, ticket availability, and travel times from one venue to another.

3.2 Feature extraction and Modeling

Our analysis modeling approach is to identify and classify the opinions or sentiments expressed in text by viewers. The first step in this analysis is to preprocess it. Since the data is mined automatically through an API, we will need to remove links, URLs and any irrelevant information.

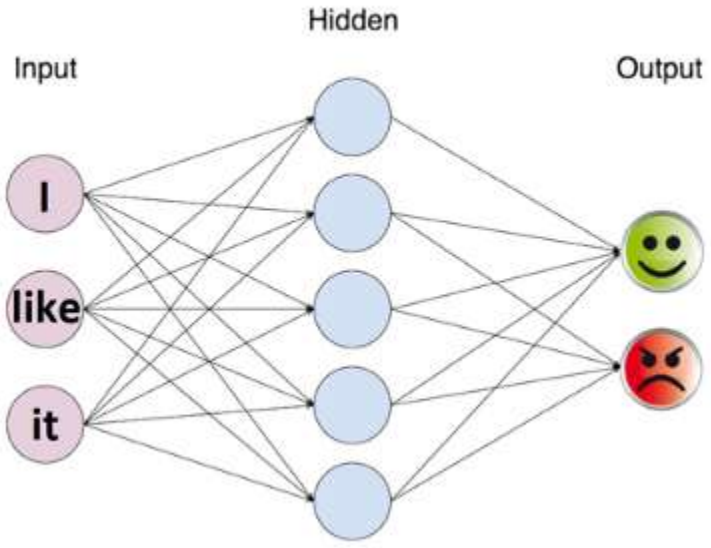
As the reader surely knows through experience, many people post social media updates with slang, hashtags, and emojis. Slang and jargon may contribute a lot to the emotion of the post. To handle these types of words, we need to construct a slang dictionary, which we can use to replace each slang word with its associated meaning. Similarly, we need to have an emoji dictionary that would have the emojis in text as keys and the associated meanings as values, which we can use

to replace the emojis in each post with their meaning. Hashtags would need to be extracted as they are.

To extract features from text, we will apply a widely used natural language processing model called a “Bag of Words”. We will process each tweets/FB post then create a vocabulary where each word has a score. We will first remove all stopwords from each list of words where every list represents a post. Stopwords are the words that do not contain much information about text like ‘is’, ‘a’, ‘the’, and many more. We will assign a score for each word on the frequency of its occurrence in a list.

Since we are modeling what fans liked or disliked about the games, we will need to maintain a list of positive and negative keywords. We will cross check our keyword lists with each list in our bag of words then we will add the count of each negative/positive keyword to our feature vector.

Now, that we have our feature vectors, we will need to split our data into training, validation and testing sets. First, we will fit a classification model using Support Vector Machines (SVM) onto our training dataset. Second, we will also architect a Neural Network (NN) model onto the training dataset. The input layer will consist of a list of words from our bag of words and the model will be constructed as seen in the chart below [3], where the output layer will indicate “liked” or “disliked” sentiments.



We can decide which model to use based on each model's performance on the validation data. Knowing that we have tons of data to train the neural net, we expect the NN model to have better performance than the SVM model. The test dataset will be used to measure the final performance of our chosen model.

3.3 Closing statement

Our sentiment analysis will allow us to distinguish between all the positive things fans said about the games, while helping us know in which areas we need to improve. This could provide a more entertaining experience for the thousands of fans attending the games and the millions of viewers tuning in.

The committee can use these results to create data-driven policies after the event. Creating new policies that are focused on making the event a pleasurable experience for the spectators will help the organizers create an all-around successful experience for everyone involved.

4. Resources

1. “World's La_NJDrgest Sports and Humanitarian Event Builds Legacy of Inclusion with Data-Driven Technology.” *SAS*, www.sas.com/en_us/customers/special-olympics-world-games-abu-dhabi.html.
2. “Big Data and Wearable Health Monitors: Harnessing the Benefits and Overcoming Challenges - Health Informatics Online Masters: Nursing & Medical Degrees.” *Health Informatics Online Masters | Nursing & Medical Degrees*, 17 Sept. 2019, healthinformatics.uic.edu/blog/big-data-and-wearable-health-monitors-harnessing-the-benefits-and-overcoming-challenges/.
3. "Sentiment Analysis: Analysis Part 3 — Neural Networks". Medium, 2020, https://medium.com/nlpython/sentiment-analysis-analysis-part-3-neural-networks-3768dd088f71.