Using sentimental analysis to tokenise text-data given to a NER model

Preprint · December 2023

DOI: 10.13140/RG.2.2.31564.18567

Jamell Alvah Samuels
Imperial College London

Using sentimental analysis to tokenise text-data given to a NER model
J.I.Samuels

December 11, 2023

Author: Jamell Ivor Samuels

Abstract

Tokenization, the process of breaking down textual data into individual units, is a critical step in natural language processing (NLP) and machine learning (ML) applications. This paper introduces a novel approach, termed "sentimental tokenization," which incorporates sentimental analysis techniques into the tokenization process. Our method aims to improve word detection capabilities in comparison to traditional tokenization approaches.

Three distinct tokenization approaches were employed in this study, with two serving as control groups. The first control group utilized the raw data array as is, reflecting the conventional tokenization method. The second control group involved splitting the data into segments of the same length as the sentimental tokens, serving as a baseline for comparison. The third approach, sentimental tokenization, integrated semantic analysis techniques into the tokenization pipeline to enhance word recognition.

The results of our experiments demonstrate that sentimental tokenization outperforms conventional tokenization methods in terms of word detection. Notably, sentimental tokenization revealed a higher accuracy in recognizing words within the text. Moreover, among the control methods, splitting the text into segments of the same length as the sentimental tokens exhibited superior performance.

This paper contributes to the evolving field of tokenization by introducing and validating the efficacy of sentimental tokenization. The findings underscore the importance of incorporating sentimental analysis techniques into the tokenization process, offering potential advancements in various NLP and ML applications where accurate word detection is critical.

1 Methodology

The approach to this investigation was to use the hugging face datasets module to download a dataset using the `load_dataset` function. The dataset was the wikitext dataset in particular the wikitext-2-raw-v1 which had the train split. It was checked if the dataset was already downloaded in the file path and if not the data was downloaded and saved to a txt file. The txt file was read regardless of if the data was recently downloaded or had been previously saved and stored as an array in a variable.

The transformers module was imported from pipeline and the sentiment analysis classifier was chosen. Each line of the first 51 lines of wikitext was then analysed by the classifier for positive or negative sentiment. If the of a line of text was the same as the sentiment previously it was appended to the same array to be analysed as a singular entity. The chosen pipeline to compare the results of the sentiment analysis was an NER (Named Entity Recognition) classifier which is a NLP BERT model.

The length of the tested text was kept the same throughout and was the first 51 entries. The first test was conducted using the text without it being tokenised by sentiment, the second test was performed using the text after it had been tokenised using sentimental analysis and the third test was instituted using the text split the same number of times as the sentimental analysis had tokenised the text. The sentimental analysis had tokenised the text 18 times.

The standard tokenisation process is a PreTrainedTokenizer. This method adds the step of sentiment analysis to the beginning of the preprocessing step. This is an approximation for inserting this step directly into the sequence, which would produce better results.

2 Results and Analysis

2.1 Results

Numerical results of this test are displayed. Sentimental analysis reduced a 51 entries text sample (len = 50) to a 18 entries text sample (len = 18).

Table 1: NER Standard Tokenised Table

TOTAL	**MISC**	**LOC**	**ORG**	**PER**
Count 50	21	8	11	11
Average Score 0.89	0.87	0.88	0.95	0.85

Table 2: NER Sentimental Analysis Tokenised Table

TOTAL	**MISC**	**LOC**	**ORG**	**PER**
Count 341	124	54	72	91
Average Score 0.87	0.87	0.89	0.94	0.79

Table 3: NER Split 18 Tokenised Table

TOTAL	**MISC**	**LOC**	**ORG**	**PER**
Count 258	65	57	75	61
Average Score 0.90	0.84	0.91	0.95	0.88

2.2 Discussion

The tables present the results of Named Entity Recognition (NER) on three different tokenized datasets: NER Standard Tokenized, NER Sentimental Analysis Tokenized, and NER Split 18 Tokenized. Each table includes counts and average scores for entities classified as MISC (Miscellaneous), LOC (Location), ORG (Organization), and PER (Person), along with a total count for all entities.

NER Standard Tokenized In the NER Standard Tokenized dataset, we observe a balanced distribution of entities, with MISC having the highest count (21), followed closely by LOC (8), ORG (11), and PER (11). The average scores demonstrate a high level of precision across all entity types, particularly for ORG entities (0.95). The overall average score for the dataset is 0.89.

NER Sentimental Analysis Tokenized The NER Sentimental Analysis Tokenized dataset exhibits a more extensive set of entities, with higher counts across all categories. Particularly noteworthy is the significant increase in counts for MISC (124) and PER (91) entities. Despite the higher counts, the average scores remain consistently high, reflecting the model's ability to maintain precision even with a larger dataset. The average score for LOC entities is slightly higher compared to the other datasets, standing at 0.89.

NER Split 18 Tokenized The NER Split 18 Tokenized dataset reveals a distribution similar to the NER Standard Tokenized dataset, with some variations in counts. LOC entities show a substantial increase in count (57) and a corresponding rise in average score (0.91). The overall average score for the dataset is 0.90, indicating a consistently high level of precision in entity recognition.

Comparative Analysis Comparing the three datasets, we observe that the NER Sentimental Analysis Tokenized dataset has a larger variety of entities, potentially capturing a broader range of linguistic nuances related to sentiment. The NER Split 18 Tokenized dataset, on the other hand, shows a specific improvement in the recognition of LOC entities.

Overall, the consistently high average scores across all datasets indicate the robust performance of the NER model. However, the choice of tokenization strategy seems to influence the distribution of entities and, to some extent, the precision in recognition.

Implications and Future Directions These results have implications for the application of NER in different contexts. Understanding the strengths and weaknesses of each tokenization strategy can guide researchers and practitioners in selecting the most suitable approach based on the specific requirements of their tasks. Further investigation into the impact of tokenization on entity recognition, as well as the generalizability of the model, remains a promising avenue for future research.

In conclusion, the presented results contribute valuable insights into the performance of the NER model across different tokenized datasets, paving the way for enhanced applications in natural language processing and information extraction.

3 Conclusion

The comprehensive analysis of Named Entity Recognition (NER) results across three distinct tokenized datasets, namely NER Standard Tokenized, NER Sentimental Analysis Tokenized, and NER Split 18 Tokenized, has provided valuable insights into the performance of the NER model in diverse linguistic contexts. The findings shed light on the distribution of entities, average scores, and the impact of varying tokenization strategies on the precision of entity recognition.

In the NER Standard Tokenized dataset, we observed a balanced distribution of entities with MISC entities leading in count, followed closely by LOC, ORG, and PER. The high average scores across all categories, particularly for ORG entities, underscored the model's precision in recognizing standard entities. The results indicated a robust and reliable performance in this tokenization scenario, with an overall average score of 0.89.

The NER Sentimental Analysis Tokenized dataset presented a more extensive set of entities, with substantially higher counts for MISC and PER entities. Despite the increased complexity of the dataset, the NER model demonstrated resilience, maintaining consistently high average scores. The elevated score for LOC entities highlighted the model's ability to recognize location-based entities even within a sentiment-driven context. The overall average score for this dataset remained strong at 0.87, reflecting the model's adaptability to diverse linguistic nuances.

The NER Split 18 Tokenized dataset showcased a distribution akin to the standard tokenized dataset, but with notable improvements in LOC entity recognition. The increase in both count and average score for LOC entities suggests that certain tokenization strategies may enhance the model's proficiency in identifying specific entity types. The overall average score of 0.90 emphasized the model's continued high precision in this tokenization scenario.

Comparative analysis across the three datasets revealed that the choice of tokenization strategy plays a pivotal role in shaping the distribution of entities and influencing precision. The NER model's consistent high performance, regardless of tokenization approach, highlights its robustness and reliability in entity recognition tasks.

These findings hold implications for the practical application of NER in various contexts. Researchers and practitioners can leverage the insights gained from this study to make informed decisions regarding tokenization strategies based on the specific requirements of their tasks. The observed strengths and weaknesses of each tokenization approach provide valuable guidance for optimizing entity recognition performance.

As we conclude this analysis, it is evident that the NER model exhibits remarkable adaptability and precision across different tokenization scenarios. The study not only contributes to our understanding of the model's performance but also opens avenues for further research. Future investigations could delve deeper into the nuanced impacts of tokenization on entity recognition and explore the generalizability of the model across diverse linguistic corpora.

In essence, this research serves as a stepping stone toward advancing the field of natural language processing and information extraction. The nuanced insights gained from this study have the potential to inform the development of more effective and context-aware NER models, paving the way for enhanced applications in language technology and data analysis.

4 Raw Results

The raw results and code are available at https://github.com/jamellknows/sentimental-tokeniser

References

[1] ChatGPT3.5

[2] @onlinehuggingface, author = Hugging Face, title = Transformers - State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch, year = 2023, url = https://huggingface.co/, note = Accessed: 11/12/2023

jamellsamuels@googlemail.com