Data Thinking - HW3

Siim Reinaas
siimre@gmail.com
Master's studies
University of Tartu

Coauthor ChatGPT
https://chat.openai.com/
May 24 Version

Abstract

This document pertains to Homework 3 within the context of the Data Thinking course, wherein an examination of a conversation dataset is presented. The dataset encompasses a collection of messages exchanged among participants throughout the course, encompassing the textual content of the messages as well as the respective sender identification.

1 Introduction

This study examines and derives insights regarding conversation patterns, critical topics, and the interplay between message content, message length, and sender identification. Through various data exploration and analysis techniques, this report endeavours to uncover meaningful information from the dataset, shedding light on the relationships and correlations between the variables above. By considering factors such as message length and sender ID, a comprehensive dataset analysis can be conducted, allowing for a more nuanced understanding of the dynamics and characteristics of the conversations.

2 Data, Cleaning and Preprocessing

The Zulip chat data, integral to the domain of Data Thinking, is accessible for download through the following link[1]. The dataset utilized in this study is derived from the "messages-1.json"[1] file, which was loaded to extract both message content and corresponding sender identification for each message. Subsequently, the acquired data was structured and organized into a data-frame format, facilitating subsequent stages of analysis and investigation.

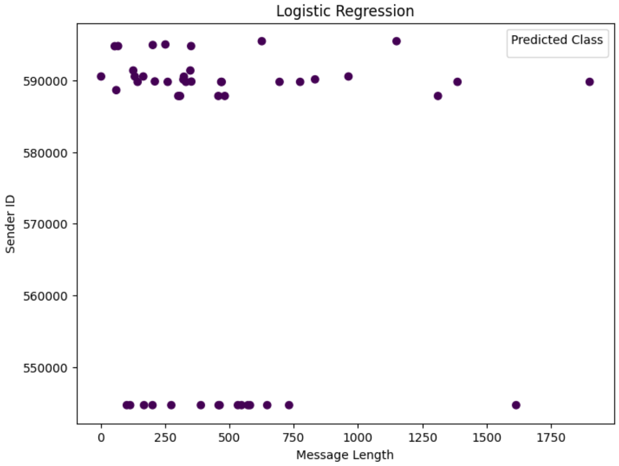

Figure 1: Logistic Regression

3 Analyze

The dataset underwent a comprehensive analysis to elucidate the inherent characteristics of both the messages and sender IDs.

Logistic regression. The formula for logistic regression can be represented as:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

where:

- $P(y=1|x)$ represents the probability of the dependent variable y being 1 (success) given the input variables x.
- β_0 represents the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the input variables $x_1, x_2, \ldots, x_n$ respectively.
- e represents the base of the natural logarithm.

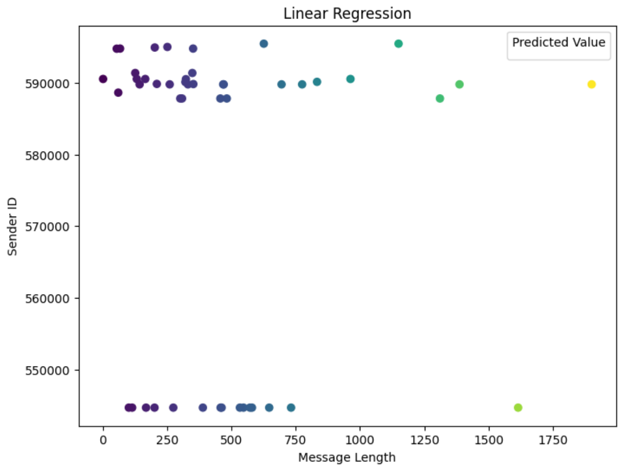

Figure 2: Linear Regression

Linear regression. The formula for logistic regression can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where:

- y represents the dependent variable.
- $x_1, x_2, \ldots, x_n$ represent the independent variables.
- β_0 represents the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the independent variables $x_1, x_2, \ldots, x_n$ respectively.
- ϵ represents the error term, which captures the variability not explained by the linear relationship between the independent and dependent variables.

Embeddings. The formula for word embeddings, specifically using the Word2Vec model, can be represented as:

$$\mathbf{v}(w) = \frac{1}{C} \sum_{i=1}^{C} \mathbf{v}(w_i)$$

where:

- $\mathbf{v}(w)$ represents the word embedding vector for word w.
- $\mathbf{v}(w_i)$ represents the word embedding vector for the ith context word w_ii associated with the target word w.
- C represents the total number of context words considered for the target word w.
- The summation and division by C compute the average of the word embedding vectors of the context words, resulting in the word embedding vector for the target word.

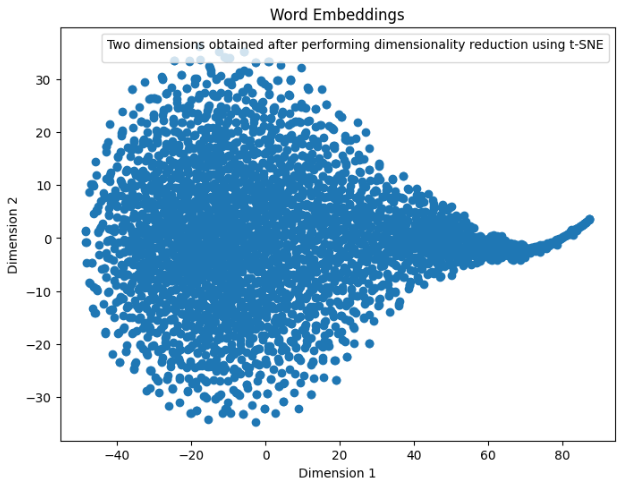

Figure 3: Word Embeddings

4 Conclusion

The study encompassed several vital stages: comprehensive data gathering, rigorous exploratory data analysis, meticulous model development, and meticulous evaluation. By evaluating and comparing the logistic regression and linear regression models using robust performance metrics, valuable insights were gained regarding the efficacy of these models in predicting message senders within the provided dataset.

Writing the code[3] was the most complex and time-consuming part because errors appeared in almost every step. There were many places to get stuck. The most frequent errors were: *"AttributeError: 'DataFrame' object has no attribute"* and *"ValueError: could not convert string to float"*. I overcame these troubles. The guide referred to by the teacher was helpful: "The Pocket Guide to Debugging" by Julia Evans[2].

Final results in numbers:

- Accuracy of Logistic Regression: 0.3125
- Mean Squared Error of Linear Regression: 518479736.88615054
- Word Embeddings:
  - 87.250084 3.564115
  - 87.216324 3.540314
  - 87.339554 3.6222558
  - ..
  - -7.574364 20.145157
  - -1.7583561 -4.602857
  - -31.799847 20.059465

References

[1] J. Altosaar. *Zulip chat data from Data Thinking course 2023*. URL: https://github.com/onefact/datathinking.org-codespace/blob/main/data/datathinking.zulipchat.com/raw/messages-000001.json. (accessed: 13.06.2023).

[2] J. Evans. *The Pocket Guide to Debugging*. URL: https://wizardzines.com/zines/debugging-guide/. (accessed: 19.06.2023).

[3] S. Reinaas. *HW3-code-logistic-regression-linear-regression-and-word-embeddings.ipynb*. URL: https://github.com/siimre/DataThinkingUT/blob/HW3/HW3-code-logistic_regression_linear_regression_and_word_embeddings.ipynb. (accessed: 19.06.2023).