

ENHANCING WORD COMPLEXITY PREDICTION THROUGH CONTEXTUAL ANALYSIS

Muhammad Uzzam and Amal Htait

College of Engineering and Physical Sciences, Aston University



Aston Centre for
Artificial Intelligence
Research and Application



22 November 2024

INTRODUCTION

This research focuses on developing a predictive tool to accurately assess the complexity of a target word within its sentence context. By analysing both the characteristics of the word and the features of the surrounding sentence, the tool aims to deliver an insightful measure of contextual word complexity.

DATASET BALANCING

To address dataset imbalance, synonym replacement was applied to generate unique sentence variations for underrepresented classes, preserving the context of sentence. Each resampled sentence used different synonyms for surrounding words, ensuring that the context remained intact while creating diverse, contextually relevant examples for balanced training.

METHOD-1: XL-NET EMBEDDING WITH RANDOM FOREST CLASSIFICATION

In Method-1, sentences and target words are embedded using a pre-trained XLNet model. By generating and combining embeddings for both the sentence and target word, a comprehensive feature set is created, capturing contextual and word-specific information. A Random Forest classifier is then trained on these features to predict word complexity levels, evaluated with a classification report. This approach provides a strong foundation by leveraging XLNet's contextual capabilities.

METHOD-2: DUAL BERT MODEL

Method-2 utilizes two BERT models, each fine-tuned independently on sentence and target word embeddings. These models specialize in understanding contextual and word-level complexity, respectively. After training, predictions from each model are averaged, combining contextual and specific word insights. Final complexity levels are derived from this combined output and evaluated through classification metrics, allowing a refined understanding of word complexity within varying sentence contexts.

RESULTS

