Classifying Yelp User Rating with Yelp User Reviews:

An application of NLP and LSTM

SI 670 Final Project, 2020 Fall
Team: Machine Liers
Team members: Yuan Cheng, Xinyi Zhao, Meixin Yuan

Introduction

Yelp (yelp.com) is an online platform that allows users to submit star ratings and narrative reviews of local businesses. As of the second quarter of 2019, Yelp has reported to have a monthly average of over 61 million unique users visiting the platform through desktop and over 76 million unique users visiting Yelp with mobile app and website (Yelp, 2019). With millions of users contributing to the local knowledge on Yelp, Yelp data has become a rich library to learn user behaviors and user sentiments for scholars and businesses.

Among all the data yelp provides, the user reviews is the richest in content that attracts many researchers interests. For instance. Ranard et.al.(2016) explored the potential of yelp reviews as a supplement information source to inform traditional surveys of patient experience using natural language processing (NLP) techniques. They found that Yelp reviews could provide a better sense of the hospital's overall performance and even wider topics of patient experiences compared to the conventional Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey. As one of the most important categories of local business on Yelp, restaurants' reviews also gain attention. Huang and colleagues (2014) used Latent Dirichlet Allocation (LDA) algorithm to describe the latent subtopics from Yelp restaurant reviews in order to help restaurant owners improve their business. Luca (2016) found that online customer reviews might become a substitute for more traditional forms of reputation and Yelp stars are correlated to the restaurant revenues using a regression discontinuity framework.

Many studies regarding user reviews have applied traditional NLP and machine learning methods as the mentioned examples. Among the different machine learning models, Naïve Bayes is a common method for language processing tasks (Jurafsky & Martin,2019, Luca, 2016, Le and Nguyen, 2019). With the invention and broader application of deep learning techniques, deep

learning methods have also been introduced in the realm of NLP (Socher, 2012). Jelodar et. al. (2020) applied LSTM (Long-Short Term Memory) recurrent Neural Network Approach to classify the sentiment and topics of the COVID-19 discussions. According to Jelodar and colleagues, LSTM units can outperform the traditional RNN in terms of avoiding gradient vanishing or exploding by adjusting the information in forget, input, and output gates for each of the cells. Other research has also confirmed the effectiveness of applying LSTM neural networks in NLP tasks (Wang et. al., 2016).

For this project, we would like to explore the relationship between yelp user reviews and yelp user ratings of restaurants. User rating is a good indicator of user attitudes towards local business. Inspired by the novel approaches that integrate NLP with deep learning, this project aims to train an algorithm that predicts the classes of user ratings with user reviews. This algorithm could be used to detect the user attitudes to restaurants from other crowdsourcing narratives (e.g. Tweets about restaurants) and can potentially expand the rating system of local restaurants.

Methods

Since our main purpose is to build a deep learning algorithm to predict the rating of Yelp Restaurants by the latent semantic information from the reviews, we plan to build feasible deep learning models and implement the classification with the numerical data converted from the text data by natural language processing. We classified the user ratings into “good” and “not-good” categories and made predictions according to these tags. After implementing basic data cleaning steps, we chose restaurant reviews in Pennsylvania as our experiment dataset as it provides an adequate amount of records but not too much to process. We conducted an exploratory data analysis to demonstrate a granular image of our data.

Three machine learning models were trained for this project. A baseline model with Naïve Bayes Classifier and two deep learning models. Our first deep learning model is based on a regular deep learning algorithm for text data classification, with two dense layers of the algorithm ‘relu’ and the last layer of the algorithm “sigmoid.” The second model is based on the advanced deep learning algorithms specific for text data analysis, which includes a Conv1D layer, a LSTM layer and three dropout layers in addition to the three regular layers. We compared the accuracy and mean absolute error (MAE) of three models.

A high-level description of code

There are three main parts of the code. Firstly, the data cleaning part involves steps to retrieve. extract and combine necessary data from the original json dataset (package: pandas, numpy, json, ect.). Secondly, some basic features are displayed and visualized and two wordcloud images were provided (package: seaborn, wordcloud, etc.).

Finally, we created three machine learning algorithms. The Naïve Bayes is a simple traditional method of feature classifying and it is included to provide a baseline to be compared with more complex deep learning models (package: sklearn, etc.). To create deep learning models, we processed the text with the natural language methods (packages: nltk.tokenize, and nltk.corpus). In model one, we used the package keras(models, layers, losses, metrics, optimizers) to build the deep learning model and do the evaluation. In model two, we used the package keras (keras.preprocessing.text, Sequential, preprocessing.sequence), and the dense layers(LSTM, Dropout, Dense, Embedding, Conv1D, MaxPooling1D, Bidirectional, metrics.

Dataset overview

The datasets are retrieved from Yelp Dataset (https://www.yelp.com/dataset). It provides detailed information of businesses, users, reviews, etc. In our study, we will focus on the review and business datasets. The most important information came from "text" (the review of the business), "stars" (the rating score given by users), as well as the "categories" of the business. The "useful", "funny", "cool" features provide us some supplemental information about the preferences of users. According to our preliminary analysis of the dataset, more than half of reviews feature with 4 and 5 stars (Figure 1). Thus, we classified the reviews with the stars 4 or 5 as "good" and the reviews with 1 to 3 stars as "not-good." In addition, since deep learning models require a substantial amount of computational power but the mode data could result in better outcomes, we need to find a subset of data with appropriate amounts of records. Figure 2 shows the count of reviews by state. We finally chose data for the state of Pennsylvania for our project, which ranks the 6th for the count of reviews across the provided states.

Evaluation and Analysis

Brief description of your study design and/or evaluation methods

We combined the information of the two datasets and created a data frame with the reviews, stars, and other related information about all the restaurants of Pennsylvania. A new column "is_good" was created with the label "1" or "0" based on the value of "stars" (>=4.0 or <4.0) to indicate the user sentiment. This column will be used as the labels to feed the deep learning model.

Figure 1. Star rating makeup of reviews

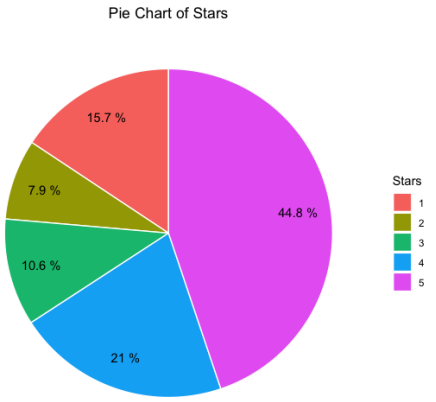
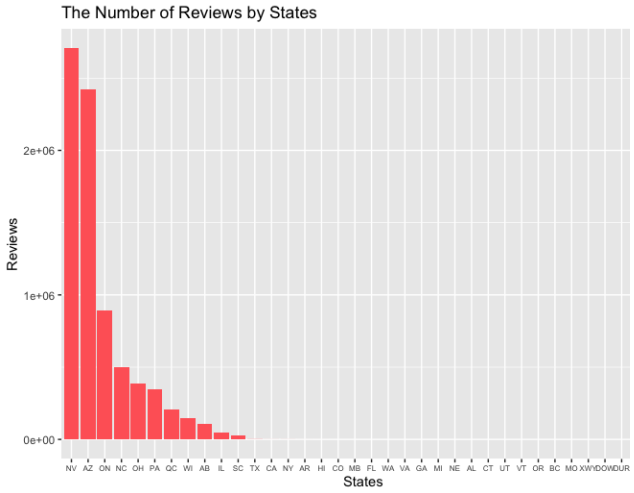


Figure 2. Count of reviews by State



A multinomial naïve Bayes (NB) classifier is set as the baseline model for this project. The multinomial NB is an appropriate approach to classify discrete features such as word counts or text classification (sklearn, 2020). To make our text data be able to be processed by multinomial NB, we first converted the reviews into a matrix of token counts with ConuntVectorizer function provided by sklearn.feature_extraction.text (sklearn, 2020).

To construct deep learning models, the content of "text" was isolated from the data frame and converted into a nest list of tokens. Only the most frequent 10000 words were kept for further study. And the list of tokens was converted into a numeral matrix, which will be used for model training. The content of "is_good" was isolated and converted into a list, which will be used as the labels for the classification. The matrix and the list of labels were split into train and test sets. The train set will be used for model training and the test will be used for evaluating the accuracy of the model. The train set was used to train the two models, and we used 70% to 30% ratio or 50% to 50% ratio to do the train validation split. The model training was implemented with the train set, and we tried multiple epochs to analyze the accuracy and loss of both train and validation sets. The test set was used to do the prediction by the model, and the labels of the test

set was used as ground truth data to evaluate the accuracy of the model. MAE has been used to evaluate the accuracy of the prediction.

Results

Exploratory Analysis

To better understand our dataset, we ran a series of exploratory analysis. After our preliminary data cleaning, we acquired a dataset of 344,253 records based on the user reviews of restaurants in Pennsylvania. The average rating of all selected reviews is 3.71 (Table.1) and the average length of reviews is 114 words. Most reviews give the restaurants five-star ratings (Figure 3) while the least number of users give 2-star reviews. According to Figure 3, we found that positive reviews tend to be shorter than negative reviews with lower mean value of length and higher concentration of shorter reviews.

Figure 4-a and b demonstrate words that are most often used in good and bad reviews. Both types of reviews include keywords such as food and place. Positive reviews usually involve positive adjectives, however, negative reviews involve more verbs which indicates that negative reviews may more likely be given by users because of poor services.

Table 1. Descriptive statistics of users in PA

	stars	**useful**	**funny**	**cool**	**Text length**
count	344,253	344,253	344,253	344,253	344,253
mean	3.71	1.23	0.37	0.56	113.69
std	1.43	2.82	1.62	2.08	104.82

Figure 3. User rating counts in PA

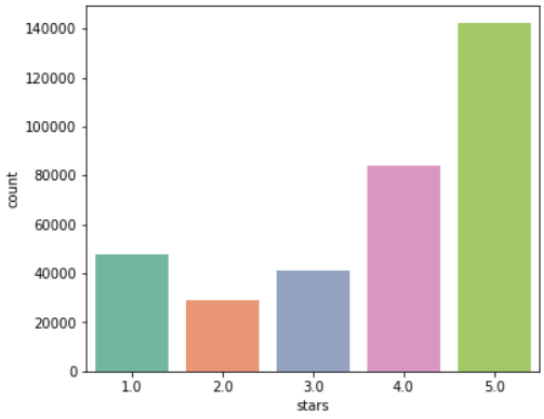


Figure 4. Text length by positive/negative reviews

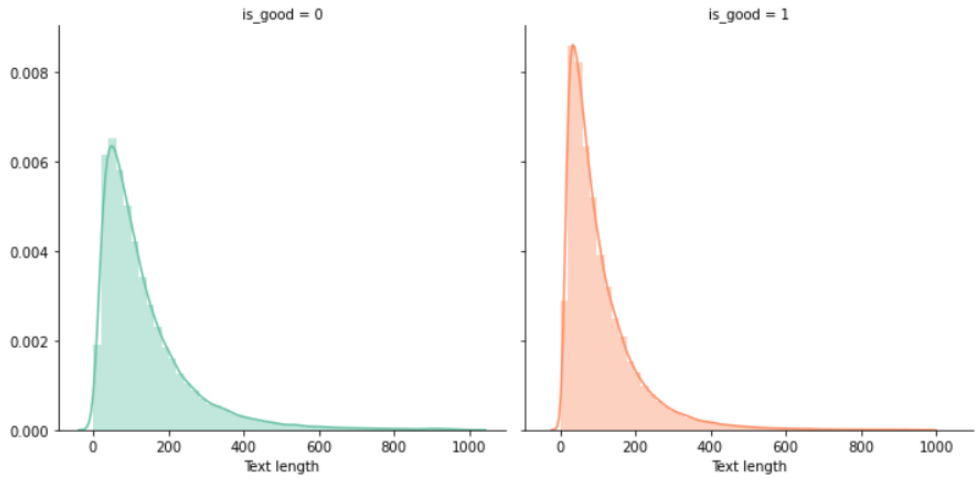


Figure 5-a Word cloud of good reviews

Figure 5-b Word cloud of bad reviews

Models

Baseline Model

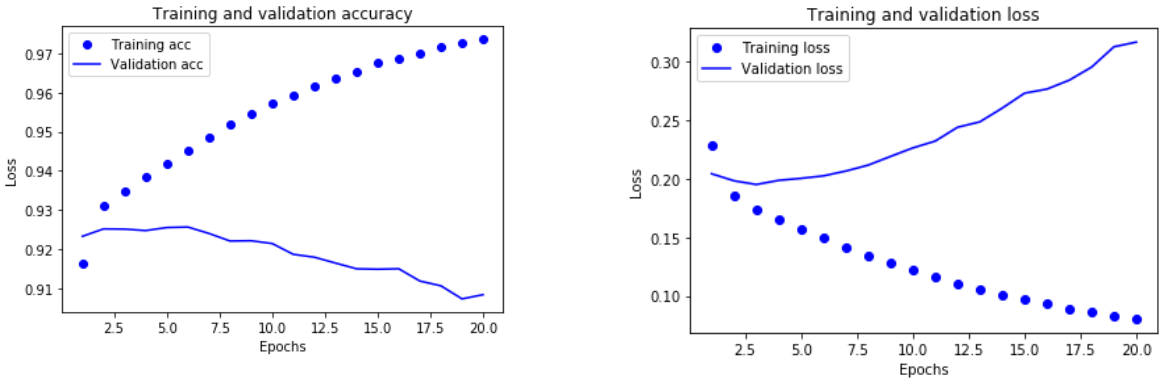
Multinomial naïve Bayes is a common and simple machine learning technique that has been applied to detect text contents and make classifications (Jurafsky & Martin,2019). For this model, we randomly split the tagged review data into 30% of test data and 70% of training data. The simple NB model achieved accuracy scores of 0.89 for the training data and 0.88 for the testing data with an MAE of 0.12.

Deep Learning Model 1

Deep Learning Model 1 just used a very basic network structure with two dense layers of the algorithm ‘relu’ and the last layer of the algorithm “sigmoid.” The review text was converted into a 10000 dimension matrix according to the most frequent words in the text. We carried out

20 epochs of the first deep learning model and achieved accuracy rates of the validation set between 0.923 and 0.908. The loss values range from 0.204 to 0.316. Based on the plot of each epoch (Figure 6), there will be an overfitting in the model from the 3 epochs on. The MAE of the predicted and test values is 0.093. This result indicates that the deep learning model is better than the regular machine learning model on predicting the text data.

Figure 6. Training and validation accuracy and loss of model 1



Deep Learning Model 2

Figure 8 shows the training and validation accuracy and losses of model 2. Model 2, 3 epochs has been carried out, the accuracy of the validation set is around 0.92, and the loss values are around 0.22. We can see that the accuracy and loose value have the best result at the second epoch. In the tuning parameter phase, we have found that the best number of epochs could be 2 or 3 and after 3 epochs, the overfitting could happen. So we have used the early stopping callback method to monitor the change of validation accuracy so that the highest validation loss will be found. We also change the value of patience so that some small fluctuations could be ignored. The running process is shown in Figure 7. The MAE between the predicted result and the ground truth result is 0.088, which means that the prediction is very accurate and model two is a better model than model one.

Figure 7. Training and validation accuracy and loss of model 2

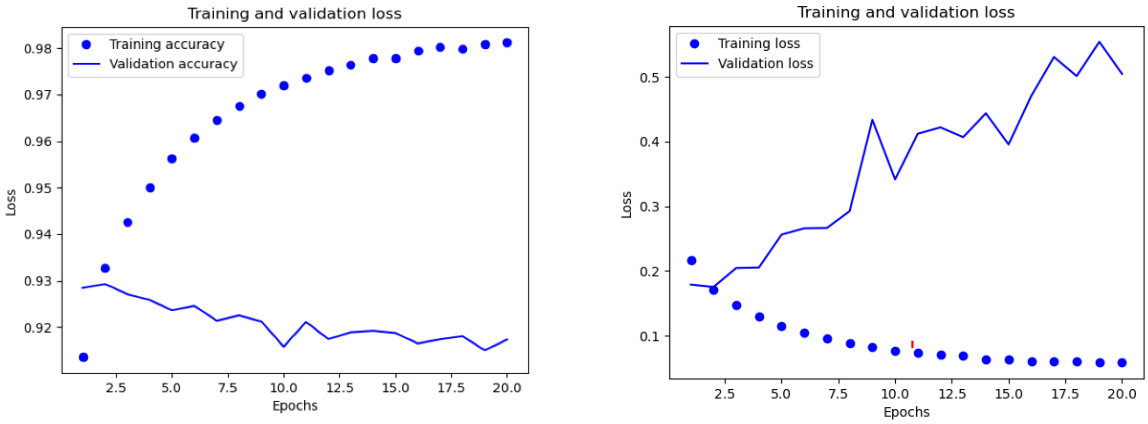


Figure 8. Model 2 processing

```
Epoch 1/20
7593/7593 [==============================] - 404s 53ms/step - loss: 0.2169 - accuracy: 0.9128 - val_loss: 0.1784 - va
l_accuracy: 0.9268
Epoch 2/20
7593/7593 [==============================] - 390s 51ms/step - loss: 0.1707 - accuracy: 0.9320 - val_loss: 0.1817 - va
l_accuracy: 0.9270
Epoch 3/20
7593/7593 [==============================] - 390s 51ms/step - loss: 0.1487 - accuracy: 0.9411 - val_loss: 0.1936 - va
l_accuracy: 0.9276
Epoch 4/20
7593/7593 [==============================] - 382s 50ms/step - loss: 0.1300 - accuracy: 0.9499 - val_loss: 0.2077 - va
l_accuracy: 0.9253
Epoch 5/20
7593/7593 [==============================] - 389s 51ms/step - loss: 0.1159 - accuracy: 0.9560 - val_loss: 0.2366 - va
l_accuracy: 0.9256
Epoch 6/20
7593/7593 [==============================] - 375s 49ms/step - loss: 0.1046 - accuracy: 0.9609 - val_loss: 0.2437 - va
l_accuracy: 0.9244
Epoch 7/20
7593/7593 [==============================] - 376s 50ms/step - loss: 0.0982 - accuracy: 0.9634 - val_loss: 0.2455 - va
l_accuracy: 0.9223
Epoch 8/20
7593/7593 [==============================] - 373s 49ms/step - loss: 0.0890 - accuracy: 0.9671 - val_loss: 0.3379 - va
l_accuracy: 0.9225
Epoch 9/20
7593/7593 [==============================] - 371s 49ms/step - loss: 0.0842 - accuracy: 0.9694 - val_loss: 0.3016 - va
l_accuracy: 0.9193
Epoch 10/20
7593/7593 [==============================] - 369s 49ms/step - loss: 0.0795 - accuracy: 0.9714 - val_loss: 0.3423 - va
l_accuracy: 0.9211
Epoch 11/20
7593/7593 [==============================] - 365s 48ms/step - loss: 0.0762 - accuracy: 0.9730 - val_loss: 0.3406 - va
l_accuracy: 0.9222
Epoch 12/20
7593/7593 [==============================] - 366s 48ms/step - loss: 0.0712 - accuracy: 0.9750 - val_loss: 0.3314 - va
l_accuracy: 0.9207
Epoch 13/20
7593/7593 [==============================] - 368s 48ms/step - loss: 0.0681 - accuracy: 0.9765 - val_loss: 0.3410 - va
l_accuracy: 0.9196
```

Related work

As discussed in the introduction section, there are several related works that inspired us to explore this topic and help us identify the effective model. Ranard et.al.(2016) used NLP methods to analyze Yelp reviews and found latent topics of hospital service that were not investigated by traditional surveys. Jurafsky & Martin (2019) provided guidance of using multinomial naive Bayes models to conduct NLP tasks. LSTM is another classic approach to conduct sentiment classification analysis (Hochreiter and Schmidhuber, 1997) and has been increasingly used in the NLP. Jelodar et. al. presented a systematic framework based on NLP that

can extract meaningful topics from COVID-19 related comments on Reddit and proposed a LSTM based deep learning model to conduct sentiment classification of the reddits. LSTM also holds great potential for text-classification tasks. Zhou et. al.(2016) has established that integrating bidirectional LSTM with two-dimensional max pooling can significantly improve text-classification accuracy. In addition, LSTM can also be integrated with traditional convolutional neural networks for sentence representation and text classification (Zhou et. al., 2015).

Discussion and Conclusion

By this study, we have learned and practiced series approaches of deep learning to predict the latent semantic information of text data. With the trial of different combinations of deep learning algorithms and the adjustment of different hyper parameters, we found LSTM is the most effective model to predict the user rating classes. It can provide fairly accurate classification of the review data with a MAE of about 0.088 or less, which significantly outperforms the conventional multinomial naive Bayes classifier. It is obvious that the seemingly orderless user reviews actually contain hidden patterns that help us detect the attitudes of the reviewers.

By doing this project, we learned the techniques of deep learning for classification and prediction on text data. It is a good chance to practice this method with a real dataset. During this process, we also got to know that deep learning is a more complex approach than regular machine learning algorithms, it needs more adjustment and tests to get the best model. Deep learning is powerful but can easily lead to overfitting, so we need to be careful about the processes and try to find the best parameters that can result in highest accuracy and lowest loss. Additionally, it is also an interesting thing to find hidden patterns and tendencies of text data. We would like to further explore natural language processing and deep learning schemes in the future.

Future work

This project only provided a coarse classification of the user reviews with a subset of yelp data for the time and computing power constraint. The future work could build on this project by expanding the dataset to include more categories of local business reviews, more states, or finer classification of the ratings. As mentioned earlier, the algorithm can potentially be applied

elsewhere to determine the sentiment of people reviewing local businesses. In the future, we could scrape twitter or reddit data regarding local businesses and then test the current algorithm towards the reviews that don't provide a direct rating.

References

Huang, J., Rogers, S., & Joo, E. (2014). Improving Restaurants by Extracting Subtopics from Yelp Reviews. https://www.ideals.illinois.edu/handle/2142/48832

Jelodar, H., Wang, Y., Orji, R., & Huang, H. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. ArXiv:2004.11695 [Cs]. http://arxiv.org/abs/2004.11695

Le, B., & Nguyen, H. (2015). Twitter Sentiment Analysis Using Machine Learning Techniques. In H. A. Le Thi, N. T. Nguyen, & T. V. Do (Eds.), Advanced Computational Methods for Knowledge Engineering (pp. 279–289). Springer International Publishing. https://doi.org/10.1007/978-3-319-17996-4_25

Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.Com (SSRN Scholarly Paper ID 1928601). Social Science Research Network. https://doi.org/10.2139/ssrn.1928601

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., & Merchant, R. M. (2016). Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. Health Affairs, 35(4), 697–705. https://doi.org/10.1377/hlthaff.2015.1030

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long ¨ short-term memory. Neural computation, 9(8):1735– 1780.

Socher, R., Bengio, Y., & Manning, C. D. (2012). Deep learning for NLP (without magic). Tutorial Abstracts of ACL 2012, 5.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 606–615. https://doi.org/10.18653/v1/D16-1058

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. ArXiv:1611.06639 [Cs]. http://arxiv.org/abs/1611.06639

Yelp. 2020. Shareholder letter Q2 2019. Retrieve from: https://s24.q4cdn.com/521204325/files/doc_financials/quarterly/2019/q2/ShareholderLetter_Q2_2019.pdf

Chollet, François, deep-learning-with-python-notebooks. (2017). Github Repository. fchollet/deep-learning-with-python-notebooks: Jupyter notebooks for the code samples of the book "Deep Learning with Python" (github.com)

Technical references:

Gabor Melis, Chris Dyer, and Phil Blunsom. On the State of the Art of Evaluation in Neural Language Models. arXiv:1707.05589v2 [cs.CL] 20 Nov 2017. URL [1707.05589] On the State of the Art of Evaluation in Neural Language Models (arxiv.org)

sklearn.feature_extraction.text.CountVectorizer:https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

sklearn.naive_bayes.MultinomialNB:https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB