Applying Symbolic Regression to Predict Human Heart Rates using Co-occurrence Graph-based Representation on Twitter Tweets

Xin Wang[1]

[1] Binghamton University, New York, USA
xwang314@binghamton.edu

Abstract

Human emotional change usually leads to internal changes in the human body, such as blood pressure and heart rate changes. Emotional change can be reflected through human-generated textual content. This study conducts the experiment through a dataset containing 2040 pairs of heart rate and Twitter tweets, collected from one 35-year-old participant. In the current stage, this study establishes co-occurrence graph-based representation, which can cover all possible emotional expressions, such as explicit adjectives and nouns with emotional meanings, in Twitter tweets. Node and edge features are generated by constructing a context-document co-occurrence matrix. To reduce the dimensionality of the feature data, a genetic algorithm is utilized for feature selection. Symbolic regression is utilized to find out a mathematical relationship between graph-based features and human heart rate. Moreover, this study applies network analysis to understand the power law behavior reflected in the proposed graph-based representation. The selected features are obtained from the genotype with the optimal fitness score, after the training of the 1000 generations by the genetic algorithm. The comparison of heart rate prediction shows that LightGBM can achieve better performance than symbolic regression. The future direction of this study is discussed based on the results in the current stage.

Introduction

Many studies have shown the relationship between the emotions from user-generated content on social media and the physical status of human beings. Early research has revealed that emotional changes will cause changes in heart rate (HR), heart rate variability (HRV), and contractility (Köbele et al., 2010). Through analyzing and establishing a model on the posted tweets, atherosclerotic heart disease (AHD) mortality can be predicted successfully at the county level (Eichstaedt et al., 2015). Most of the related studies utilized sentiment classification methods, and their results rely on the labeling quality of the emotional corpus. Shaver et al. (1987) outlined a three-level hierarchy of emotions and categorized words within the hierarchy. However, the richness of people's emotional language expression is improved with the progress of the areas. For example, "Black Friday" can represent exhilaration and happiness.

Another limitation influencing the performance of sentiment analysis performance is caused by the traditional text representation method, such as the bag-of-words model and vector space model, which cannot capture the semantic relationships in words (Bijari et al., 2020; Wael and Arafat, 2023). To understand the context of social media in-depth, this study utilizes a co-occurrence graph-based representation to characterize the relationships between various components of the Twitter user's posts in an entire timeline.

Methodology

Data

The heart rate and tweet sentiment experimental data are obtained from the study of Salamon et al (2017). They recorded one participant's daily tweets and collected measured heart rate 24×7 hours per day by a wearable monitor in the observed 50 days. The participant in their study is a 35-years healthy male having high blood pressure treatment. The heart rates of this participant in various circumstances are demonstrated in Table 1.

Circumstances	Heart rate (bpm)
Regular physical activity	94
Physical inactivity (rest HR)	72
Higher physical activity (sports)	100

Table 1: Heart rate in different circumstances of the participant.

Their experiment duration is divided into two different periods. One is from May 10th to June 28th, 2016, and the other is from August 15th to October 3rd, 2016. The data statistics for heart rate and tweets are presented in Table 2.

Information	Period 1	Period 2
Starting time	2016-05-10 00:00:05 CET	2016-08-15 00:00:00 CET
Final time	2016-06-29 23:58:55 CET	2016-10-04 23:59:00 CET
Total number of heart rate records	411,799	69,941
Average number of heart rate records per day	8074	1371
Total number of tweets	1029	1017

Average number of tweets per day	20.56	20.32
Number of days with over 20 posted tweets	35/50(70%)	35/50(70%)

Table 2: Statistics of heart rate sensor data and participant-posted tweets.

As shown in Figure 1, each tweet was recorded along with a short timeline, during which the participant edited his tweet and posted it on Twitter. This study extracts all heart rate data for each timeline of a tweet and selects the maximum heart rate during that period as the target heart rate that will be predicted through the regression model built on textual emotion features. The reason for choosing the maximum heart rate is that this study wants to predict the heart rate aroused by certain emotions. After this data extraction process, this study obtains 2040 tweets along with the corresponding heart rates.

Figure 1: Timeline of participant's Twitter posting.

Data Processing

Since the text data from social media is not in normal writing and contains typos and miss spellings, it is necessary to go through the text cleaning process. This study utilizes multiple filters to clean the text data as presented in Table 3. The cleaned texts are used to create graph-based representations.

Filter	Function
Http filter	Remove url
Username filter	Remove the attached username in text
Space filter	Remove all additional white spaces
Emoji filter	Remove all emojis, emoticons, symbols, pictographs, transport, map symbols, flag symbols
General filter	Remove suffix with stem
Specific filter	Replace abbreviated expressions
Case filter	Change all to lowercase

Table 3: Filters applied in text cleaning process.

Co-occurrence graph-based representation for Twitter tweets

Since the Twitter tweets are not computable textual data, feature extraction is a crucial step for natural language processing tasks, which can transform textual data into its features for modeling purposes. This study applies co-occurrence graph-based representation, which can map the richness of Twitter tweets, as an alternative to traditional feature extraction methods, such as the commonly used vector representation.

For the feature extraction task on Twitter tweets, a non-directed graph presentation based on a co-occurrence graph is established. The goal is to obtain node and edge features through the interaction of terms on the lexical-syntactic structure of Twitter user's tweet texts in an entire timeline. The proposed graph is described by $G = (V, E)$, where:

1. $V = \{v_1, v_2, \dots, v_n\}$ is a finite set of nodes that contains all terms tokenized from all Twitter tweets.
2. $E = \{e, e_2, \dots, e_n\}$ is a finite set of edges which represents the connection relationship between two nodes in each Twitter tweet.

Figure 2 shows an example of this graph, which is generated through three sample Twitter tweets as listed in Table 4. It is clear to see that each node is connected following the syntactic sequence of each Twitter tweet. Besides, several edges, such as ("it", "is"), are thicker than the other edges, which means that the two node terms are next to each other several times in the Twitter tweets.

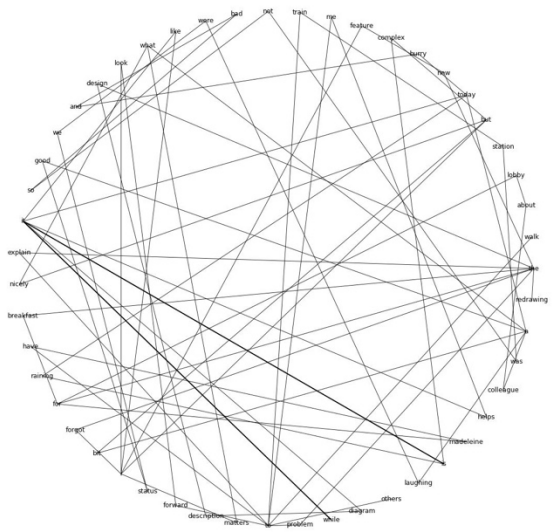

Figure 2: Example of the proposed co-occurrence graph.

No.	Text
1	"Walk to train station wasn't so bad and hurry today. It's raining today a bit, but nicely. What matters I forgot the breakfast. #p #xfb"
2	'What a good status. We were laughing a colleague about lobby for the new feature. I look forward to have Madeleine for breakfast. #p #xfb'
3	"Redrawing the design description diagram. It's complex, but I like it, while it helps me to explain the problem to others. #p #xfb"

Table 4: Examples of Twitter tweets.

Based on the constructed graph representation, this study further generates the context-document co-occurrence matrix, in which the context consists of the nodes and edges obtained from the graph-based representation, and the document corresponds to each Twitter tweet. For each document, a context-document co-occurrence statistic is conducted, as shown through an example in Table 5.

No.	i	like	dislike	summer	(computer, science)	(bad, news)
Doc_1	1	0	1	2	0	1
Doc_2	1	2	0	0	1	1
Doc_3	2	1	1	1	0	1

Table 5: Example for context-document co-occurrence matrix.

To alleviate the sparsity of this generated context-document co-occurrence matrix, a normalization is further conducted on the matrix using scikit-learn StandardScaler before the modeling process.

Genetic algorithm implementation for feature selection

In the real situation, a graph-based representation built from user-generated content on social media can have millions of nodes and edges, which are much greater than the total number of the user's posts. Thus, this study designs a genetic algorithm approach to implement feature reduction, to reduce the dimensionality of the generated context-document co-occurrence matrix.

Based on natural genetics and biological evolution, the genetic algorithm is a stochastic method prosed by Holland in 1975 (Hayes-Roth, 1975), which has been commonly applied for optimization problems, such as feature selection (Zhou & Hua, 2022; Khan et al., 2022). In this study, the features are the nodes and edges. Thus, the goal of feature selection is to find out an optimal combination of these features, which can be applied for regression model trained to predict human heart rates corresponding to certain Twitter tweets. To simplify the feature selection process, this study creates a unique ID number for each node or edge feature, so that the feature selection problem can be transformed to find out an optimal number sequence, regarded as a genotype in the genetic algorithm, to represent a feature combination.

In this study, the genetic algorithm approach is designed with five processes, including initialization, evaluation, selection, recombination, and mutation. The initialization process will generate the population with random number sequences of feature ID. The evaluation process will calculate fitness for each number sequence. Since these features will be used for regression model prediction, the root mean squared error (RMSE) is set as a fitness function, which means the best-performing genotype can achieve the best prediction accuracy for human heart rate prediction based on node and edge features. The roulette wheel method will be conducted to select two parents in the selection process. Through recombination, two parents can produce two offspring with a certain probability. Each item, a feature ID, in a genotype can choose to mutate to any other feature ID that has not appeared in this genotype, with a certain mutation rate.

When each generation goes through the evaluation process, heart rate data and a part of feature data are selected via the genotype containing feature IDs and divided into training and testing sets. The training sets are used to train a LightGBM regression model, and the model prediction is performed on the testing set. The RMSE is calculated between the predicted and actual heart rates from the testing set.

Heart rate prediction using symbolic regression

Symbolic regression model is usually used to search the space of mathematical expressions that can perfectly align with a given dataset. This study further performs a symbolic regression model on the optimal features selected by the genetic algorithm approach. The goal is to explore a mathematical expression relationship between a person's habitual language expression and specific heart rate status.

Results

Network analysis

To further understand the proposed graph-based representation, this study analyzes the degree distribution of this network with 3,713 nodes and 20,503 edges, as shown in Figure 3 (a). Several small hubs can be seen, which are caused by the nodes with a much higher degree than most other nodes. The average degree of this network is about 24. Based on the shape of the degree distribution, it is possible to say that a power law behavior exists on this network. Figure 3 (b) shows its complementary cumulative distribution function (CCDF), in which the plot line seems straight. This study further estimates the scaling exponent of the power law from the distribution by a simple linear regression, as shown in Figure 3 (c).

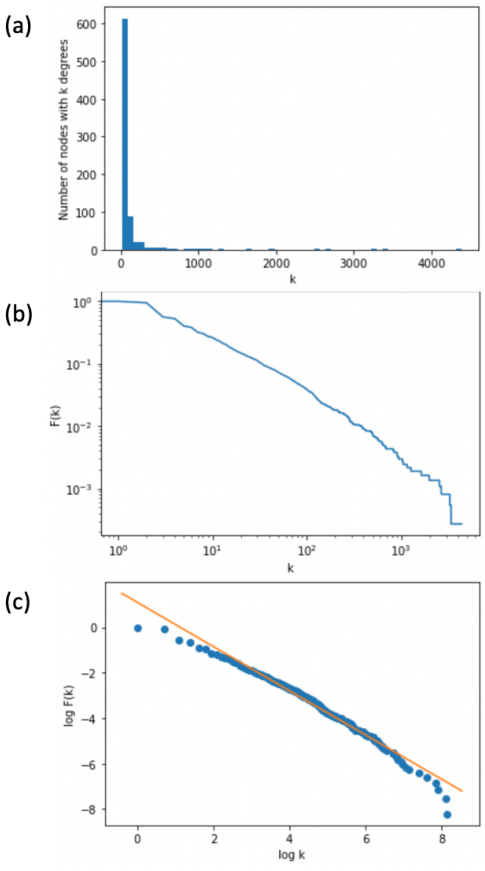

Figure 3: Network analysis.

The CCDF is presented in Formula 1. According to this estimated result, the CCDF has a negative exponent of -0.973, and the actual scaling exponent γ is about 1.973.

$$F(k) = 2.987\, k^{-0.973} \quad (1)$$

Feature evolution analysis

This study sets the genetic algorithm parameters as shown in Table 6. After all the iterations for all the generations, the selected features are contained in the generated genotypes.

GA parameters	Value
Population size	10
Number of generations	1000
Genotype size	5
Recombination rate	0.5
Mutation rate	0.3

Table 6: Genetic algorithm parameters.

Based on the feature IDs, this study further decodes the sequence of feature IDs in each generated genotype. The word cloud in Figure 4 presents the exact node features. The font size and color indicate the frequency of a feature in all the genotypes of the last generation.

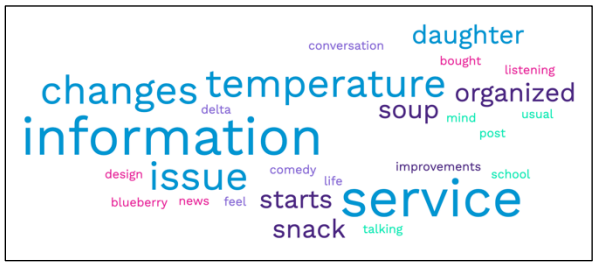

Figure 4: Word cloud on the selected features.

Heart rate prediction performance analysis

One genotype in the last generation is selected with the best fitness score, to further analyze the heart rate prediction performance through the model built by the best group of features. Table 7 shows the performance comparison between symbolic regression and LightGBM. The R2 score of the symbolic regression model is a negative value, which means that the model's predictions are not ideal.

Model	**RMSE**	**R2 score**
Symbolic regression	0.999	-0.051
LightGBM	0.973	0.003

Table 7: Prediction performance.

Discussion

Even though the regression models cannot achieve good results for human heart rate prediction in the current stage, the possible directions still can be found to optimize the results. To further release the problems on sparsity and high dimensionality of the co-occurrence graph-based representation, this study will set window size, so that those terms having co-occurrence inside a distance scope can all be considered.

From the perspective of the prediction model, the selected features in Figure 4 have proved that certain nouns that are not adjectives still can contribute to the human man heart rate change. That means that these nouns also have emotional attributes. To find out those emotional-related nouns or events from a person's Twitter posts in the entire timeline, this study will create word-word co-occurrence graph-based representation and use a ant clustering algorithm to observe the clustering results of the emotional terms.

Conclusion

Instead of using the traditional text representation methods, this study implements a co-occurrence graph-based representation of the short user-generated content on Twitter. The results of this study have shown that this co-occurrence graph-based representation will be able to interpret the richness and complexity in the emotional language expression of user-generated content on social media in an entire timeline.

References

Bijari, K., Zare, H., Kebriaei, E., & Veisi, H. (2020). Leveraging deep graph-based text representation for sentiment polarity applications. *Expert Systems with Applications*, 144, 113090.

Etaiwi, W., & Awajan, A. (2023). SemanticGraph2Vec: Semantic graph embedding for text representation. *Array*, 100276.

Hayes-Roth, F. (1975). Review of" Adaptation in Natural and Artificial Systems by John H. Holland", The U. of Michigan Press, 1975. *ACM SIGART Bulletin*, (53), 15-15.

Khan, A. H., Sarkar, S. S., Mali, K., & Sarkar, R. (2022). A genetic algorithm based feature selection approach for microstructural image classification. *Experimental Techniques*, 1-13.

Köbele, R., Koschke, M., Schulz, S., Wagner, G., Yeragani, S., Ramachandraiah, C. T., ... & Bär, K. J. (2010). The influence of negative mood on heart rate complexity measures and baroreflex sensitivity in healthy subjects. *Indian Journal of Psychiatry*, 52(1), 42.

Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. (2013). Practical graph mining with R, chapter Introduction. Chapman & Hall/CRC, pp. 1–7.

Salamon, J., & Mouček, R. (2017). Heart rate and sentiment experimental data with common timeline. *Data in brief*, 15, 851-861.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6), 1061.

Zhou, J., & Hua, Z. (2022). A correlation guided genetic algorithm and its application to feature selection. *Applied Soft Computing*, 123, 108964.