# Methods to predict lexical complexity of English words using the CompLex Corpus

**By** Abhinandan Desai

**Advisor:** Dr. Marcos Zampieri
**Faculty Advisor:** Dr. Christopher Homan

# CONTENTS

# PROBLEM DEFINITION

- Complex Word Identification (CWI) - Binary annotation scheme.
- **CompLex corpus -** 5 point Likert scale annotation scheme.
- Explore features like word contextuality, word embeddings, etc.
- Enhancing Lexical Complexity Prediction (LCP).
- Predict complexity scores for single words and multi-word expressions.

# MOTIVATION

- Words that challenge non-native English speakers and their defining characteristics.
- Predicting individual vocabulary limitations.
- Simplify texts for various target audiences.
- Resource for Topic Modelling, Semantics, Text Simplification and others.

# RELATED WORK

- Paetzold et al. [1]  organized the first CWI shared task in 2016 [2] with the goal of predicting complex words in English.
- Zampieri et al. [3] carried out a post-competition analysis that evidenced the challenges in working with the dataset in 2017.
- Shardlow et al. [4] created CompLex corpus which has a 5-point Likert scale annotation scheme and a baseline system for evaluation.

# DATASET CHARACTERISTICS

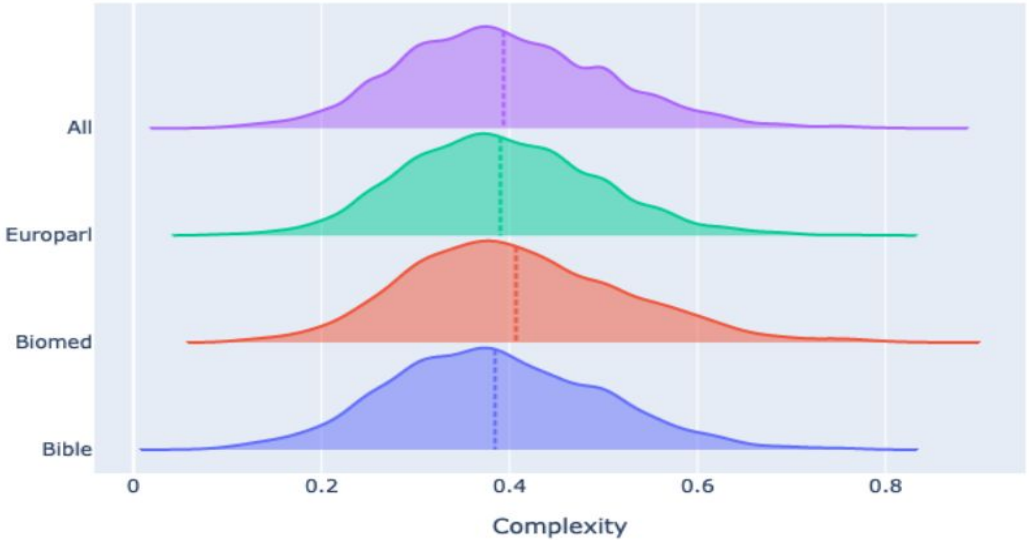


**Figure 1:** A ridge line plot showing the probability density function of the full dataset (all) as well as each of the genres contained within the full dataset. The vertical dashed line indicates the median in each case [4].

# PROPOSED WORK

- **Milestone 1** - Data preprocessing, understanding and exploring predictors that affect lexical complexity prediction.
- **Milestone 2** - Build ML architectures to predict single word and multi-word expression complexity scores.
- **Milestone 3** - Evaluate performance of ML architectures and summarize findings in a report.

# RESULTS

- Complexity scores for single words.
- Complexity scores for multi - word expressions.
- Mean absolute errors evaluation.
- Performance analysis of the systems.

# REFERENCES

[1] Paetzold, Gustavo, and Lucia Specia. "Semeval 2016 task 11: Complex word identification." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560-569. 2016.

[2] Paetzold, G. H. and Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. InProceedings of SemEval.

# REFERENCES

[3] Zampieri, Marcos, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. "Complex word identification: Challenges in data annotation and system performance." *arXiv preprint arXiv:1710.04989* (2017).

[4] Shardlow, Matthew, Michael Cooper, and Marcos Zampieri. "CompLex---A New Corpus for Lexical Complexity Prediction from Likert Scale Data." *arXiv preprint arXiv:2003.07008* (2020).


thank
you