K-Rater: Unveiling Korean Drama Ratings Through Review Analysis

Ewa Słowińska

Department of Philosophy, Linguistics and Theory of Science (FLoV)

Machine Learning For Statistical NLP: Advanced LT2326

University of Gothenburg

Abstract

Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling computers to understand and generate human language. The importance of NLP lies in its capacity to bridge the gap between humans and computers, thus faciliating tasks such as text classification or sentiment analysis. Previous research applied such tasks to numerous domains including tourism (Puh & Babac, 2022), medicine (Agarwal, 2019) and entertainment (Shen et al., 2017). The current project aimed to develop a Bidirectional Long Short-Term Memory model (BiLSTM) in order to conduct a multiclass score classification and sentiment analysis on Korean drama reviews. Four experiments were conducted using the same model skeleton: (1) multiclass score classification with 19 classes, (2) multiclass score classification with 10 classes, (3) sentiment analysis with weights to account for class imbalance, (4) sentiment analysis with undersampling to account for class imbalance. The results showed that the model's performance improved when the number of classes decreased. In the initial experiment, the model achieved an accuracy of 17% with a corresponding mean absolute error (MAE) of 3.4. Subsequently, the second experiment exhibited improved performance, yielding an accuracy of 23% and a MAE of 1.7. The third experiment showcased a substantial accuracy increase to 78%, while the fourth experiment maintained an accuracy level of 58%.

Background

Natural Language Processing (NLP) is a field of artificial intelligence dating back to 1950s when the first electronic computers started emerging. Initially employed for language translation, NLP rapidly expanded its applications, covering diverse tasks such as text classification and sentiment analysis. The former is a task which involves categorising pieces of text into predefined classes, making it easier for humans to retrieve and analyse relevant information from vast amounts of data. In contrast, sentiment analysis, a subfield of text classification, sets its goal to classify text based on its emotional tone.

Text classification and sentiment analysis find their applications in various domains. In Agarwal's project from 2019, the researcher created a medicine-oriented model which classified pill reviews based on the reviewer's condition. Another study by Puh and Babac in 2022 aimed to construct a multiclass classification model categorising hotel reviews by their score. Additionally, Shen et al. utilised an IMDB dataset in 2017 to categorise movie reviews based on their sentiment.

In each of the aforementioned projects, the researchers consistently opted for the Bidirectional Long Short-Term Memory (BiLSTM) model for its adept handling of sequential data, ability to account for long-term dependencies and excellent understanding of contextual intricacies.

This project aims to leverage the capabilities of BiLSTM in conducting score classification and sentiment analysis of Korean drama reviews. Korean dramas, television series produced in South Korea, have recently gained popularity both domestically and internationally. As they cover a wide range of themes and genres, they cater to various age groups. Korean drama reviews are a good source of data for text classification as they contain an emotionally-charged language, potentially aiding the performance of machine learning models.

This project encompasses four different experiments conducted on the same model skeleton: (1) score classification with 19 or (2) 10 classes as well as (3) binary sentiment analysis with either weight tensor or (4) undersampling implemented to account for the class imbalance.

The upcoming sections provide an overview of the dataset employed in this project. The methodology in then explained in fine detail, shedding light on the model's architecture. Following this, the results of every experiment are presented, setting the stage for a comprehensive discussion of the findings and conclusions drawn from this study.

Data

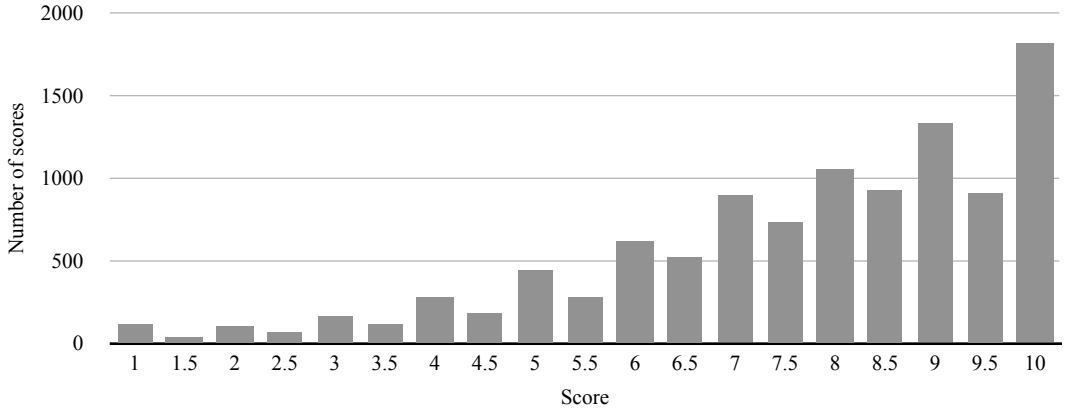
The dataset employed in this project originates from Kaggle, a prominent platform for data science and machine learning enthusiasts. Created in 2023 by Charuchinda, the dataset consists of Korean drama reviews spanning the years 2015 to 2023. See Figure 1 for an example review. The data was extracted from MyDramaList, which describes itself as a community-driven platform providing a space for fans to connect and share their passion for Asian dramas and movies.

Figure 1: Example user review of “Sing My Crush”

Sing My Crush	*“it's a cute healing kbl drama Oh My god, I loved this drama so much, and that last scene with the spotlight being on Hantae and Baram being in the darkness and then him slowly walking into the light facing Hantae, representing that in Baram's darkness and times of difficulties, Hantae was the light that was always with him. It is such a beautiful story of healing and love and overcoming past scars together. The music was also really good, I especually loved the last song. and i actually really like the second couple. They were so healthy and secure”*	*9*

The dataset consists of 10,625 reviews, out of which six had to be dropped due to the lack of their corresponding score data. The reviews exhibit diversity in length, with the briefest containing a single token and the most extensive extending to 17,509 tokens. On average, each review comprised of approximately 1,922 tokens.

Figure 2: Score distribution



The scores, ranging from one to ten with half-point increments, constitute 19 distinct classes. Notably, the score distribution is uneven, characterised by a surplus of positive ratings as compared to the negative ones. Moreover, users exibit a clear tendency to assign “full” points rather than “halves”. The most frequent score of 10 occurs more than 42 times as often as the least frequent score of 1.5. To address this class imbalance in the ensuing experiments, one of two strategies was implemented: the assignment of weight tensors to the loss function or the undersampling of the dataset. These methods will be further explained in the subsequent sections.

Methodology

In experiment 1, the original number of 19 classes was retained. Experiment 2 narrowed down the number of classes to 10, by merging “full” points with their “halved” counterparts (e.g. 1 with 1.5). For experiments 3 and 4, a binary classification approach was adopted, categorizing scores from 1 to 5.5 as "negative" and scores from 6 to 10 as "positive," resulting in two classes.

For experiment 1-3, weight tensors were implemented to counter the class imbalance. The weights were assigned based on a simple inverted frequency calculation. Conversely, experiment 4 employed undersampling of the majority class, reducing it to 1819 reviews to align with the minority class.

Post-cleaning, the dataset underwent a split into train, validation and test sets with a customary ratio of 70:15:15. Subsequently, a custom dataloader was crafted to preprocess the data by removing stopwords, lowercasing the text and tokenising it. A sample review before and after preprocessing can be seen in Figure 3. The dataloader object returned score and review vocabularies of lengths 19/10/2 and 140,252, respectively as well as train, validation, and test iterators.

Figure 3: Example review before and after preprocessing

I really love this drama because it's so cuteee! From the first airing, I'm already fall in love with it because of the light story But after a few episodes, the story is not light like I think before The plot is like make viewers guessing what happen here and make me really curious what happens next! I like the plot that make me curious Amazing acting from the casts. Especially Choi Woo Sik and Lee Soo Kyung. Uee and Seulong acting is improved too here And I love almost all of the soundtrack Well, the drama is really worth to watch if you want a romance comedy although Do Do Hee character is a bit annoying at first. But I love it!"	*['really', 'love', 'drama', "it's", 'cuteee!', 'first', 'airing,', "i'm", 'already', 'fall', 'love', 'light', 'story', 'episodes,', 'story', 'not', 'light', 'like', 'think', 'plot', 'like', 'make', 'viewers', 'guessing', 'happen', 'make', 'really', 'curious', 'happens', 'next!', 'like', 'plot', 'make', 'curious', 'amazing', 'acting', 'casts.', 'especially', 'choi', 'woo', 'sik', 'lee', 'soo', 'kyung.', 'uee', 'seulong', 'acting', 'improved', 'love', 'almost', 'soundtrack', 'well,', 'drama', 'really', 'worth', 'watch', 'want', 'romance', 'comedy', 'although', 'hee', 'character', 'bit', 'annoying', 'first.', 'love', 'it!']*

Across all experiments, the number of training epochs was set to 50, with early stopping implemented, exhibiting patience for ten epochs to prevent overfitting. The model processed the data in batches of either 8 or 32, depending on the condition. During training, the model iteratively converted reviews into embeddings, passed them through a Bidirectional Long Short-Term Memory (BiLSTM) layer, applied a dropout of 0.3, and forwarded the output to a linear classification layer, producing predicted scores. Next, a CrossEntropyLoss function and Adam optimizer were applied and the training continued until the last epoch was reached or until an early stopping was triggered.

Results

Experiment 1 encompassed five distinct conditions, each characterized by unique parameter settings. A comprehensive overview of the results can be found in Table 1. Optimal accuracy and recall emerged in the second condition, where the batch size equaled 32, embeddings' size was set at 256, the hidden layer measured 512, and the learning rate stood at 0.001. Condition 5 yielded the highest macro F1 score and Mean Absolute Error (MAE), maintaining the same parameter configuration as condition 2, except for a shift in batch size to 8 and a reduction in the hidden layer's size to 64. Meanwhile, the best precision was achieved in condition 1, where the parameters of condition 5 were copied, but with a hidden layer's size reverting to 512.

Table 1: Results of experiment 1

Condition	Accuracy	Macro F1	Precision	Recall	MAE
batch size = 8 embeddings' size = 256 number of hidden layers = 1 hidden layer's size = 512 learning rate = 0.001	16.8%	6.0%	14.4%	16.8%	4.5
batch size = 32 embeddings' size = 256 number of hidden layers = 1 hidden layer's size = 512 learning rate = 0.001	16.8%	2.1%	11.9%	16.8%	4.5
batch size = 8 embeddings' size = 256 number of hidden layers = 1 hidden layer's size = 512 learning rate = 0.1	15.8%	3.9%	10.5%	15.8%	4.0
batch size = 8 embeddings' size = 64 number of hidden layers = 1 hidden layer's size = 512 learning rate = 0.001	16.8%	2.8%	12.0%	16.8%	4.5
batch size = 8 embeddings' size = 256 number of hidden layers = 1 hidden layer's size = 64 learning rate = 0.001	14.2%	6.2%	13.8%	14.2%	3.6

After choosing the optimal parameter configuration from the results of the first experiment, the remaining three experiments were conducted with the settings of condition 1. This entailed the batch size of 8, embeddings' size at 256, a hidden layer of 512, and a learning rate of 0.001. Experiment 2 yielded relatively improved results, featuring an accuracy of 23.4%, a macro F1 score of 16.9%, precision at 23.9%, an augmented recall of 23.4%, and a relatively stable Mean Absolute Error (MAE) holding at 1.7.

The third experiment exhibited even more commendable performance, with accuracy surging to an impressive 77.7%, a macro F1 score escalating to 63.1%, precision reaching 79.1%, and recall attaining 77.7%.

In the fourth experiment, the results were poorer. Accuracy dropped to near-chance 57.9%, macro F1 score plummeted to 57.7%, precision contracted to 58% and recall decreased to 57.9%.

Discussion

The results emphasise the pivotal role of class count in multiclass classification tasks. It appears that a higher number of classes decreases the model's' ability to generate correct predictions. This may be caused by several factors. Firstly, with the rise in the class count, the complexity of the task increases, thus necessitating the model to learn a more extensive array of information. What follows, with a higher class count, the boundaries between the classes blur, leading to more ambiguity. Moreover, increasing the class count may often lead to bigger data sparsity and thus diminish the model's exposure to learning material.

The greatest difference between experiments 3 and 4 lies in the volume of data presented to the model. The model from experiment 3 showcased a significantly better performance from its counterpart in experiment 4, where undersampling was applied. This highlits the importance of prominent datasets and implies that opting for weight tensors might be a better strategy than undersampling when confronted with class imbalance.

Another noteworthy revelation from this project is that machine learning models can be relatively sensitive to parameter configurations. Changing the parameters in the first experiment indeed influenced the results, if only by marginal percentage shifts. This highlights the importance of the nuanced understanding of the interplay between various parameters.

Overall, the model's performance, while sufficient, falls short of the desired level. Although the results were generally higher than chance, there remains room for much development. Surely, more time and developing is needed to reach an objectively commendable performance.

Conclusion

This project focused on developing a machine learning model geared towards conducting multiclass score classification and sentiment analysis on Korean drama reviews. Inspired by the previous research in the field of NLP, BiLSTM was employed as a model of choice. The exploration revealed a number of valuable insights into the nature of machine learning and the chosen tasks. Firstly, it accentuates the importance of class count and its impact on the model's performance. Moreover, it stresses the significance of prominent datasets and the sensitivity of machine learning models to parameter configuration. In summary, this projects not only presents a functional machine learning model for Korean drama reviews but also contributes valuable insights applicable to broader applications in NLP and sentiment analysis.

References

Agarwal, R. (2020, March 16). *Multiclass text classification - pytorch.* Kaggle. https:www.kaggle.com/code/mlwhiz/multiclass-text-classification pytorch

Puh, Karlo & Bagic Babac, Marina. (2022). Predicting sentiment and rating of tourist reviews using machine learning. Journal of Hospitality and Tourism Insights. 6. 10.1108/JHTI-02-2022-0078

Qianzi Shen, Zijian Wang, Yaoru Sun. Sentiment Analysis of Movie Reviews Based on CNN-BLSTM. 2nd International Conference on Intelligence Science (ICIS), Oct 2017, Shanghai, China. pp.164-171, ff10.1007/978-3-319-68121-4_17ff. ffhal-01820937f