A Machine Learning-Based Lifestyle Recommendation System using Natural Language Processing

Senbagavalli M[1], Shashi kant Gupta[2]
[1]Post Doctoral Researcher, LinColn University College, Malaysia; [2]Adjunct professor, LinColn University College, Malaysia.
[1]pdf.Senbagavalli@lincoln.edu.my, senba1983@gmail.com ; [2]shashigupta@lincoln.edu.my

---

Abstract: The Sustainable Living Advisor is an intelligent recommendation system powered by AI for processing an individual's lifestyle narratives to provide insights on sustainability through Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system takes textual data on a person's habits, consumption patterns, and environmental behaviors and predicts a Sustainability Score between 0 and 10. The model checks a range of environmentally positive and negative behaviors using TF-IDF vectorization, sentiment analysis via TextBlob, and keyword-based sustainability indicators. A synthetic dataset of sustainable, unsustainable, and mixed lifestyles was generated to train regression-based models, namely Random Forest, Gradient Boosting, Linear Regression, and Support Vector Regression (SVR). Among these, one is chosen as the best for predicting lifestyle scores and interpreting them based on R^2 and MSE scores. The suggested system provides a numerical sustainability score along with a detailed analysis of the environmental footprint, which ranks under waste management, transportation, energy consumption, food choices, and general consumption. It gives small eco-friendly recommendations to aid users in more sustainable options. Such associations of environmental sciences and artificial intelligence provide a unique means to promote sustainability from the perspective of personalized data analysis. The Sustainable Living Advisor works towards raising awareness and enabling analytical decision-making for a greener, more responsible lifestyle by analyzing real behavioral reflections in text.

Keywords: TF-IDF (Term Frequency - Inverse Document Frequency), Machine Learning, Natural Language Processing, Support vector Regression, Gradient Boosting

---

Introduction

The transformation of lifestyle practices into sustainable living has, by far, become an increasingly hot topic around the world, with water pollution, climate change, resource depletion, and environmental degradation regarding the importance of humankind in the development of societies on an environmental footprint. An individual having an awareness of his sustainability can initiate real change, but in practical terms, most people have no real option to measure, at least in academic consideration, the sustainability of their lifestyle. By using a variety of artificial intelligence (AI) and natural language processing (NLP) methods to evaluate lifestyle trends and offer astute suggestions on eco-friendly living, the Sustainable Living Advisor project sought to close that gap.

Text analysis, machine learning, and sentiment scoring of self-descriptive lifestyle inputs—like "I recycle my waste" or "I use disposable plastic bottles"—are used in the study to score sustainability. Eco-positive and eco-negative nomenclature are represented by cut-and-dried dictionary entries, which provide the theoretical basis for the system's interpretation of the given response. Advanced tokenization, lemmatization, and TF-IDF vectorization are examples of state-of-the-art preprocessing techniques that reduce the likelihood of any information going through text being incorrectly encoded for machine interpretation. In order to highlight any emotional or attitudinal tones users convey through their words, it incorporates the examination of sentiment polarity and subjectivity.

The System was developed and trained using a synthetic data set developed with three distinctive behaviors to assess an individual's sustainable behavior: sustainable, unsustainable and mixed behavior. The model was then trained using multiple different regression algorithms such as Random Forest, Gradient Boosting, Linear Regression and Support Vector Regression, to predict the sustainability score from a combination of TF-IDF and custom environmental features. After thorough evaluation and testing using multiple metrics including R2, MSE and MAE, the best performing model was chosen. The proposed system merges data and technology with environmental learning, creating a dual-function system. This system provides users with both a sustainability rating service as well as an educational portal that will offer users recommendations or insights on how to change their behavior to help them live in a way that is more sustainable. Additionally, this project will combine the fields of artificial intelligence (AI), natural language processing (NLP), behavioral science, and sustainability research to support the emerging field of AI applications that promote environmental responsibility and the concepts associated with the Sustainable Development Goals.

Related work

Explainable Natural Language Processing for Corporate Sustainability — Ong et al. advocate for the application of explainable NLP (XNLP) techniques to enhance clarity in analyzing corporate sustainability disclosures. The authors argue that explainability is needed for stakeholders to comprehend sustainability judgments made via models and propose methods to highlight model rationales (important phrases, attributions) that are consistent with sustainability taxonomies [1]. Social Media Data for Environmental Sustainability (review) Ghermandi et al. analyze all systematic evidence for using social media in environmental and sustainability research, covering topic detection, sentiment analysis, geospatial mapping, and behavior change studies. They conclude that social media yields large-scale behavioral signals and detects public responses to sustainability campaigns; however, they highlight biases and representativeness issues [2].

Advancing Sustainability via Recommender Systems — A survey for 2024/2025 analyzes how recommender systems can be redesigned to promote sustainability (e.g., recommending low--carbon products or nudging low-impact choices) and presents research avenues such as multi- objective optimization and lifecycle-aware recommendations. The review presents design patterns and evaluation metrics that are pertinent to the building of the advisor's recommendation module (how to rank and present green alternatives) [3]. This qualifies as one of the many Natural Language Processing Methods for Scoring CSR / Sustainability Reports. - Gutierrez-Bustament et al. applying TF-IDF, supervised classifiers

and lexicon-based methods to score corporate social responsibility (CSR) content and retrieve topic- level signals from reports. They prove that word-level features combined with domain lexicons improve the detection of sustainability topics and produce reasonably robust automated quantification. These findings support the TF-IDF + keyword approach, with even suggestions for improvements such as domain lexicons and topic-level scoring [4].

ClimateQA / Analyzing Sustainability Reports Using NLP — Climate focused QA works on the development of a question-answering model that serves the purpose of addressing long financial/sustainability reports in the scope of climate and sustainability content. They show how specific applications of custom QA or information-extraction models could retrieve quite fine- grained facts from such long documents (emission targets or mitigation actions, for example): a pattern that you will probably want to adapt for extracting specific behaviors from long user descriptions or transcripts [5]. Usage of Text Mining and NLP in the Analysis of Sustainability Reports—A Systematic Review- Mohammadrezaei et al. conduct a systematic review of text mining/NLP applications for sustainability reporting, compiling the choice of their preprocessing methods, features engineering, and prevalent common evaluation procedures given different studies. Key findings reveal a spread leeway for TF-IDF and word embeddings, a recurring domain-specific lexicon, a widespread need for benchmark datasets, and it provides a solid methodological reference for your pipeline features and synthetic-data assignments and their evaluation [6].

This article by Felfernig et al. provides a summary of the scholarly literature on recommending systems that have been developed to assist the accomplishment of sustainability goals as well as define considerations for determining system sustainability (the energy input for the recommendation system) and user-side sustainability (changes made by users of the recommendation system). Key issues for multi-criteria evaluation (user utility and environmental impacts) and the discussion on explainability versus nudging will need to be considered when selecting the recommending engine for the advisors' recommendations and communicating the trade-offs between the recommendation to potential end users. New CID/empirical research has demonstrated that companies can identify claims of sustainable development in private sector sources (e.g., documents, web pages) using methods like TF-IDF, logistic regression, and careful stopword selection. The authors of this study emphasize the importance of selecting relevant features and augmenting the dictionary to improve the ability for the model to differentiate marketing terminology from real company sustainable development initiatives. As a result, this article makes a strong case for using well-defined keyword lists and closely monitoring false positive errors in the keyword scoring process when developing targeted sustainable actions for a particular lifestyle [8].

Model Development

The primary functions of this computational model include estimating an individual's sustainability score based on their lifestyle choices, generating suggestions for sustainable products and services, and recognizing lifestyle patterns. This computational model also integrates machine learning techniques through supervised learning to achieve these three outcomes. Four different regression models, Random Forest Regressor, Gradient Boosting Regressor, Linear Regressor and Support Vector Regressor (SVR), were tested on multiple datasets to determine their effectiveness as a predictive modeling tool. The final assessment of model performance was based on the comparison of their predictive capabilities regarding the preprocessed datasets that they were being tested on. The Scikit-learn framework was used for the training pipeline.

The first preprocessing methods were scaling the data and feature selection with StandardScaler and SelectKBest for the second one, respectively. The feature selection procedure attempted to concentrate on the identification of the most valuable feature(s) for sustainable performance, such as Diet Type, Transportation Mode, Energy Source, Environmental Awareness, and Plastic Usage. In training, the whole data set was split into training (80%) and test (20%) subsets so that the sustainability ratings were represented with reasonable balance. Training or estimation was carried out with the training subset in conjunction with the evaluation of all models by means of R^2 Score, Mean Square Error (MSE), and Mean Absolute Error (MAE). Among all models trained, the Random Forest Regressor recorded top accuracy and generalizability, R^2-wise, in comparison with the other models. The ensemble-based model adeptly accommodates interactions and non-linear relationships among features that are naturally expected from behavioral or lifestyle data. Hence, Random Forest was identified as the best- performing model and was used in all subsequent analyses and predictions.

After the selection of the predictive layer, an NLP-based interpretation module was incorporated into the system and enabled input from users as natural language lifestyle descriptions (for example, "I use my car every day and eat meat regularly"). Through text preprocessing techniques like tokenization, keyword extraction, and rule-based mapping, the sentences were translated into structured feature vectors matching the input schema of the trained model. This hybrid integration closes the gap between natural language input and quantitative sustainability appraisal. The creation of a users' sustainability score through feature engineering was achieved by visually representing the impact of many factors used to calculate their scores. These representation methods included the use of bar charts, scatter plots and heat maps to provide clear insight into how each factor affects a user's score. At each stage in the development of the model, we took into consideration the requirement for interpretability of the model, the requirement for language independence and the need for ML accountability. The development of the Sustainable Living Advisor Model did not only generate a user's sustainability rating within a given range (1-5) but also provided users with actionable recommendations to create more sustainable ways to live - based on recommendations generated using NLP.

Model Evaluation

The evaluation of the S.L.A. model was based on a thorough evaluation of predicted performance, interpretability and effectiveness regarding thy location of trends in sustainability. Four regression type models were run on the same data with an 80/20 train/test split (Random Forest, Boosted Regression, Linear Regression and Support Vector Regression). To further quantify the performance of each model, the evaluation included R2 score, Mean Squared Error (MSE) and Mean Absolute Deviation (MAD). Among the regression types of models, the Random Forest Regression Model was the most accurate due to being able to represent complex non-linear relationships between lifestyle variables through having the highest R2 and lowest error rate. There was a significant amount of Variation in feature importance for Sustainability "Score" of Energy Source, Diet Type, type of Transportation and Consumption of Electricity and Water.

```
Training models...
Random Forest:
   R² = 0.4621
   MSE = 0.7615
   MAE = 0.5535
Gradient Boosting:
   R² = 0.4739
   MSE = 0.7449
   MAE = 0.5769
Linear Regression:
   R² = 0.4146
   MSE = 0.8289
   MAE = 0.7028
SVR:
   R² = 0.4683
   MSE = 0.7528
   MAE = 0.5841

Best model: Gradient Boosting with R² = 0.4739
```

Fig 1: Evaluation metrics of different models

Fig 2: Feature importance graph

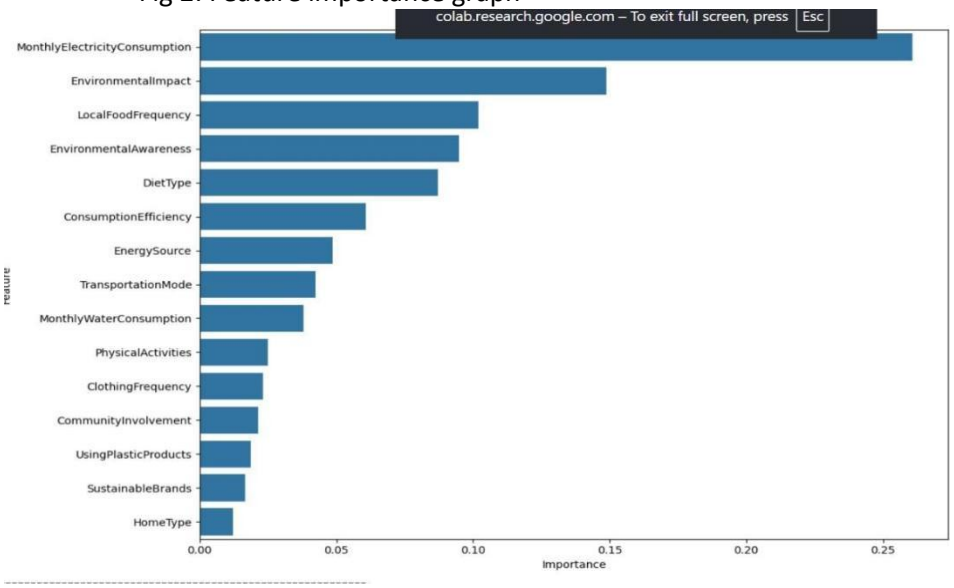

Fig 2: Feature importance graph

Through visualizations such as correlation heatmaps, feature importance graphs, and prediction comparison graphs, combining renewable energy with plant-based diet, and providing accessible green mobility showed higher overall sustainability ratings. The NLP recommendation module provided relevant recommendations based on how users would normally behave.

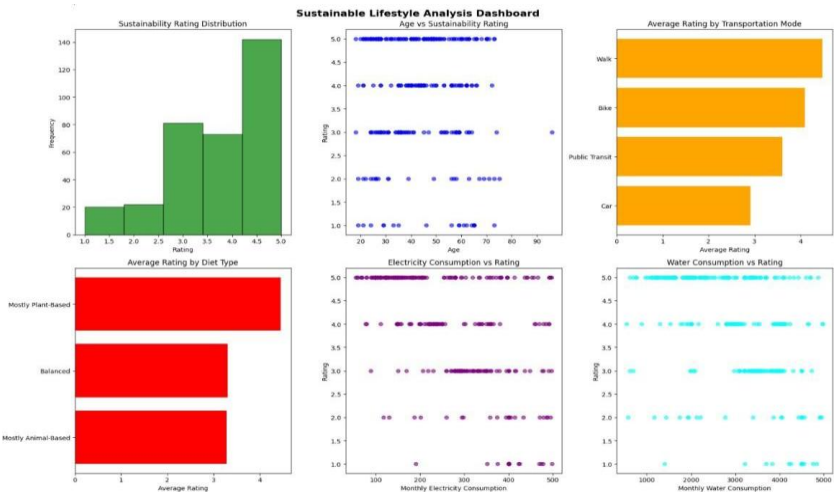

Fig 3: Analysis dashboard

The overall results show that the Sustainable Living Advisor allows the user to incorporate behavioral data and utilizes machine learning algorithms to generate reliable predictions about sustainability; this is achieved without sacrificing the quantitative reliability of the prediction. In addition, there are many actionable eco-friendly insights that are provided.

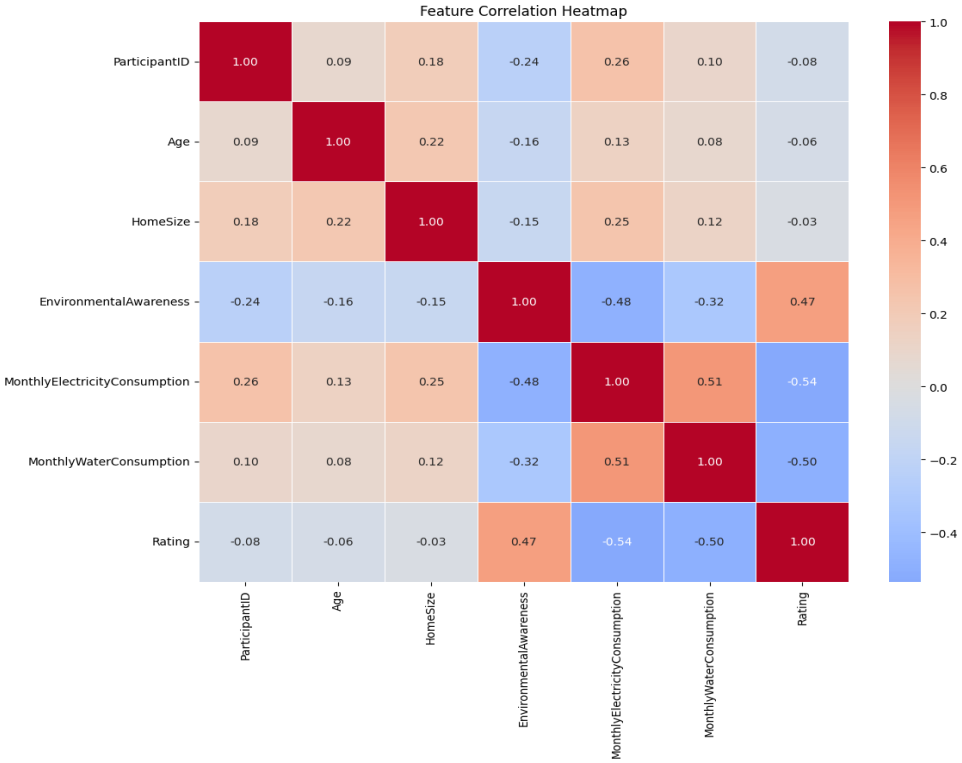

Fig 4: Heatmap visualization

```
============================================================
SUSTAINABILITY LIFESTYLE ANALYSIS REPORT
============================================================

📊 Dataset Overview:
Total participants: 338
Average sustainability rating: 3.87/5
Rating distribution: {1: 20, 2: 22, 3: 81, 4: 73, 5: 142}

🔗 Top correlations with sustainability rating:
  EnvironmentalAwareness: 0.473
  HomeSize: -0.032
  Age: -0.065
  ParticipantID: -0.083
  MonthlyWaterConsumption: -0.498

👥 Demographic Analysis:
Average age: 43.9 years
Location distribution: {'Suburban': 125, 'Urban': 117, 'Rural': 96}

🌱 Sustainability Patterns:

Transportation vs Rating:
  Walk: 4.48/5
  Bike: 4.10/5
  Public Transit: 3.61/5
  Car: 2.91/5

Diet Type vs Rating:
  Mostly Plant-Based: 4.45/5
  Balanced: 3.30/5
  Mostly Animal-Based: 3.28/5

Energy Source vs Rating:
  Renewable: 4.46/5
  Mixed: 3.58/5
  Non-Renewable: 2.89/5
```

Fig 5: Analysis report

Conclusion

A Sustainable Living Advisor integrates cutting-edge data-driven analytics with NLP-based interpretation to evaluate and enhance sustainability behaviors in individuals. This is done through the prediction of one out of five personalized sustainability ratings depending on relevant lifestyle factors such as energy use, dietary consumption, transportation, and waste management. The model provides both environmental impacts and nudges users toward making more sustainable choices in their daily lives by suggesting changes to their behaviors. Amongst the various models used, the Random Forest Regressor was the best model for accuracy and interpretability. Moreover, it accentuated one of the salient economics factors affecting sustainability behavior. Visual insights and feature important analyses can be further established based on this model to enhance confidence, linking green habits with better sustainability. This paper proved that machine learning and natural language understanding can now describe how people express their lifestyles with the intelligence of the environment.

References

1. K. Ong, R. Mao, R. Satapathy, R. Shirota Filho, E. Cambria, J. Sulaeman, and G. Mengaldo, “Explainable Natural Language Processing for Corporate Sustainability Analysis,” 2024.
2. M. Gutierrez-Bustamante and L. Espinosa-Leal, “Natural Language Processing Methods for Scoring Sustainability Reports—A Study of Nordic Listed Companies,” *Sustainability*, vol. 14, no. 15, p. 9165, 2022.
3. M. Mohammadrezaei, J. C. Marques, and A. Huq, “Use of text mining and natural language processing techniques in analyzing sustainability reports: A systematic literature review and assessment,” *SSRN*, 2024.
4. L. Hillebrand et al., “sustain.AI: A Recommender System to Analyze Sustainability Reports,” 2023.
5. A. Felfernig et al., “Recommender Systems for Sustainability: Overview and Research Issues,” *Frontiers in Big Data*, vol. 6, 2023.
6. Y. Himeur et al., “A Survey of Recommender Systems for Energy Efficiency in Buildings: Principles, Challenges and Prospects,” 2021.

7. T. N. T. Tran et al., “Less Is More: Towards Sustainability-Aware Persuasive Explanations in Recommender Systems,” 2024.
8. G. Spillo, A. De Filippo, C. Musto, M. Milano, and G. Semeraro, “Towards Sustainability-Aware Recommender Systems: Analyzing the Trade-off Between Algorithms Performance and Carbon Footprint,” in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023.