NLP project: prediction of like probability of quotes

Darya Kryzhanovskaya

May 2024

Аннотация

In this article, we explore the application of machine learning algorithms for analyzing and predicting the probability of like by posts of MIPT professors' quotes in the popular Russian social network VK in social community `https://vk.com/prepod_mipt`.

Code: `https://github.com/kryzhanovskaya/like_quotes_probability`

1 Introduction

The ability to predict user engagement on social media platforms is a crucial aspect for content creators, marketers, and community managers. It enables them to anticipate how well their content will perform, allowing for better planning, targeted content creation, and optimized posting schedules.

The specific problem tackled in this research is crucial due to several reasons:

1. Educational Impact: By engaging students with relatable and insightful content, educational institutions can enhance learning experiences and create a supportive academic community.

2. Content Strategy Optimization: Predictive insights allow for the strategic timing and theming of posts, which can lead to increased interaction and reach within the community.

3. Community Engagement: Active engagement in academic communities can help maintain and boost student morale and motivation, particularly important in remote or hybrid learning setups.

1.1 Team

Darya Kryzhanovskaya

2 Related Work

In my research aimed at predicting the popularity of quotes on the social network VK, I used the BERT (Bidirectional Encoder Representations from Transformers), which was introduced in the paper [Devlin et al., 2019]. The BERT model represents an innovative approach in the field of pre-training deep bidirectional transformers for natural language understanding, making it ideally suitable for text analysis tasks.

3 Model Description

My model for predicting the likelihood of a quote receiving a like processes data as follows.

3.1 Preliminary text processing

The original text of the quote is cleansed of tags. This is done to ensure the text is as "clean"as possible and suitable for further analysis.

3.2 Using the BERT model

The cleaned text is fed into the BERT model. This model, trained on a large dataset of texts, returns vector representations (features) of the text. These representations contain a deep semantic and syntactic understanding of the text, making them particularly valuable for the task of text analysis.

3.3 Tag processing

Tags initially removed from the text are now processed separately, being converted into categorical variables. This process involves assigning a unique number to popular tags, allowing such data to be used in numerical form.

3.4 Concatenation of features

The vector features obtained from the BERT model are concatenated with the vectorized categorical features of the tags. This combination creates a single feature vector that includes both a deep understanding of the text and information about the tags associated with the text.

3.5 Neural network for predicting the likelihood of a like

The formed feature vector is fed into a neural network consisting of two linear layers. This network is trained to predict the ratio of likes to the number of views for the quote.

Thus, my model integrates a deep understanding of the text obtained through the modern NLP model BERT, with contextual data in the form of tags converted into categorical features. This allows for a more accurate prediction of user reactions to the published content.

4 Dataset

To collect the dataset I utilized the VK API. The data collection process was organized as follows:

I used specific API methods designed for extracting information from VK communities. These methods enable the retrieval of lists of all posts published in a specific community or group. In addition to the text of the posts themselves, I also collected meta-information for each entry. Important data included the number of likes and views. After extracting the data, it was organized and saved in CSV format. During the data collection process, I also considered possible limitations and the usage policy of the VK API, such as limits on the number of requests per unit of time, to avoid access blocking.

Train size: 2688 objects, Test size: 672 objects.

5 Experiments

5.1 Metrics

I used the CTR (click-through rate) as the target, which is the ratio of the number of likes to views. The metric was MSE.

5.2 Experiment Setup

Data Splitting: The dataset was randomly split into training and testing sets. This random selection ensures that the data split offers a representative sample of the general dataset for both training and evaluation, preventing any bias that could affect the model's performance on unseen data.

Model Architecture: The regression model is based on the "DeepPavlov/rubert-base-cased"configuration. This model is a variant of the BERT model pre-trained on Russian language texts and is particularly suited for Russian text data such as that from VK.

Model Hyperparameters:

- Learning Rate: Set to 1e-5, a rather low rate, which helps in fine-tuning the pre-trained model without causing rapid changes in the weights that could disrupt the pre-learned representations.

- Number of Epochs: Set to 100 to allow sufficient iterations over the dataset for the model to converge.

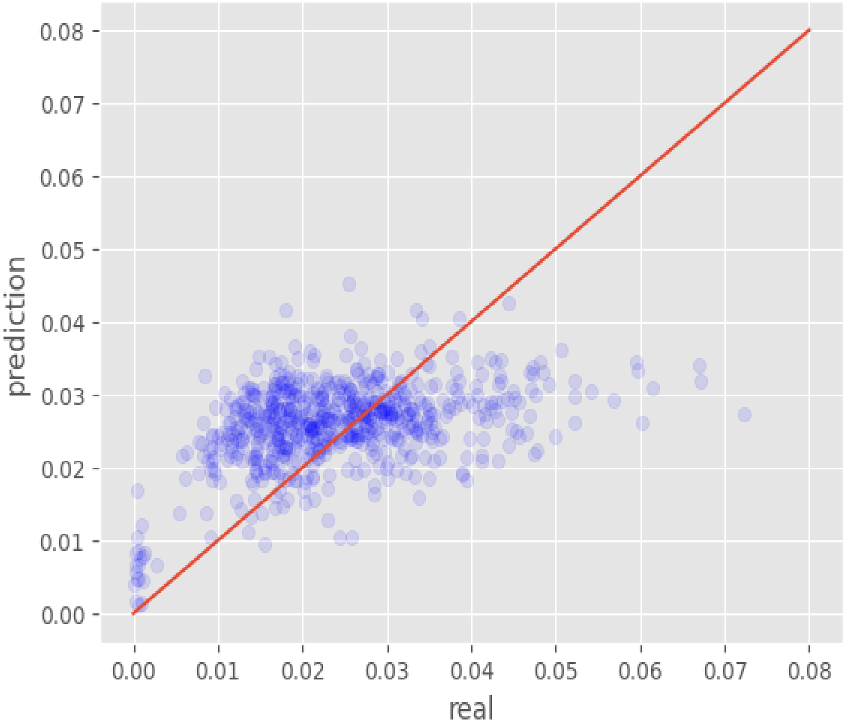
- Weight Decay: Set at 1e-5 to regularize the model and prevent overfitting.

5.3 Baselines

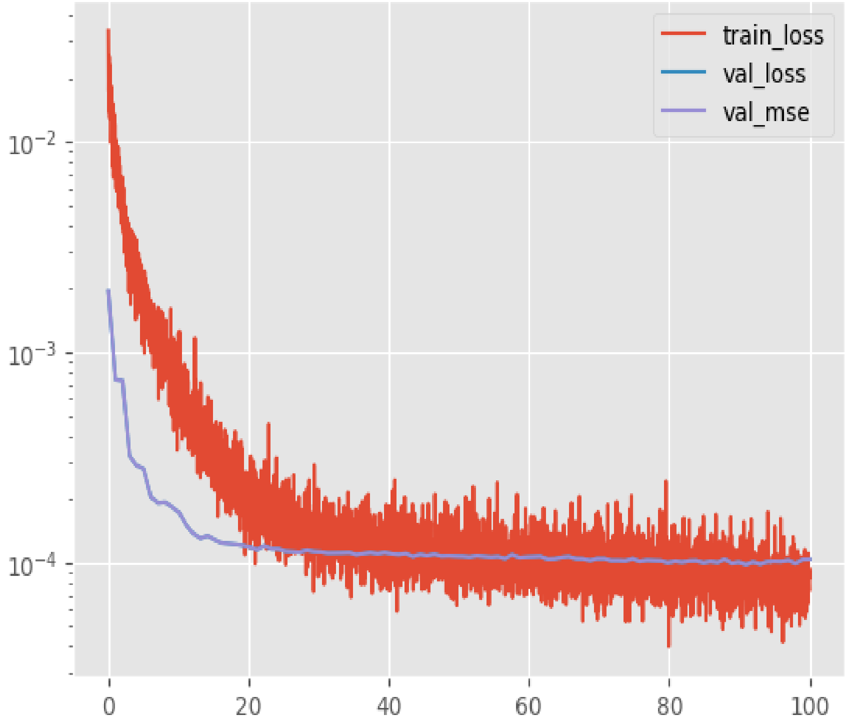
For the baseline I used Linear Regression over TF-IDF. MSE was 0.000178.

6 Results

On the picture below is presented a scatter plot of the result we achieved on test. MSE was 0.0001037, against 0.000178 which we had on baseline.



Кривая обучения:



6.1 Examples

Promo post Below is presented the text. The real ctr: 0.00039; prediction: 0.0039

```
'Знаешь всё про Big Data?\n
С 24 по 26 июня SENSE Group проведёт
онлайн-хакатон DATA HACK, а ГК «Иннотех»
выступит партнёром битвы IT-умов!\n\nВыполни
задание одного из трёх кейсов хакатона и
получи 100 000 рублей!\n\n Даты хакатона:
24-26 июня 2022 года\n Дедлайн регистрации:
22 июня 23:59\n Регистрация:
https://bit.ly/3H5e0Ta\n\n
Одной из задач хакатона станет
разработка статического анализатора Spark
SQL-кода. Также среди испытаний:\n-
разработка генератора фейковых данных
для сложных запросов;\n- создание
прототипа ETL-движка из Postgres,
Oracle, ClickHouse в HDFS на Spark,
который будет шаблонизирован через
конфигурацию.\n\nПризовой фонд -
300 000 рублей!\n\nПодробности и
регистрация: https://bit.ly/3H5e0Ta’
```

Short quote The real ctr: 0.03295; prediction: 0.02276

```
’Рациональные числа - они убогие\n#Давтян_mipt’
```

Yet another quote The real ctr: 0.03367; prediction: 0.02235

```
’Как у собак Павлова выделялась слюна при виде
горящей лампочки, так у вас при виде квадратного
трехчлена должен выделяться полный квадрат.\n
#Чубаров_mipt’
```

Message from the commumity administrator The real ctr: 0.0192; prediction: 0.0133

```
’Нашему цитатнику уже больше года.\nКогда мы
только хотели его сделать, многие считали,
что он вообще не нужен.\nНо сейчас, кажется,
он у многих вызывает интерес.\nНадеемся, что
будем и дальше оперативно постить цитаты преподавателей.’
```

7 Conclusion

In conclusion, it is noteworthy that there has been a significant improvement in prediction accuracy compared to the baseline model. However, it should be considered that the outcomes of the predictions proved to be quite noisy, which may be associated with the high variance of the target variable values. This indicates the variability of the audience’s interest in different posts. Additionally, it was observed that a substantial number of quotes receive a low predicted level of likes, which likely may be indicative of promotional posts. This aspect suggests the need for further analysis and possible adjustments to the model to enhance its ability to recognize such features in the data.

Список литературы

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.