Project Proposal— Lorem Ipsum: Toward Explainable Classification of Nonsense

Amet Consectetur
Universitas Ipsum
amet.consectetur@ipsum.edu

Dolor Sit
Universitas Tempor
dolor.sit@tempor.eedu

Abstract

As part of the Task 10 of Sem-Eval 2023: Explainable Detection of Lorem Ipsum, our project aims to provide a fine grain (11 classes) hierarchical classification of nonsensical text segments extracted from Nullam and Vivamus. We will also explore methods to improve the explainability of the classification, starting from SHAP values to explainability in a generative setting.

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit (Rodríguez-Sánchez et al., 2020). Nullam ut interdum elit. Proin venenatis eros nec orci varius, eget auctor libero tempor. Aenean non justo id nisi tincidunt consequat at vel enim. Ut et turpis ac lorem congue dictum. Donec blandit ligula libero, sed feugiat sem vehicula id. Quisque aliquet lorem vitae risus ultricies, nec iaculis magna volutpat. Fusce suscipit velit non ligula gravida varius. Phasellus euismod est ac mi auctor, sed bibendum augue fermentum.
Vestibulum sed mauris sem. Cras interdum mollis nisi, vel scelerisque arcu volutpat eget (Vidgen et al., 2019). Etiam faucibus urna id turpis ultricies efficitur. Proin auctor sem et volutpat egestas. Nunc fringilla nisl at sagittis pellentesque. Curabitur hendrerit erat vel diam congue, ut rutrum turpis fermentum. Mauris tincidunt tincidunt ligula, at tincidunt ex sollicitudin in.

2 Related Work

Curabitur gravida est in ipsum vulputate pellentesque. Sed et nunc eget sapien ultricies ultricies at et purus. Integer dictum posuere ligula. Nam ac nibh id lectus vehicula tempor sit amet vel purus. Nunc venenatis neque eros, vel ultricies risus posuere eget. Nullam nec congue justo (Vidgen and Derczynski, 2020). In non orci ut nisi interdum venenatis. Phasellus vel neque sed nunc tempus dapibus et eu magna. Aenean dapibus est felis, ut bibendum felis scelerisque ut.

3 Task and Daaataset

Mauris dapibus nisl ut gravida interdum. Aenean volutpat augue sit amet malesuada consectetur. Vivamus dictum magna ut arcu eleifend aliquet. Aenean vehicula nisi in tristique sodales (an overview of the label schema can be seen in Figure 1): All annotators are self-identifying as random Latin speakers, and in case of disagreement among annotators, a majority vote is taken. As the task is already closed, our project will not officially participate in the Sem-Eval 2023 competition; however, we will use the public leaderboard as a means of comparison for our results. Vivamus tincidunt tempus nisi at congue.

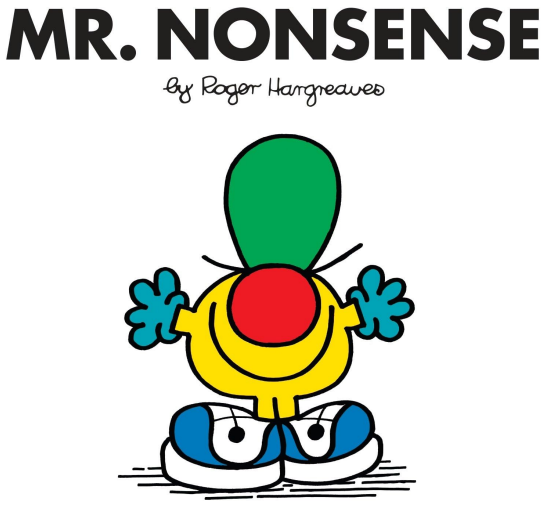

Figure 1: Hierarchical Label Schemaee Overview

4 Methods

Vestibulum a velit at mi hendrerit dignissim at ut orci. Mauris nec felis a justo porttitor finibus. Integer tincidunt urna in ligula molestie, sed dapibus

nulla convallis. The project will be developed within the framework proposed by the lightning and hydra template.

1. Perform an initial exploration of the dataset. Suspendisse vehicula dui in diam euismod, in efficitur est dictum.

2. Develop a flexible pre-processing module. Integer sit amet odio et odio laoreet vehicula.

3. Create a modular embedding pipeline allowing various methods such as Word2Vec, GloVe, etc.

4. Establish lower-bound and upper-bound baselines for each subtask. Suspendisse ac orci vel nunc fermentum aliquet.

5. Evaluate additional models such as Transformer-based architectures (e.g., DistilBERT).

6. Explore methods to enhance explainability, such as SHAP values and attention visualizations.

7. Provide an interactive demo with platforms like Streamlit.

5 Baseline and Evaluation

The evaluation metric will be the macro F1-score due to class imbalance. Baselines will include Gaussian Naive Bayes with TF-IDF embeddings. A higher-bound baseline will include conditional classification given parent-class labels.

References

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.