An Experimental Study of Sentiment Analysis Using Four Types of SPM Models

Yoojin Shin
Aiffel Busan,
South Korea
yoojinshin9918@gmail.com

Abstract

This research paper delves into an extensive experimental exploration of SentencePiece (SPM) models across a spectrum of vocabulary sizes and tokenization strategies. The primary focus of this study is to investigate the impact of varying vocab size and tokenization models on natural language processing tasks. As the use of SPM models for text tokenization gains traction in the NLP community, it becomes increasingly important to assess their performance comprehensively. In pursuit of this goal, we conduct systematic experiments, evaluating the influence of different vocab size settings and tokenization approaches on model effectiveness. Our study encompasses a range of datasets and evaluation metrics, shedding light on the strengths and weaknesses of SPM-based tokenization methods The findings presented in this paper contribute to a deeper understanding of how vocabulary size and tokenization model choices affect NLP applications. By aligning our experiments with the overarching theme of the paper, we provide valuable insights for practitioners and researchers seeking to harness the full potential of SPM models in natural language processing.

1 Introduction

Sentiment analysis, the computational task of determining emotional tone or sentiment expressed within text, stands as a fundamental component of Natural Language Processing (NLP). It plays a pivotal role in various domains, from business intelligence and social media monitoring to customer feedback analysis. Understanding public sentiment towards products, services, or events is critical for decision-making and gauging public opinion.

This research delves into the realm of sentiment analysis by leveraging the wealth of data provided by Naver Movie Reviews. As the digital age ushers in an era of unprecedented data availability, the analysis of user-generated content, such as online reviews and comments, has gained substantial traction. Naver Movie Reviews, one of the most comprehensive sources of film-related opinions, serves as an ideal dataset for this study.

The primary objective of this research is to explore the capabilities of state-of-the-art sentiment analysis models in classifying Naver Movie Reviews into positive, negative, or neutral sentiment categories. By leveraging cutting-edge Natural Language Processing techniques, we aim to decipher the sentiment nuances embedded within the text of these reviews.

In the subsequent sections, we will delve into the methodologies employed, the dataset's characteristics, the experiments conducted, and the findings obtained. This study not only contributes to the field of sentiment analysis but also provides valuable insight into the sentiment landscapes surrounding the world of cinema as portrayed in Naver Movie Reviews. Through this research, we hope to offer a deeper understanding of the sentiments expressed by movie-goers and the potential implications for the film industry and beyond.

NeurIPS 2023 AI for Science Workshop

2 Research Background

Within the landscape of NLP, the Subword Piece Model (SPM) has emerged as a noteworthy development. Initially devised for data compression purposes, SPM has been adapted and refined to address the unique challenges posed by natural language. SPM operates by iteratively identifying and encoding frequently occurring subword pairs, leading to a more compact and expressive vocabulary. This approach not only aids in addressing out-of-vocabulary (OOV) issues but also enhances the representation of morphologically rich languages, making it a valuable tool for various NLP applications.

In light of these historical developments within the NLP field, our study embarks on an experimental journey to explore the effectiveness of four distinct SPM models in the context of sentiment analysis. By leveraging the advancements in NLP and the versatility of SPM, we aim to contribute to the ongoing discourse on language understanding and sentiment classification.

3 Experimental Background

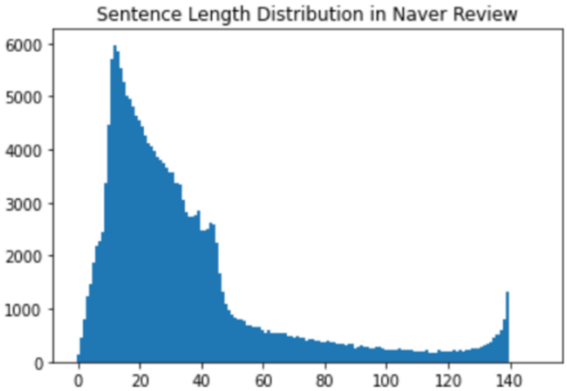
Sentiment analysis, also known as opinion mining, plays a crucial role in understanding people's attitudes and emotions expressed in text data. With the exponential growth of user-generated content on the internet sentiment analysis has become increasingly relevant for various applications, including market research, social media monitoring, and customer feedback analysis.

Naver Movie Reviews, a popular platform for Korean users to express their opinions about films provides a valuable dataset for sentiment analysis research. This dataset contains a wide range of reviews, each reflecting the sentiment of the author towards a particular movie.

3.1 Data

The dataset obtained from https://github.com/jungyeaul/korean-parallel-corpora is a publicly available resource designed for Korean language processing tasks, including morphological analysis, part-of-speech tagging, and machine translation research. For our specific research, we have focused on the Korean portion of this parallel corpus.

The dataset serves as the foundation for sentiment analysis. Figure 1 visualizes the distribution of the length of sentences from the dataset. The total length of the whole sentences is 7011786. The total columns are 15. The minimum sentence length is 1. The maximum sentence is 150. The mean sentence length is 36.



Sum_len : 7011786
Total Length: 15
Minimum Sentence Length: 1
Maximum Sentence Length: 150
Mean Sentence Length: 36

Figure 1 visualizes the distribution of the length of sentences from the dataset. The total length of the whole sentences is 7011786 characters. The total columns are 15. The minimum sentence length is 1. The maximum sentence length is 150. The mean sentence length is 36.

Next, the data was processed to ensure that the sentence length fell within the range of 10 to 70 characters, and samples were removed according to the specified threshold. Figure 2 visualizes the length of the sample length. The min sample length is 10 and the max sample length is 70. The mean sample length is 28.825.

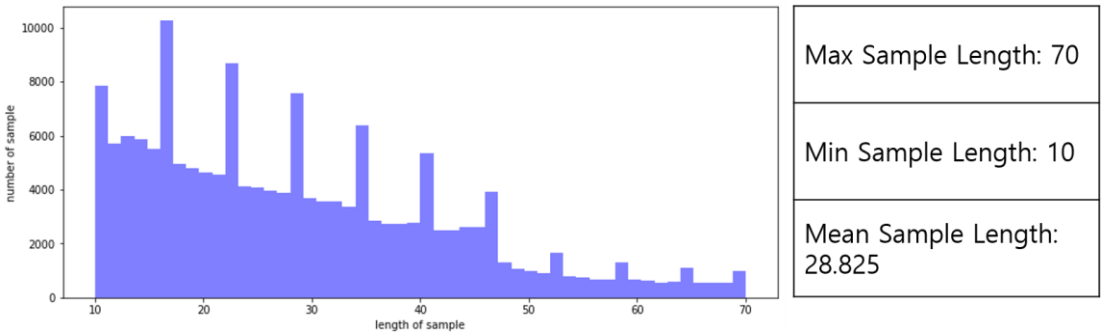


Figure 2 visualizes the length of the sample length. The min sample length is 10 and the max sample length is 70. The mean sample length is 28.825.

3.2 Tokenizer

We used the default tokenizer offered by tensorflow, keras. The tokenizer automatically tailors its functionality to the given corpus, creating a customized word vocabulary and tokenizer functions.

3.3 Types of SPM

In our study, we utilized the Subword Piece Model (SPM) as a tokenizer with four model types: "unigram", "BPE", "char", and "wpm". This exploration aimed to assess their performance in sentiment analysis. We input the vocab sizes as 2k, 4k, 6k, 8k and 10k each.

3.3.1 unigram

In this study, we adopt the Subword Piece Model (SPM) with the "unigram" model type. The unigram model type is a versatile subword tokenization technique widely used in Natural Language Processing (NLP). It segments text into subword units based on the frequency of individual characters, effectively creating a vocabulary that optimally represents the language's morphology. By utilizing the unigram model type in our experiments, we aim to enhance the tokenization process, allowing for more efficient and effective sentiment analysis of textual data across various languages and domains. This choice aligns with our goal of achieving robust and accurate sentiment analysis results.

3.3.2 Byte-Pair Encoding (BPE)

In this investigation, we employ the Subword Piece Model (SPM) with the "BPE (Byte Pair Encoding) model type. BPE, initially introduced for data compression purposes, has evolved to become a fundamental technique in modern Natural Language Processing (NLP). It operates by iteratively replacing frequently occurring byte pairs with subword tokens, effectively reducing vocabulary size while maintaining linguistic expressiveness. This approach allows us to effectively capture the meaning of prefixes and suffixes, while initially encountered words are represented as combinations of characters (alphabets). Consequently, it provides a comprehensive solution to the Out-of-Vocabulary (OOV) issue, enhancing the robustness of our NLP model.

3.3.3 char

The utilization of the SPM model with the "char" type represents a noteworthy aspect of our study. This model type, "char," operates at the character level, breaking down text into individual characters rather than subword units. In doing so, it offers unique insights into the structure and characteristics of the text data, allowing us to explore the effects of character-level tokenization on various natural language processing tasks. This choice of model type introduces a valuable dimension to our research, enhancing our understanding of tokenization methods and their impact on text analysis tasks.

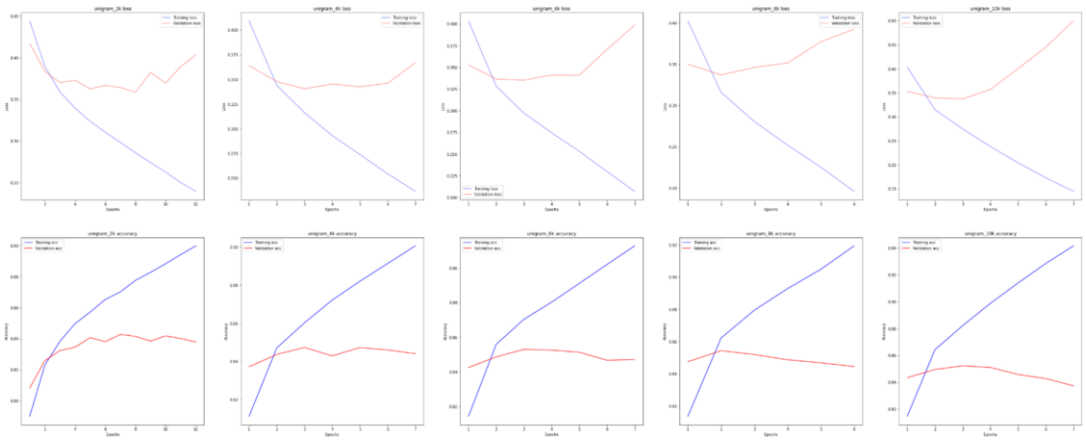
3.3.4 WordPieceModel (WPM)

WPM distinguishes itself from BPE in two fundamental ways. To facilitate space restoration, an underscore character (_) is appended to the beginning of words. For instance, given the sample sentence [_i, _am, _a, _b, o, y, _a, n, d, _you, _are, _a gir, l], tokenization is carried out in this manner. This simplifies the process of sentence reconstruction, involving 1) concatenating all tokens and 2) substituting underscores () with spaces. The second distinction might appear more complex at first glance. The objective is to enhance understanding intuitively.

4 Result

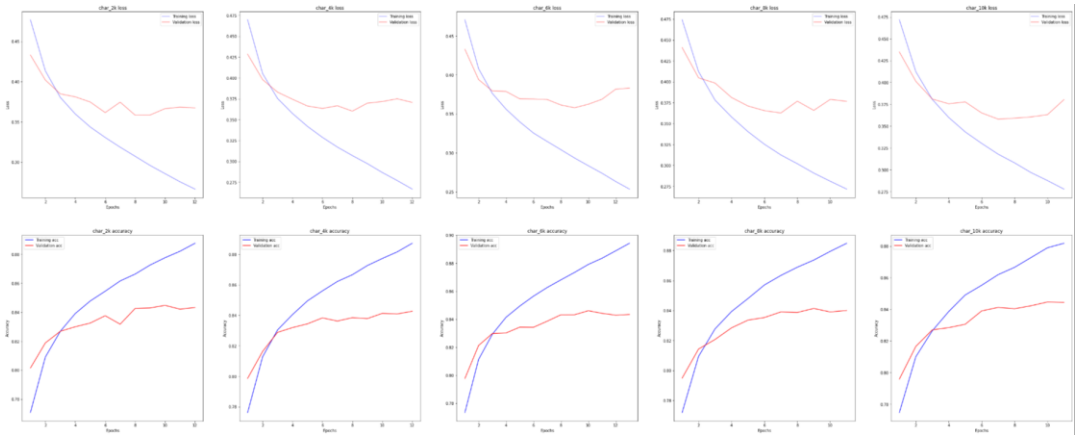
The results show the loss and the accuracy of each model and accuracy.

4.1 Unigram

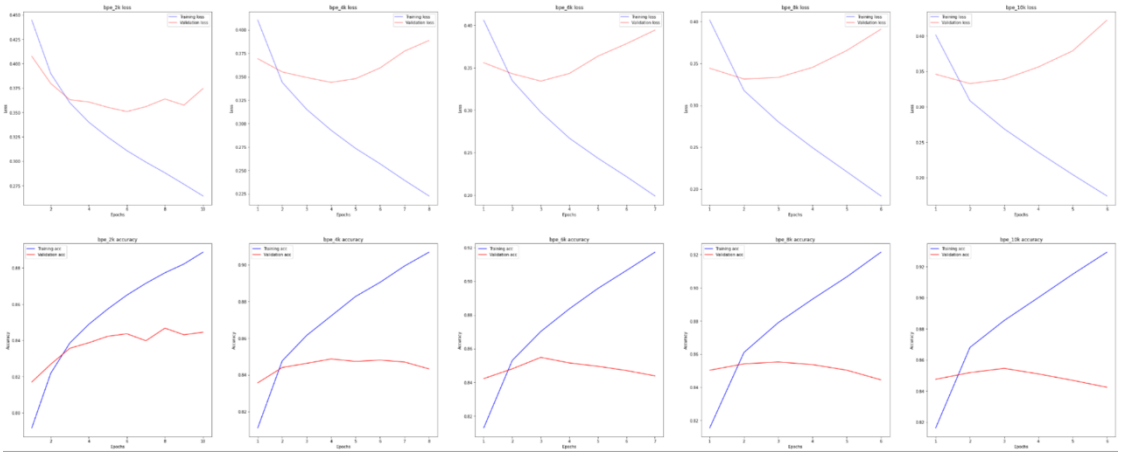


Unigram

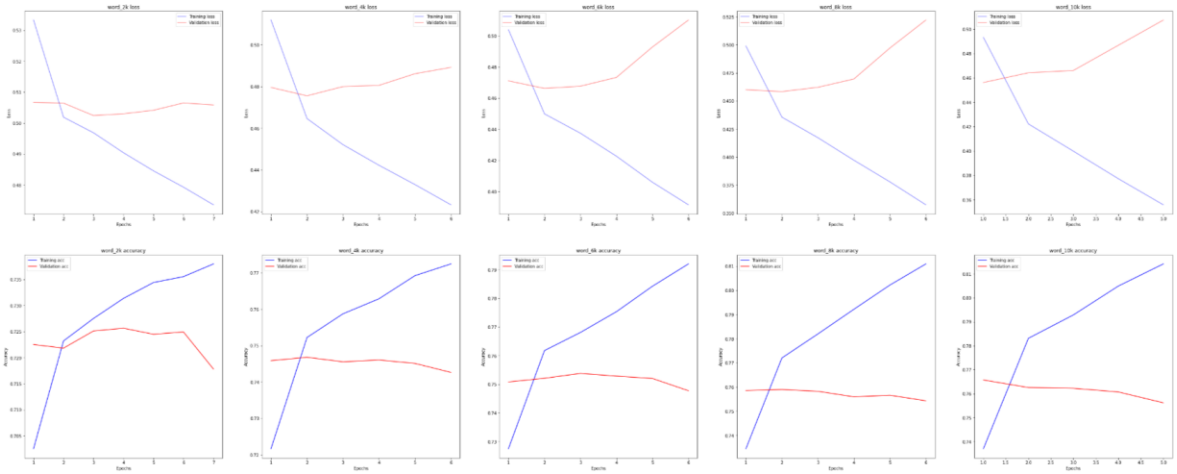
4.2 Bps



4.3 Char



4.4 word



As a result of 4 types of Sentence Piece Models, using word tokenizers with a vocab size of 10k. Table 1 shows the highest accuracy. shows the accuracy of each model with different vocab sizes. The number of total parameters is constant at 344897.

Accuracy	Unigram	BPE	Char	Word
2k	0.513036	0.505128	0.497729	0.740526
6k	0.493032	0.506225	0.500970	0.768008
8k	0.491466	0.500979	0.504815	0.784803
10k	0.501879	0.504032	0.490096	0.798622
Total Parameters	344,897	344,897	344,897	344,897

Table 1. shows the accuracy of each model with different vocab sizes. The number of total parameters is constant at 344897.

5 Conclusion

In this empirical study, we conducted sentiment analysis using four different Subword Piece Models (SPM) with varying vocabulary sizes, including unigram, BPE (Byte Pair Encoding), char (character-level), and word-level tokenization. The vocabulary sizes ranged from 2,000 (2k) to 10,000 (10k) tokens.

Our findings reveal that among the SPM models, the word-level tokenization with a vocabulary size of 10,000 tokens consistently achieved the highest accuracy in sentiment analysis tasks. This result underscores the importance of fine-grained subword representations in capturing nuances of sentiment within text data. The larger vocabulary size enabled the model to better handle out-of-vocabulary (OOV) words and capture semantic subtleties, leading to superior performance.

Furthermore, our study demonstrates that the choice of tokenization strategy plays a pivotal role in sentiment analysis tasks. While word-level tokenization excelled in our experiments, it is essential to consider the specific requirements and characteristics of the NLP task at hand when selecting an appropriate tokenization method.

In conclusion, our research highlights the effectiveness of word-level SPM tokenization with a 10,000-token vocabulary in sentiment analysis, emphasizing the significance of subword modeling in NLP applications. These insights contribute to the ongoing efforts to enhance the accuracy and robustness of sentiment analysis systems in real-world scenarios.

References

[1] Kudo, Taku, and John Richardson. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing." *arXiv preprint arXiv:1808.06226* (2018).

[2] Kudo, Taku. "Subword regularization: Improving neural network translation models with multiple subword candidates." *arXiv preprint arXiv:1804.10959* (2018).

[3] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." *arXiv preprint arXiv:1508.07909* (2015).