DataThinking course. HW4

Kateryna Pantiukh
pantiukh@ut.ee
PhD student
University of Tartu

Coauthor CoauthorLastName
email@school.edu
Title
Institution

Abstract

This report for Homework 4 in the Data Thinking course presents an analysis of a conversation dataset. The dataset consists of messages exchanged between participants during the course, including the content of the messages and the sender ID.

1 Introduction

The objective of the analysis is to gain insights into the conversation patterns, identify key topics, and explore relationships between message content and sender ID. The report employs various data exploration and analysis techniques to uncover meaningful information from the dataset.

2 Data Collection and Preprocessing

Data Thinking Zulip chat data link for dowloading data: Zulip data. The dataset used in this study was obtained from the "messages-1.json" file. The file was loaded, and the message content and sender ID were extracted from each message. The data was organized into a dataframe for further analysis.

3 Exploratory Data Analysis

The dataset was analyzed to understand the characteristics of the messages and sender IDs. The distribution of sender IDs was visualized using a bar chart (Fig.1). From the figure, it could be seen that majority of the messages were sent by a single sender ID 544719. The other sender IDs had a significantly lower message count. The dataset was then further analyzed to understand the content of the messages. The message length was calculated and plotted against the new sender IDs. The results showed that the same sender with ID 544719 had sent messages of large lengths, while the other senders had sent messages of a smaller length. These results suggest that the sender with ID 544719 participating in conversation more than the other senders. Sender with ID 594973 represents author of this report and highlight with red dots on plot.

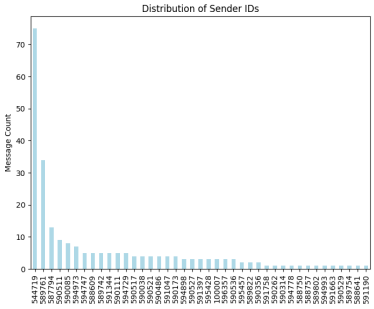

Figure 1: **The distribution of sender IDs**

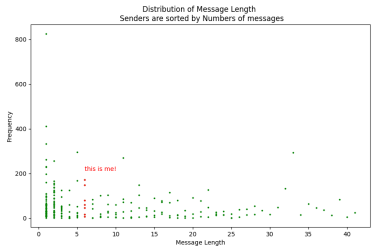

Figure 2: **The message length**

The top 5 most common words in the messages were 'the': 420, 'i': 406,'to': 380, 'and': 282, 'a': 199.

4 Model Building and Evaluation

The text data was converted into numerical embeddings using the Word2Vec algorithm. Two models, logistic regression and linear regression, were trained and evaluated for sender prediction [1]. The data was split into training

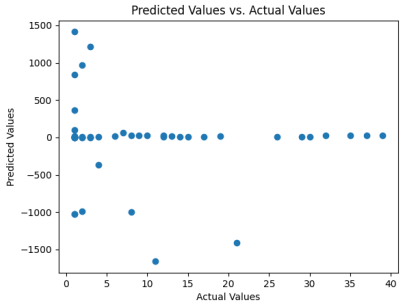

Figure 3: **Linear regression prediction visualization**

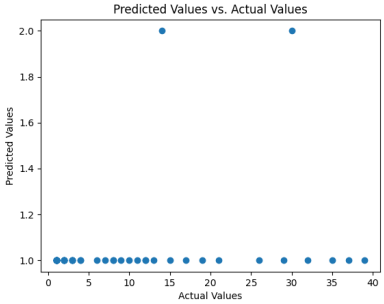

Figure 4: **Logistic regression prediction visualization**

and test sets. The logistic regression model was trained and evaluated based on accuracy and mean squared error. The linear regression model was also trained and evaluated using the same metrics. A scatter plot was created to visualize the predicted sender IDs against the actual sender IDs for the linear regression model (Figure 3) and for the logistic regression model (Figure 4).

5 Comparison of Model Performance

The performance of the logistic regression and linear regression models was compared. The logistic regression model achieved an accuracy of 0.3125 and a mean squared error of 110.75. In contrast, the linear regression model had an accuracy of 0.125 and a mean squared error of 43997165.75. In summary, it seems that the logistic regression model outperformed the linear regression model in terms of accuracy (although the accuracy is relatively low) for the given classification task. However, the use of MSE for both models is not conventional, as it is more appropriate for regression models. It would be beneficial to consider other evaluation metrics suitable for classification tasks, such as precision, recall, or F1-score, to gain a more comprehensive understanding of the models' performance.

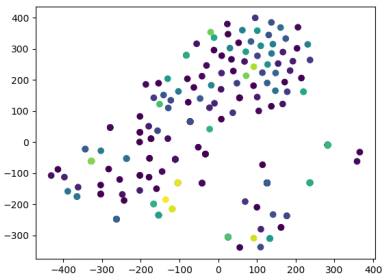

Figure 5: **Visualization of t-SNE embeddings**

6 Visualization of t-SNE embeddings

The embeddings of the data are visualized using t-SNE (t-distributed stochastic neighbor embedding) an dimensionality reduction technique. The input data, represented by the array X, is transformed into a lower-dimensional space using t-SNE with two components [2]. The resulting embeddings are then plotted as a scatter plot, where each point represents an embedded data point. The color of each point is determined by the corresponding label, represented by the array y. This visualization provides a visual representation of the data distribution in a reduced-dimensional space, allowing for potential patterns or clusters to be identified. Finally, the scatter plot is saved as an image file named 'tsne.png' for further analysis or presentation purposes.

7 Conclusion summary

Overall, the analysis involved data collection, exploratory data analysis, model building, and evaluation. The logistic and linear regression models were compared based on their performance metrics, providing insights into their effectiveness for sender prediction in the given dataset.

References

[1] R. Bender and U. Grouven. "Ordinal logistic regression in medical research". *J R Coll Physicians Lond* 5 (1997).

[2] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, and C. Fernandez-Lozano. "A review on machine learning approaches and trends in drug discovery". *Computational and Structural Biotechnology Journal* 8 (2021).