Analysis and Classification of Drug Ratings

Classification of Drug Ratings Using Deep Learning Models

By: Ziv Fenigstein

Introduction

This study examines the use of BERT models for classifying drug ratings from textual reviews into 10-class and 3-class systems. The aim is to leverage deep learning to understand and predict user sentiments expressed in drug reviews accurately. Results indicate high effectiveness in classification, showcasing the models' capabilities in handling nuanced textual data.

The evaluation of drug reviews presents significant challenges due to the subjective nature of user-generated content. This project aims to classify drug ratings accurately using BERT models, improving the understanding of consumer feedback on medications.

Dataset

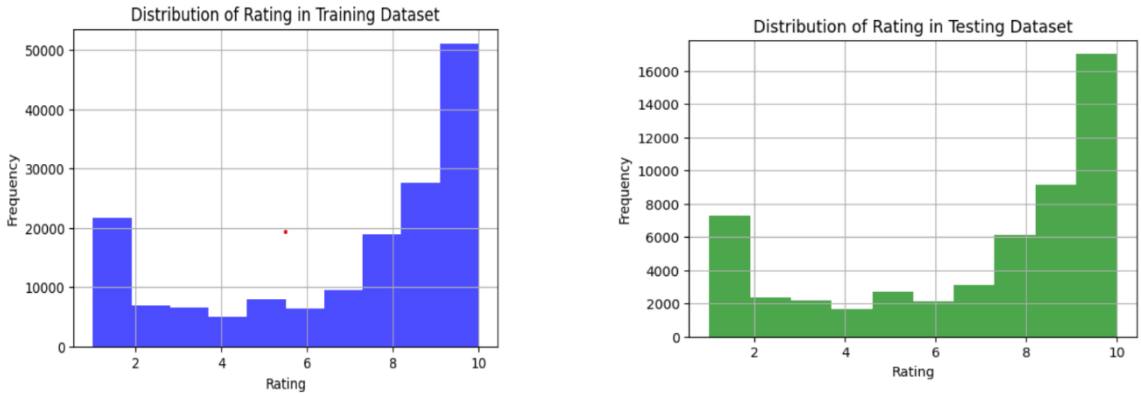
The Drug Review Dataset from Drugs.com, hosted by the UCI Machine Learning Repository, is a rich collection of patient reviews on various drugs alongside related medical conditions. It features a 10-star patient rating system that reflects overall patient satisfaction. The dataset includes 215,063 instances across six features, including drug names, conditions, reviews, ratings, review dates, and a count of how many found the review useful.

Data analysis

The dataset comprises a pandas DataFrame containing 161,297 entries and seven features, each providing valuable insights into drug reviews. These features include an index column ('Unnamed: 0') for row identification, the names of drugs reviewed ('drugName'), and the corresponding medical conditions ('condition'). While most entries are complete, the 'condition' column exhibits 899 missing values. Other features such as 'review', 'date', and 'usefulCount' present textual reviews, review dates, and counts of how many users found the review helpful, respectively, all fully populated without any nulls. Lastly, the 'rating' feature denotes integer ratings given by users, with no missing values. This succinct overview underscores the dataset's comprehensive structure, with the noted absence of data

in the 'condition' column requiring attention for certain analyses or applications.

Class distribution



The histograms show the rating distribution in training and testing datasets, both displaying a similar bimodal pattern with peaks at the lowest and highest ratings. Most ratings are at 10, indicating a prevalence of highly favorable reviews, while mid-range ratings (3 to 7) are less common, forming the minority. This consistent pattern across both datasets suggests a skew towards extreme ratings, which could influence the bias of machine learning models trained with this data.

First try

In my first attempt, I built a logistic regression model without using the review feature. Instead, I replaced the original date feature with a new one called 'years_passed' to focus on how time influences the outcomes. This approach helped simplify the model and explore the impact of time on the data.

New dataset Example:

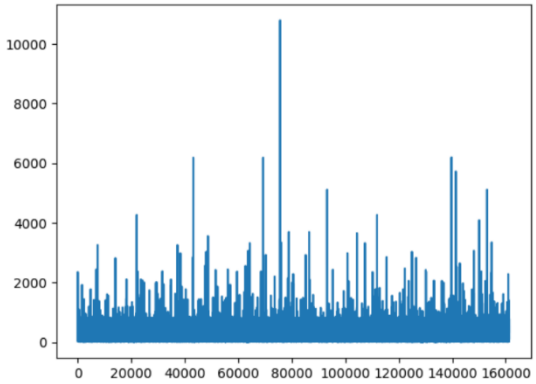
	Unnamed: 0	drugName	condition	rating	usefulCount	years_passed
0	206461	Valsartan	Left Ventricular Dysfunction	9	27	12
1	95260	Guanfacine	ADHD	8	192	14
2	92703	Lybrel	Birth Control	5	17	15
3	138000	Ortho Evra	Birth Control	8	10	9
4	35696	Buprenorphine / naloxone	Opiate Dependence	9	37	8

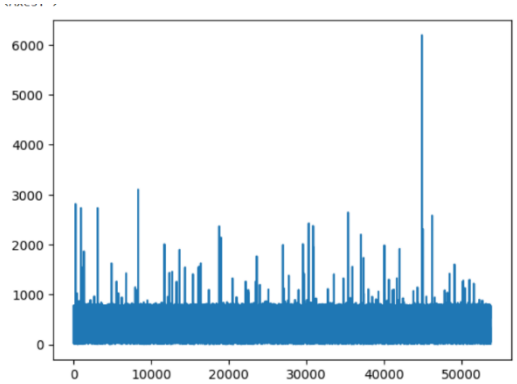
The logistic regression model yielded poor results with an accuracy of 32.11% and an F1 score of 21.80%, indicating significant prediction challenges. The lack of review data, rich in contextual and sentiment insights, likely contributes to this underperformance. Incorporating natural language processing (NLP) could enhance accuracy as NLP models excel at analyzing text to capture nuanced sentiments and contexts, which are missed by simpler models like logistic regression.

```
Accuracy: 0.32107873527588343
F1 Score: 0.21800413570571026
```

Analysis and Classification of Drug Ratings

Analysis before BERT





The histograms illustrate the distribution of word counts in comments for both training (first graph) and testing (second graph) datasets, with the x-axis representing the number of comments and the y-axis showing the word count. In both datasets, most reviews consist of fewer than 2,000 words. To simplify the data processing and make the model more manageable, reviews exceeding 2,000 words have been filtered out. This decision likely helps in reducing noise and computational complexity, enabling more efficient training and testing of the machine learning models by focusing on more concisely written reviews.

10 class classification with BERT

In this project, the BERT model is fine-tuned for sequence classification using the Transformers library and PyTorch. Initially, text data is preprocessed through tokenization and encoding into formats suitable for BERT, including input IDs and attention masks. These are wrapped in PyTorch datasets and loaded into DataLoaders for efficient batch processing during training. The model, BertForSequenceClassification, is configured with dropout and hidden state outputs tailored for the classification task. Training is conducted over multiple epochs with an AdamW optimizer and a learning rate scheduler to optimize performance. Each epoch involves processing batches, calculating loss, and updating model weights, with gradient clipping employed to stabilize training. Model performance is monitored through loss, accuracy score and F1 scores, and checkpoints are saved periodically. This structured approach leverages BERT's pre-trained capabilities, enhancing its applicability to specific classification tasks while ensuring robust generalization to unseen data.

Analysis and Classification of Drug Ratings

Epochs comparison

The training results for the BERT model across ten epochs reveal substantial improvements across all key performance metrics. Initially, the model began with a validation loss of 1.30143, which steadily decreased to 0.596518 by the tenth epoch, indicating a significant reduction in prediction error as the model adapted to the dataset. Concurrently, the F1 Score, measuring the model's accuracy in identifying the positive class, improved from 0.443593 to 0.806019. This improvement reflects enhanced balance and precision in the model's predictions. Accuracy also saw a progressive increase, beginning at 0.50605 and escalating to 0.811697, demonstrating the model's growing proficiency in correctly classifying both classes over time.

Epoch	Validation Loss	F1 Score	Accuracy
1	1.30143	0.443593	0.506055
2	1.20107	0.477667	0.53478
3	1.06572	0.56457	0.595288
4	0.917055	0.645548	0.66336
6	0.73586	0.737759	0.744865
7	0.590592	0.806642	0.810374
8	0.721695	0.745709	0.751106
9	0.6356	0.792754	0.799132
10	0.596518	0.806019	0.811697

Results on test

The results on the test set indicate that the BERT model achieved an accuracy of 64.06% and an F1 Score of 63.83%. These metrics suggest that while the model demonstrates a reasonable ability to generalize to new data, there is still room for improvement. The disparity between the training performance and the test results may point to issues such as overfitting during the training phase or limitations in the model's ability to handle the diversity or complexity of the test data. Future efforts could focus on strategies such as more extensive data preprocessing, exploring different model architectures.

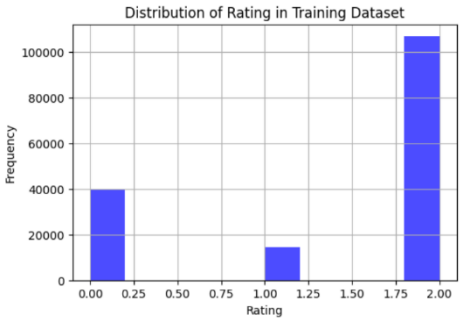
```
Accuracy on test set: 0.6406279061116691
F1 Score on test set: 0.6382534573506756
```

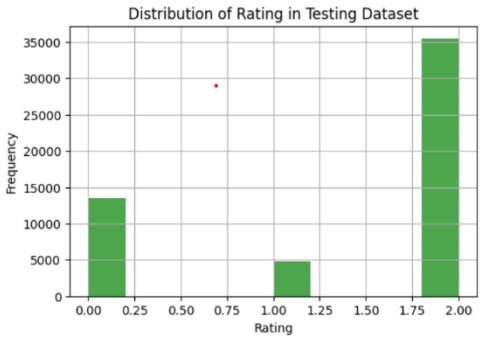
Analysis and Classification of Drug Ratings

3 class analysis

The histograms show the distribution of ratings in the three-class system for both the training and testing datasets, where the classes are divided as follows: ratings less than or equal to 4 as class 0, ratings between 4 and 7 as class 1 and ratings greater than or equal to 7 as class 2 .

Both datasets exhibit a similar distribution pattern, indicating a consistent approach to class division between training and testing.





3 class classification with BERT

In the three-class classification scenario using BERT, the dataset ratings were categorized into three distinct classes based on their values: ratings less than or equal to 4 were labeled as class 0, ratings between 4 and 7 as class 1, and ratings greater than or equal to 7 as class 2. This approach aimed to simplify the output space and potentially improve the model's ability to differentiate between low, medium, and high sentiments. The preprocessing for this classification was largely similar to that of the 10-class configuration, maintaining consistent methods for tokenization and encoding. However, a significant modification was the adoption of the BioBERT model instead of the standard BERT. BioBERT, a domain-specific variation of BERT pre-trained on biomedical literature, was chosen potentially for its enhanced semantic understanding of context, which could be advantageous depending on the nature of the text data. This adjustment in the model architecture aimed to leverage BioBERT's specialized training to better capture and classify the nuanced differences among the three defined rating classes.

Analysis and Classification of Drug Ratings

Epochs comparison

The validation results from the 3-class classification using the BioBERT model over five epochs indicate a strong and relatively stable model performance. The F1 Score and Accuracy consistently increased from the first to the fourth epoch, peaking at an F1 score of approximately 0.899 and an accuracy of nearly 89.89%. However, both metrics saw a slight decline in the fifth epoch, suggesting the model reaching its performance limit with the given training data and hyperparameters.
Overall, the model shows high effectiveness in classifying the data into the three specified classes with both high accuracy and F1 scores, indicating a successful adaptation of BioBERT to this specific task.

Epoch	Validation Loss	F1 Score	Accuracy
1	0.39235	0.840725	0.859169
2	0.337654	0.874895	0.878313
3	0.347275	0.88889	0.889932
4	0.399487	0.899463	0.898863
5	0.472309	0.887634	0.892909

Results on test

The BioBERT model's performance on the test set for the 3-class classification task shows impressive results, achieving an accuracy of 90.08% and an F1 Score of 90.16%. These metrics indicate a high level of precision and reliability in classifying the ratings into the three predefined categories. The consistency of both accuracy and F1 Score above 90% suggests that the model not only correctly identifies the majority of classes but also maintains a balanced precision and recall across the dataset. This strong performance on the test set confirms the model's ability to generalize well to new, unseen data, underscoring its effectiveness for practical applications in this classification domain.

```
Accuracy on test set: 0.9008018455470596
F1 Score on test set: 0.901572313747015
```

Conclusions

In this comprehensive project, we explored the application of advanced NLP models for sentiment analysis, specifically focusing on classifying textual reviews into multiple classes. Initially, a 10-class classification model was developed using BERT, which demonstrated substantial improvements across various metrics through the epochs but exhibited room for improvement in generalizing to the test set. Following this, a shift to a more focused 3-class classification system utilizing the specialized BioBERT model yielded highly encouraging results, with both validation and test performances showcasing accuracies and F1 scores above 90%. This transition underscores the efficacy of using domain-specific models like BioBERT for more targeted classification tasks, which significantly enhanced the model's predictive accuracy and reliability.

References

1. UCI Machine Learning Repository. (n.d.). Drug Review Dataset (Drugs.com). Retrieved from https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com
2. Paul, R. (n.d.). FineTuning BERT for Multi-Class Classification Turkish. Retrieved from https://github.com/rohan-paul/LLM-FineTuning-Large-Language-Models/blob/main/Other-Language_Models_BERT_related/FineTuning_BERT_for_Multi_Class_Classification_Turkish/Multi-class_Classification.ipynb
3. Saha, R. (2020). Multi-class text classification with deep learning using BERT. Towards Data Science. Retrieved from https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613
4. Hugging Face. (2020). Transformers v3.0.2: BertModel. Retrieved from https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#transformers.BertModel
5. Girish, K. (2019). Sentiment Analysis Using Deep Learning BERT. Medium. Retrieved from https://medium.com/@girish9851/sentiment-analysis-using-deep-learning-bert-adf975232da2
6. Shiju, A., & He, Z. (2022). Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models. IEEE International Conference on Healthcare Informatics. doi: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9744636/