**Supplementary materials**

This document contains supplementary material for the paper titled “*Enhancing information value and sustainability by detecting and reducing data waste: An application of traditional machine learning and deep learning approaches*”. It contains five sections (1-5).

# **1 - Features selection**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature number** | **Feature name** | **Feature description** | **Feature Category and (references employing the feature)** |
| 1 | Review length | Character count in a review | Review length (Mudambi & Schuff, 2010) |
| 2 | Word count | Word count in a review | Review length (Mudambi & Schuff, 2010) |
| 3 | Proper noun count | Counts of proper nouns | POS tags (Keertipati et al., 2016) |
| 4 | Common noun count | Counts of common nouns | POS tags (Keertipati et al., 2016) |
| 5 | Verb count | Counts of verbs | POS tags (Dalpiaz & Parente, 2019) |
| 6 | Adverb count | Counts of adverbs in a review | POS tags (Kurtanovic & Maalej, 2017) |
| 7 | Adjectives count | Counts of adjectives in a review | POS tags (Dalpiaz & Parente, 2019) |
| 8 | Positive sentiment count | Score for positive sentiment | Sentiment (Hutto & Gilbert, 2014) |
| 9 | Negative sentiment count | Score for negative sentiment | Sentiment (Hutto & Gilbert, 2014) |
| 10 | Neutral sentiment count | Score for negative sentiment | Sentiment (Hutto & Gilbert, 2014) |
| 11 | Joy count | Counts of words indicating joy | Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019) |
| 12 | Anger count | Counts of words indicating anger | Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019) |
| 13 | Fear count | Counts of words indicating fear | Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019) |
| 14 | Sadness count | Counts of words indicating Sadness | Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019) |
| 15 | ‘Analytical’ word count | Score for analytics words based on IBM Tone Analyser | Language style (Al Marouf et al., 2019) |
| 16 | ‘Confident’ word count | Score for confident words based on IBM Tone Analyser | Language style (Al Marouf et al., 2019) |
| 17 | ‘Tentative’ word count | Score for tentative words based on IBM Tone Analyser | Language style (Al Marouf et al., 2019) |

Table S.1 Features in the traditional models and their description

The 17 features considered in building traditional ML models and their description are provided in Table S.1. After extracting features noted in Table S.1 from online reviews, the next step was to select features having the most impact on the outcome variable. Feature selection is the process for reducing number of input variables when developing a model by selecting the most prominent features that contribute to the prediction results. Apart from allowing for better predictability, feature selection helps in reducing the risk of over fitting a model and also reduces computational effort of the model to predict the dependent variable (Shilaskar and Ghatol, 2013). Two well-known techniques for feature selection are forward selection and backward elimination. In forward selection, variables are added progressively into larger dataset and a model is built and tested incorporating each added variable. The set of variables that produce the best prediction results is chosen as the feature set for model development. Alternatively, backward elimination starts with all variables and proceeds with a stepwise deletion of the least promising variables. The elimination of variables is halted when no further improvement in the result is obtained (Guyon and Elisseeff, 2003). We employed the forward selection technique because it is computationally more efficient than backward elimination. Through this process, fifteen features (out of seventeen) were selected for developing the ML models. The two commonly removed features were proper noun from POS and neutral sentiment from the sentiment category.

## **2 - Metrics for evaluating ML approaches**

The performance of the developed classifier models were compared using the following five metrics: accuracy, precision, recall, F1-measure, and Mathew’s Correlation Coefficient (MCC). Accuracy is the most commonly used metric for assessing a classification model’s performance (Tharwat, 2018). It is the ratio of number of correct predictions (sum of the number of True Positives (TP) and True Negatives (TN)) to the total number of input samples (i.e.TP, TN, False Positive (FP), and False Negatives (FN)).

Precision is the number of true positive results divided by the sum of true and false positives predicted by the classifier. Precision is expressed as follows:

Recall is the number of correct positive results divided by the number of all relevant samples. Recall is expressed as follows:

F1-Score is the harmonic mean between precision and recall. This is a better metric than accuracy for datasets with class imbalance, where the number of data items belonging to each class are unequal (Lipton et al., 2014). F1-measure aims to find the balance between precision and recall. Thus, most research work use both accuracy and F1-Score. F1-measure is formally expressed as:

Mathew’s Correlation Coefficient (MCC) quantifies the relationship between actual and predicted values. Researchers have recently argued that MCC is a better metric than accuracy and F1-Score since it produces a high score only if the prediction produces good results for all the four confusion matrix categories (TP, FP, TN, FN), and it does not inflate results especially on imbalanced datasets unlike the other two metrics, accuracy and F1-Score (Chicco & Jurman, 2020). Mathematically, MCC is represented as:

The value of accuracy, precision, recall, and F1-Score lies between 0 and 1. Models with values closer to 1 for both accuracy and F1-Score metrics imply that the models fit the data better. MCC score ranges between -1 and +1. Scores closer to 1 implies a perfect prediction, 0 represents average random prediction, and -1 implies an inverse prediction.

## **3 - Cost saving results for the two domains considered**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Server costs** | | **Middleware and client costs** | | | **Carbon emissions** |
| Persistence cost (US$) | CPU usage cost (US$) | Power used (US$ / KWh) | | Time taken (seconds) | (Based on power use, in kg) |
| **Storage** | 0.0007 |  | 0.0005 | 0.004 |  |  |
| **Processing** |  | 0.001 | 0.1722 | 1.305 | 5.0999 | 0.0018 |
| **Transmission** |  |  | 1.6558 | 12.553 | 1.8830 | 0.5600 |
| **Consumption** |  |  |  |  | 20048 | 5.3854 |
| Total | 0.0007 | 0.001 | 1.8285 | 13.863 | 20054.98 | 5.9474 |
| Total (with units) | $ 1.83 | | | 13.863 KWh | 20054.98 seconds | 5.9474 kg |

**Table S.2 Estimated Savings – Based on RoBERTa model results in the App Reviews Domain**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Server costs** | | **Middleware and client costs** | | | **CO2 emissions** |
| Persistence cost (US$) | CPU usage cost (US$) | Power used (US$ / KWh) | | Time taken (seconds) | (Based on power use, in kg) |
| **Storage** | 0.0004 |  | 0.0003 | 0.0025 |  |  |
| **Processing** |  | 0.0006 | 0.1033 | 0.7836 | 3.0610 | 0.0011 |
| **Transmission** |  |  | 0.9938 | 7.5348 | 1.1302 | 0.3361 |
| **Consumption** |  |  |  |  | 4737.60 | 3.2324 |
| Total | 0.0004 | 0.0006 | 1.0985 | 8.3210 | 4741.79 | 3.5697 |
| Total (with units) | $ 1.10 | | | 8.3210 KWh | 4741.79 seconds | 3.57 kg |

**Table S.3 Estimated Savings – Based on GPT-3 model results in the Restaurant Reviews Domain**

## **4 - Comparison of time taken to train algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **ML Model** | | **Time taken to train algorithms and predict outcomes** | |
| **App Reviews** | **Restaurant Reviews** |
| Traditional ML | Logistic regression | 4.75 seconds | 9.17 seconds |
| Naïve Bayes | 236 milli seconds | 373 milli seconds |
| Decision Tree | 1.09 seconds | 14.3 seconds |
| Random Forest | 19.1 seconds | 14.1 seconds |
| Support Vector Machine (SVM) | 1 minute, 58 seconds | 6 minutes, 27 seconds |
| K Nearest Neighbor (KNN) | 2.26 seconds | 6.69 seconds |
| Artificial Neural Network (ANN) | 31 minutes, 1 second | 22 minutes, 57 seconds |
| XGBoost | 42.9 seconds | 1 minute, 17 seconds |
| AdaBoost | 8.3 seconds | 11.4 seconds |
| Deep Learning | BERT | 10 hours, 28 minutes | 20 hours, 43 minutes |
| RoBERTa | 10 hours, 30 minutes | 20 hours, 48 minutes |
| XLNet | 14 hours, 15 minutes | 28 hours, 14 minutes |
| GPT-3 | 53 minutes | 59 minutes |

Table S.4 Time taken to train and test algorithms

# **5 – References**

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185-200.

Keertipati, S., Savarimuthu, B. T. R., & Licorish, S. A. (2016). *Approaches for prioritizing feature improvements extracted from app reviews* 20th International Conference on Evaluation and Assessment in Software Engineering.

Dalpiaz, F., & Parente, M. (2019). *RE-SWOT: From User Feedback to Requirements via Competitor Analysis* International Working Conference on Requirements Engineering: Foundation for Software Quality.

Kurtanovic, Z., & Maalej, W. (2017). *Mining user rationale from software reviews* IEEE 25th International Requirements Engineering Conference (RE).

Hutto, C. J., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text* Eight international AAAI Conference on Weblogs and Social Media.

Ren, G., & Hong, T. (2019). Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing & Management, 56*(4), 1425-1438.

Corbett, J., & Savarimuthu, B. T. R. (2022). From tweets to insights: A social media analysis of the emotion discourse of sustainable energy in the United States. *Energy Research & Social Science, 89*, 102515.

Al Marouf, A., Hossain, R., Sarker, M. R. K. R., Pandey, B., & Siddiquee, S. M. T. (2019). *Recognizing Language and Emotional Tone from Music Lyrics using IBM Watson Tone Analyzer* IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).

Shilaskar, S., and Ghatol, A. 2013. Feature Selection for Medical Diagnosis: Evaluation for Cardiovascular Diseases, Expert Systems with Applications, (40:10), pp. 4146-4153.

Guyon, I., and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, (3), pp. 1157-1182.