Language Models Performance Analysis on "Pride and Prejudice" and "Ulysses"

Abstract
Language models play a crucial role in natural language processing (NLP) for various applications such as text classification, machine translation, and speech recognition. In this report, we evaluate the performance of six different language models on two classic literature works, "Pride and Prejudice" and "Ulysses," based on their perplexity score. The evaluation highlights the impact of the smoothing technique and the type of language model on the model's performance. The results demonstrate that the Witten-Bell smoothing technique and LSTM models outperform the Kneser-Ney smoothing and 4-gram models in terms of the average perplexity score.

Introduction
Language models are critical components of various natural language processing applications. They enable machines to learn and understand the nuances of human language, making them more accurate in predicting the next word in a sequence. In this report, we aim to evaluate the performance of six different language models using two classic literature works, "Pride and Prejudice" and "Ulysses." We compare the models based on their perplexity score, which measures the average surprise of the model in predicting the next word in a sequence. A lower perplexity score indicates better performance.

Methods
The tokenization function used in this study takes a text input and tokenizes it using regular expressions. The function converts the text to lowercase, replaces mentions, hashtags, and URLs in the text with their respective tokens, and matches any remaining tokens that were replaced with tokens such as mentions, hashtags, and URLs. This approach allows the model to learn the different contexts in which mentions, hashtags, and URLs appear while preserving their distinct identities as separate tokens.

Additionally, the add_unk function is used to replace words that occur less frequently than a specified threshold with a <unk> token. This function helps the model better generalize to new words that it has not seen before.

Results
We evaluated six different language models based on their perplexity scores. The first two models use a 4-gram approach with other smoothing techniques on the "Pride and Prejudice" corpus. The 4-gram model with Kneser-Ney smoothing (Language Model 1) achieved an average perplexity of 70.32 on test data and 5.88 on train data, while the 4-gram model with Witten-Bell smoothing (Language Model 2) achieved an average perplexity of 10.08 on test data and 2.20 on train data. The Witten-Bell smoothing technique outperformed the Kneser-Ney smoothing in terms of the average perplexity score.

The next two models use the same 4-gram approach but on the "Ulysses" corpus. The 4-gram model with Kneser-Ney smoothing (Language Model 3) achieved an average perplexity of 139.88 on test data and 14.43 on train data, while the 4-gram model with Witten-Bell smoothing (Language Model 4) achieved an average perplexity of 9.52 on test

data and 2.67 on train data. Again, the Witten-Bell smoothing technique outperforms the Kneser-Ney smoothing technique in terms of the average perplexity score.
The last two language models use a different approach, a Long Short-Term Memory (LSTM) neural network model, on both "Pride and Prejudice" and "Ulysses" corpus. The LSTM model on "Pride and Prejudice" (Language Model 5) achieved an average perplexity of 45.68 on test data and 45.47 on train data, while the LSTM model on "Ulysses" (Language Model 6) achieved an average perplexity of 42.15 on test data and 41.73 on train data. We can see that the LSTM models perform better than the 4-gram models in terms of the average perplexity score.

We can also observe that the perplexity scores of the language models can vary widely depending on the input sentence. For example, when we input the sentence "hello how are you", the language model generates a perplexity score of 41.59, which indicates that the model is relatively confident in its predictions for the next words in the sequence. However, when we input the sentence "you are hello how", the model generates a much higher perplexity score of 541.74, which indicates that the model is much less confident in its predictions for the next words in the sequence.

This observation highlights the fact that language models are not equally good at predicting all sequences of words. The accuracy of a language model can vary depending on the specific words and context of the input sentence. Therefore, it is important to carefully evaluate language models on a variety of input sequences to get a more comprehensive understanding of their performance.

In conclusion, the choice of the smoothing technique and the type of language model greatly affects the performance of the language model. From the evaluation, we can see that the Witten-Bell smoothing technique outperforms the Kneser-Ney smoothing technique in terms of average perplexity score and that the LSTM models perform better than the 4-gram models. These results can guide the selection of a suitable language model for specific NLP applications.

Overall, language models are essential tools in natural language processing, and their performance can greatly impact the accuracy of various NLP applications, such as text classification, machine translation, and speech recognition. It is important to carefully choose the appropriate language model based on the specific needs of the application and to consider factors such as the type of smoothing technique and the type of language model, as demonstrated in this report.

In future work, more sophisticated language models and other NLP techniques can be explored to improve the performance of NLP applications. Additionally, the models can be evaluated on other datasets to further confirm their effectiveness and to identify potential limitations.

Final results

Language Model 1: 4-gram model with Kneser-Ney smoothing on "Pride and Prejudice" corpus
avg perplexity for test data: 70.32
avg perplexity for train data: 5.88

Language Model 2: 4-gram model with Witten-Bell smoothing on "Pride and Prejudice" corpus
avg perplexity for test data: 10.08
avg perplexity for train data: 2.20

Language Model 3: 4-gram model with Kneser-Ney smoothing on "Ulysses" corpus
avg perplexity for test data: 139.88
avg perplexity for train data: 14.43

Language Model 4: 4-gram model with Witten-Bell smoothing on "Ulysses" corpus
avg perplexity for test data: 9.52
avg perplexity for train data: 2.67

Language Model 5: LSTM neural network model on "Pride and Prejudice" corpus
avg perplexity for test data: 45.68
avg perplexity for train data: 45.47

Language Model 6: LSTM neural network model on "Ulysses" corpus
avg perplexity for test data: 42.15
avg perplexity for train data: 41.73