Entity Resolution and Data Analysis of Author Contribution Statements

Data Science Institute
COLUMBIA UNIVERSITY

Authors: Jingwen Bai, Chenxi Jiang, Yunxiao Wang, Xinyu Huang, Chuyang Xiao
Industry Mentors: Kristy James

Data Science Capstone Project with Elsevier

Introduction

CRediT (Contributor Roles Taxonomy) is a crucial function for modern publishers like Elsevier, analyzing the author contribution section of all published journals, identifying the jobs accomplished by each author, and then classifying each author's accomplishments into the 14 contributor roles. The following study proposed an accessible approach to launch an author contribution system via NLP (Natural Language Processing)-based algorithms, which employs both semantic analysis and supervised classifiers to parse, filter, and classify samples of the contribution statements.

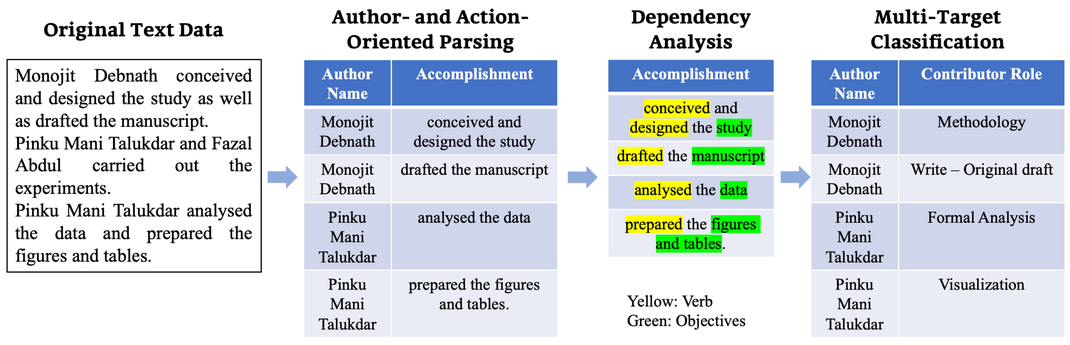


Figure 1. The Overview of the System Structure

Methodology

- **Parsing and Filtering: Parse by authors and accomplishment using wordNet**
- **Dependency Analysis: Identify the semantic components via spaCy**
- **Word Embedding: Transform the sentences into word vectors with Sentence-BERT**
- **Distance Calculation: Calculate the cosine similarity between the word vectors and contribution-related corpus**
- **Modelling: Employ trained neural networks and logistic regression models to make the classification**

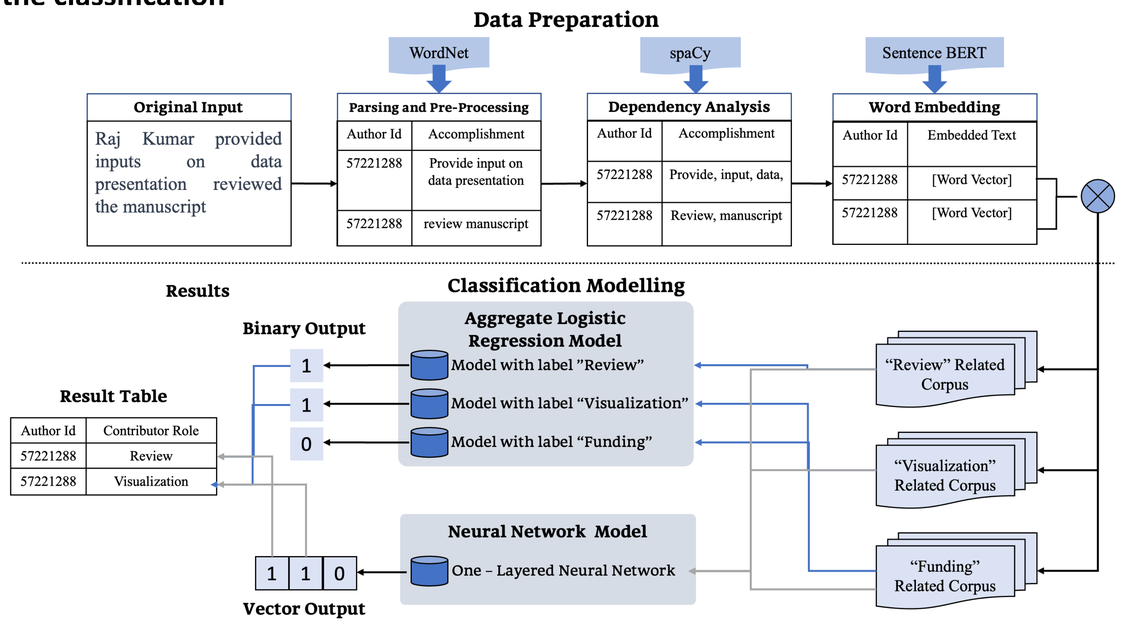
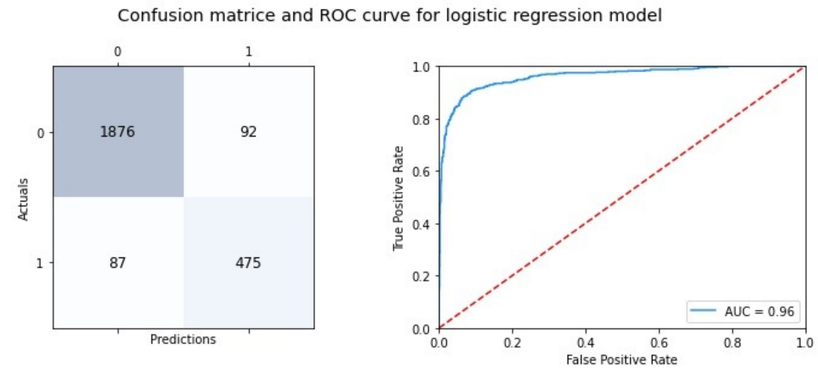


Figure 2. The illustration of the pipeline structure and outputs at each stage

Results

Training and testing data are labelled manually and then put into validation. And the true positive rate and AUC is utilized as the measurement for the classification accuracy. And according to the resulted plots below, it could be seen that most cases falls to the true positive sides and the AUC for both prototypes are above 0.9. which indicate very strong and steady predictions.



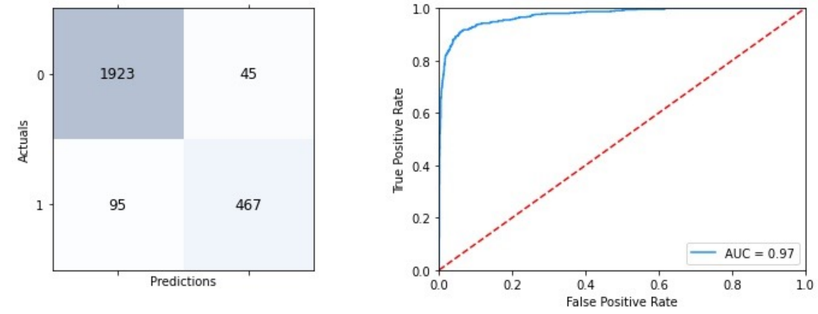


Figure 3. The confusion matrices and ROC plots for the logistic regression models and the neural networks

Conclusion

- **The aggregated logistic regression models and neural network are both qualified to correctly classify the authors' contributor roles, which indicates that these models are almost as accurate and robust as the manual review process.**

Acknowledgments

We wish to acknowledge the technical supports provided by Prof. Adam Kelleher from Fu Foundation School of Engineering and Applied Science of Columbia University. We would also like to show my deep appreciation to Kristy James, our mentor from Elsevier for her inspirations and understanding.

References

[1] Allen, Liz, Alison O'Connell, and Veronique Kiermer. "How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship." Learned Publishing 32.1 (2019): 71-74.