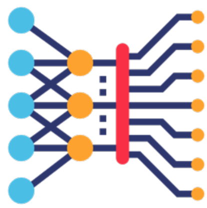
Athens University of Economics and Business
Department of Management Science and Technology

Machine Learning and Content Analytics
Instructor: Haris Papageorgiou

Detecting Heavy Drinking Episodes Using Accelerometer Data



Maria Zafeiropoulou
A.M: P2822113

Marianna Konstantopoulou
A.M: P2822122

M.Sc. Business Analytics Part Time 2021-2023
Athens, 28/08/2022

Table of Contents

Abstract...………………………………………….…………………...iii

Introduction..………………………………….……………………..iv

Our Vision……..……………………..………………………….....v

Dataset Overview…………………….……………………..….vi

Data Processing...…………………….………………………..….vii

Model Development & Evaluation…………………..………...…ix

Conclusion & Comments…………………..…………..……...…x

Members & Time Plan……………………..…………..……......xi

Bibliography………………………………………………………...xii

Abstract

The harmful use of alcohol is responsible for 5.1% of the worldwide illness burden. Recent research targeted at encouraging healthy drinking habits has showed promise for the effectiveness of just-in-time adaptive interventions delivered on mobile phones immediately before the commencement of heavy drinking episodes. This project evaluated deep learning architectures and built a reliable classifier that recognized incidents of excessive drinking using just non-sensitive accelerometer data from mobile phones.

Introduction

Frequent alcohol drinking may pose a major hazard to people's health. In reality, physicians and social workers want to reduce alcohol intake among young individuals. For example, monitoring the Transdermal Alcohol Content (TAC) can help a person stop drinking in real time. Furthermore, wearable technology might be utilized to create an alcohol detection system based on inertial signals from accelerometers in cellphones.

A recent study that offered hourly mobile treatments to participants during drinking events found no meaningful reduction in the amount of alcohol ingested, indicating that too frequent messaging might limit intervention efficacy. This highlights the need to deliver precise and focused messages to participants when drinking incidents are taking place.

This project tests an approach for combining sub-window information for an alcohol detection system based on wearable technology and deep learning, resulting in significant improvements for long windows (10 seconds). The dataset Bar Crawl: Detecting Heavy Drinking Data Set was used for this project. It includes acceleration data from 13 people during a university event as well as TAC sensor measurements. Raw accelerometer data is not sensitive; thus, it will be much easier to adopt than sensitive location, call, and keystroke data, which may create privacy concerns.

For this project we explored a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) that detect excessive drinking episodes using mobile accelerometer data.

Our Vision

Our vision regarding the business implementation of our project was that this innovation could be adopted by a huge technological company like Apple, Samsung, Xiaomi or Huawei, which are the Top 4 Wearable Device Companies by Market Share for Q1 2022 and aim to influence the fields of health, medicine, fitness and self-awareness in general. This technology could be implemented in the wearable devices and provide the users with accurate, targeted messages when a heavy drinking episode is detected therefore behaviors such as Driving under the influence (DUI) or doing recklessly activities that need concentration and clear mind would be avoided. That will result in reduction of the harmful use of alcohol and will contribute to minimizing accidents due to intoxication.

Dataset Overview

We used free source data from the UCI Machine Learning Repository: Bar Crawl: Detecting Heavy Drinking Data Set. The dataset includes acceleration signals captured at 40 Hz from a sensor integrated in smartphones, as well as TAC measures obtained using an ankle bracelet. TAC=0.08 g/dl was chosen as the degree of discrimination between drunk and sober subjects (TAC >= 0.08). Participants took part in drinking activities without being informed. We divided the participants into two groups for the categorization task: drunk and not drunk. The data set contains 14057567 instances gathered from 13 people. Accelerometer data were taken from 11 iPhones and 2 Android phones, with 5 columns: a timestamp, a participant ID, and a sample from each accelerometer axis. TAC data was obtained at 30 minute intervals using SCRAM ankle bracelets.

Data Processing

We utilized the raw accelerometer data in the x,y,z axis. The timestamp is recorded in milliseconds, with each second having a different number of readings. To ensure that all seconds had the same number of instances, each second had 20 measurements taken e.g. instances at second t are [x_t , y_t, z_t] a 20 X 3 matrix.

However, whether or not there is excessive drinking at second t is closely associated to the most recent accelerometer data, therefore we constructed an overlapping sliding window display for every 10 seconds. In particular, the total features for the second t are:

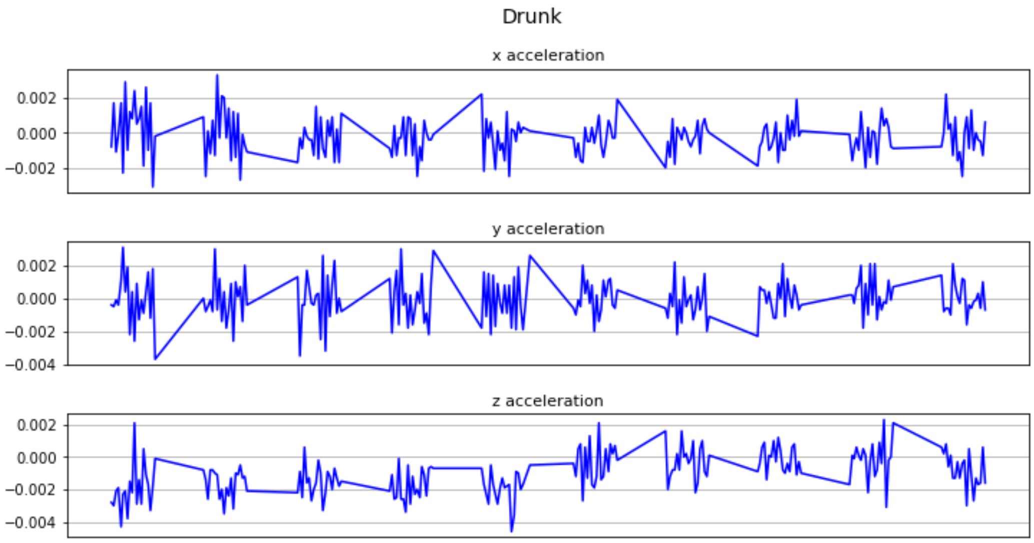
[x_{t-9} , y_{t-9} , z_{t-9}
x_{t-8} , y_{t-8} , z_{t-8}
...
x_t , y_t , z_t], a 200 X 3 matrix.

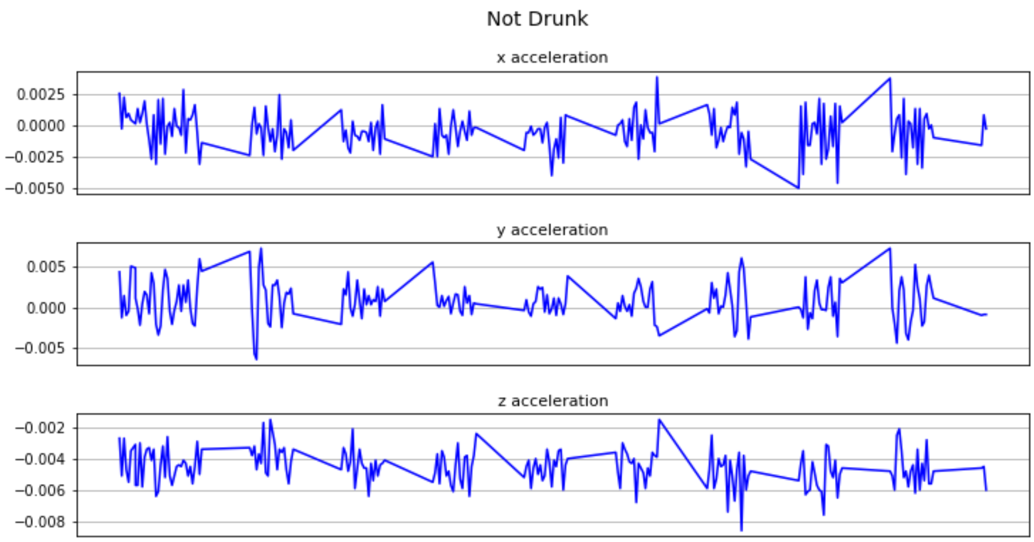
In order to reach our desirable input form, we first loaded the accelerometer data and created CSVs for the different "pid" codes. We then chose one specific pid for the training to avoid bias and created the data frame that contained time, pid, x, y, z. Following that, we added a column with seconds as the timestamp was in milliseconds and we wanted seconds for our analysis.

Next step was to load the clean TAC data set for the specific pid chosen that contained the timestamp and the TAC reading. We then combined the two data sets (Accelometer data and TAC) based on the TAC time intervals.

For our classification we would need the TAC values to be either 0 or 1 using a threshold of 0.08, with TAC values below 0.08 being not drunk and values over or equal to 0.08 meaning drunk, so we replace the TAC value column with 0 or 1 accordingly. Next, for every second we picked the 20 first instances and we added them into a new data set. Then, after creating a copy of this data set and dropping the unwanted columns, we kept the same number of rows for the drunk and not drunk and then we combined them into one data frame. We additionally checked for null values and general data information to get to know our data frame's characteristics.

For visualization reasons we created plots of x,y,z with respect to time for TAC = 1 (drunk) and TAC = 0 (not drunk), with a frame size of 20 for the sliding window:





Labeling the data set was important for the input. We dropped the time column for the training model, and we used the LabelEncoder library to label our data. We then standardized data into X and Y for training, with X being the x,y,z axis columns and Y being the label (TAC value 0,1). Finally, to match the input of the Tensor we had to transform the data set using the StandardScaler function to calculate z-score values for x, y, z and normalize the 3 readings.

Model Development & Evaluation

After doing the train-test split it was finally time to first build the Convolutional Neural Network Model and then build the Long Short Term Memory model.

Convolutional Neural Network (CNN): Our input size was 200 x 3 and we put them into two Conv1D layers with 64 filters, filter size 2 and applied activation function called Rectified Linear Unit (ReLU), then we fed them to a dropout layer with p = 0.5 and a max pooling layer with pool size equal to 2. Then we flattened the resulting array to 1D array and finally used a fully connected layer with 128 features before the final output while using a SoftMax activation function on the output of the fully connected layer. We use the binary cross entropy as the loss function and Adam as the optimizer. It received an accuracy of 91% on the test set. Tuning the filter size to 5 pushed the accuracy to 95.3% which is the best model for this project. By removing the drop out layer the training accuracy increases from 95.5% to 97.5% while the test accuracy drops from 95.3% to 91.5% which indicates a slight overfitting. The increase of hidden units' value for the last fully connected layers also shows signs of overfitting and fails to improve the model performance.

Long Short-Term Memory Neural Network (LSTM): For the LSTM our input was also size 200 x 3 and was fed in an LSTM layer with 64 units and a drop out layer with p = 0.5. Then we flattened the result to 1D and used a fully connected layer with 128 units using a SoftMax activation function and finally, as for the CNN, we used the binary cross entropy as the loss function and Adam as the optimizer. It received an accuracy of 71% which is worse than the CNN model.

An explanation for the poor performance of the LSTM is that the prediction aim is to predict drinking episodes and since the accelerometers in the previous 10 seconds all weigh equally, meaning that any abnormal pattern in the 10 second window may suggest a heavy drinking episode. As a result, CNN extracts more efficient features than LSTM, which weights the final row higher than the first rows and hence may disregard signals in the first seconds.

Conclusion & Comments

In this project we explored 2 deep learning techniques a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM). The best results came from the CNN model with accuracy 94 %. The model contained two consecutive Conv1D layers with 64 filters, filter size 5 and ReLU activation function, then a dropout layer (p = 0.5) and a 2 x 2 max pooling layer followed by a fully connected layer with 128 features and the final output layer.

The performance of CNN on classifications on inharmonious patterns using pure accelerometer data was impressive and could be performed for similar experiments utilizing the accelerometer data.

It is highly fascinating to detect alcohol intake using wearable technology and deep learning in order to avoid accidents. By presenting a way to analyze long periods of time, this project contributes to the supervision of alcohol consumption using acceleration signals. It is possible to limit the amount of parameters to be trained in the CNN by using these sub-windows, as well as the danger of overfitting.

Members

Our team consists of two members:

- **Maria Zafeiropoulou:** Maria is currently working as a Data Scientist at Unilever, the global consumer goods company. She pursues the postgraduate degree of Business Analytics at Athens University of Economics and Business. Maria also holds a bachelor's degree in Mathematics and Applied Mathematics at University of Crete and a master's degree with honors in Mathematical Finance and Risk Analysis at National & Kapodistrian University of Athens.

- **Marianna Konstantopoulou:** Marianna is currently working at Intrum Global Business Services as a Data Project Coordinator. She also has a 2-year experience in Automotive industry where she worked at the Analysis & Reporting department of a well renowned Global Automotive Business Intelligence corporation. In terms of academics, Marianna is currently a postgraduate student in the program of Business Analytics at Athens University of Economics and Business. She also holds a bachelor's degree in Mathematics with a specialization in Statistics and Operational Research from the University of Patras.

Time Plan

Project Conceptualization (final decision)	30/06/2022
Dataset search	Until 30/06/2022
Project Proposal	26/07/2022
Data preprocessing	15/07/2022-30/07/2022
Construction of models	10/08/2022-18/08/2022
Evaluation	18/08/2022-22/08/2022
Report	22/08/2022-28/08/2022

Bibliography

[1] Killian, J.A., Passino, K.M., Nandi, A., Madden, D.R. and Clapp, J., Learning to Detect Heavy Drinking Episodes Using Smartphone Accelerometer Data. In Proceedings of the 4th International Workshop on Knowledge Discovery in Healthcare Data co-located with the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) (pp. 35-42).

[2] https://archive.ics.uci.edu/ml/datasets/Bar+Crawl%3A+Detecting+Heavy+Drinking

[3] Manuel Gil-Martín, Rubén San-Segundo, Cristina Luna-Jiménez, Alcohol Detection over Long Periods Using Smartphone Accelerometer Data, SIGNAL 2022: The Seventh International Conference on Advances in Signal, Image and Video Processing

[4] Bo Yang, Deep Learning to Detect Heavy Drinking Episodes Using Smartphone Accelerometer Data, Department of Computer Science Stanford University

[5] Nikhil Doifode, Yasha Singh, Implementation of Paper: Learning to Detect Heavy Drinking Episodes Using Smartphone Accelerometer Data