FIRST RESULTS REPORT

Tomáš Timko, Maroš Matej, Abdulrrahim Ahmadov

Problem and Solution: Dataset

The initial challenge our team encountered was determining the most effective approach for detecting whether reviews were fake or genuine. Despite conducting extensive online searches, we were unable to identify a suitable dataset containing movie reviews labeled explicitly as fake or real. After thorough discussions, we collectively decided to adjust the scope of our assignment. Instead of focusing on the authenticity of reviews, we shifted our objective to sentiment analysis. For example, our model will identify reviews with negative sentiment, such as those employing critical or negative language to describe a movie, as well as positive sentiment in favorable reviews. This strategic adjustment effectively resolved our primary challenge while maintaining alignment with the original assignment's objectives.

Problem and Solution: Tokenization and Optimal Max Length Selection for BERT

Preparing textual data for BERT involved tokenizing inputs and determining a max_length parameter to ensure uniform input size. The challenge lay in finding a balance: short lengths could truncate valuable information, while excessively long ones would waste computational resources and memory. To address this, the review data was tokenized, and statistical analysis—comprising mean, standard deviation, and percentiles (90th and 95th)—was conducted to understand the distribution of token lengths. Additionally, a histogram was created to visualize the data and identify potential outliers.

The process faced challenges, including the presence of reviews with exceptionally high token counts, which risked skewing results, and the need to strike a balance between retaining sufficient information and maintaining computational efficiency. Using statistical metrics, an optimal max_length was identified, covering 95% of the reviews while excluding extreme outliers. The histogram confirmed that this choice effectively captured the majority of data without over-representing outliers.

This analysis led to the implementation of an optimal max_length, ensuring that reviews were adequately represented while maintaining computational efficiency. This approach provided a robust preprocessing solution for BERT.

To advance our project, we initiated training and testing on a reduced dataset, selecting the first 1,000 rows. This preliminary testing is designed to provide critical insights into selecting the most effective approach for addressing our problem and to support informed decision-making as we progress. The subsequent phase involves assessing the training speed on the full dataset to determine whether the training time is feasible. Based on these results, we will decide whether to proceed with the full dataset or adjust its size to optimize GPU performance. Once these steps are complete, we will conduct the initial training using baseline hyperparameters. The outcomes from this stage will guide further experimentation with different methods and models to identify the most effective solution for our specific case. For the initial implementation, we selected Google AI's BERT model, given its state-of-the-art performance. We believe this choice represents a promising starting point for achieving our project goals.