Domain adaptation of the transformer model for the film industry

Vsevolod Vlasov, Yuriy Kim

May 2023

Abstract

This report explores the domain adaptation of transformer models in the film industry. In this work, we collect film-industry-specific data and perform domain adaptation of model, comparing the results obtained with model without domain adaptation for the task of predicting the sentiment of movie reviews.
Project code: `https://github.com/Yuki-53/domain-adoptation-nlp`.
Our dataset: `https://www.kaggle.com/datasets/yuriykim/movie-reviews`.
Test dataset: `http://ai.stanford.edu/~amaas/data/sentiment/`.

1 Introduction

Just like in other areas, the use of natural language processing and machine learning has become crucial in the film industry. However, pretrained transformer models often fail to perform well in domain-specific tasks due to the lack of domain-specific data. Even such complex and large models as transformers, which have proven themselves in many areas of NLP, do not cope with the task well enough without proper training on domain-specific data.
So, in this work, we will perform transfer learning of the model for the task of predicting the sentiment of movie reviews.

1.1 Team

This project was carried out by Vsevolod Vlasov and Yuriy Kim. Areas of responsibility in the project:

Vsevolod Vlasov: research, dataset, report writing
Yuriy Kim: research, modeling, report writing

2 Related Work

There are many works describing domain adaptation for other areas.

[von Boguszewski et al., 2021] This research explores methods for detecting hate speech in movies. The lack of domain-specific data often hinders the performance of pre-trained transformer models, so the researchers use a small amount of domain-specific data to improve the model performance. The results demonstrate that transfer learning from the social media domain is effective in identifying hate and offensive speech in movies through subtitles.

[Araci, 2019] - domain adaptation in the financial sphere. Results of this research show improvement in metric for two financial sentiment analysis datasets. Even with a smaller training set and fine-tuning only a part of the model, FinBERT outperforms state-of-the-art machine learning methods.

[Barbieri et al., 2020] - set of baselines trained on such domain-specific data. This paper proposed a new evaluation framework consisting of seven heterogeneous Twitter-specific classification tasks.

3 Model Description

Base model is a fine-tune checkpoint of DistilBERT, fine-tuned on SST-2. This model was further trained on the task of classifying the sentiment of movie reviews into 2 classes. In this case, a self-assembled dataset was used. You can read more about it in section 4.

Further, the weights of all layers were frozen and the last layer was replaced, since the test dataset is divided into 2 classes. Fine tuning took place on the training data of the test dataset and the results obtained on the test were compared with the results of the fine tuned model, but without domain adaptation.

4 Dataset

Film reviews play a crucial role in understanding audience opinions and sentiments towards movies. In this study, we present a large-scale dataset of film reviews sourced from kinopoisk.ru, a prominent Russian film review platform. The dataset provides valuable insights into user sentiments and preferences, enabling various research applications, such as sentiment analysis, recommendation systems, and opinion mining.

4.1 Dataset Collection

The dataset was collected over a period of approximately three weeks using a python srcipt running on virtual private server (VPS). A browser instance was launched, and requests to visit specific review pages were sent using the selenium library. This approach allowed automated browsing and data retrieval from the Kinopoisk website. The review data, in the form of raw HTML files, were then parsed using the Beautiful Soup library, and the relevant information was extracted and saved as CSV files.

4.2 Dataset Description

The Kinopoisk Movie Reviews Dataset consists of the following attributes for each review:

Column Name	Desctiption
Film ID	The unique identifier of the movie to which the review corresponds
Film Title	The name of the movie
Review ID	The unique identifier of the review
Author ID	The unique identifier of the review author
Author Username	The username of the review author
Review Title	The title of the review
Review Type	The sentiment type of the review (Positive, Neutral, or Negative)
Likes Count	The number of likes received by the review
Dislikes Count	The number of dislikes received by the review
Review Text	The textual content of the review
Review Date	The date when the review was written

Table 1: Kinoposik Movie Reviews Dataset Structure.

4.3 Preprocessing Steps

To ensure the quality and consistency of the dataset, the raw reviews underwent several preprocessing steps:

1. Removal of Newline Characters: Any leading or trailing newline characters (\n) were removed from each review.

2. Hyperlink Removal: All hyperlinks present in the reviews were eliminated using regular expressions.

3. HTML Tag Removal: HTML tags were stripped from the review text using the Beautiful Soup library.

4. Punctuation Restoration: Punctuation marks between sentences were restored to enhance readability and grammatical correctness.

5. Unicode to ASCII Conversion: Unicode characters such as emojis, diacritical marks, hieroglyphs, mathematical symbols, and other special characters were converted to ASCII encoding using the "unidecode" library.

As an additional step, the first 5000 characters of each review were translated into English using the googletrans library. This translation process enhances the accessibility and usability of the dataset for English-speaking researchers or those requiring English text for analysis.

4.4 Dataset Statistics

The Kinopoisk Movie Reviews Dataset encompasses a significant number of reviews, movies, and authors, providing a rich resource for analysis and research. Here are detailed statistics about the dataset:

Statistics Name	Statistics Value
Total reviews	793,763
Total movies with reviews	71,420
Total authors	133,435
Positive reviews count	526,170
Negative reviews count	146,799
Neutral reviews count	120,794
Total likes count	16,482,530
Total dislikes count	11,633,218
Total characters count	1,855,710,934
Total words count	276,090,972

Table 2: Statistics of our reviews dataset.

The first review was written on July 4, 2004, the last review in the dataset is dated March 31, 2023. The distribution of reviews by time is located below:

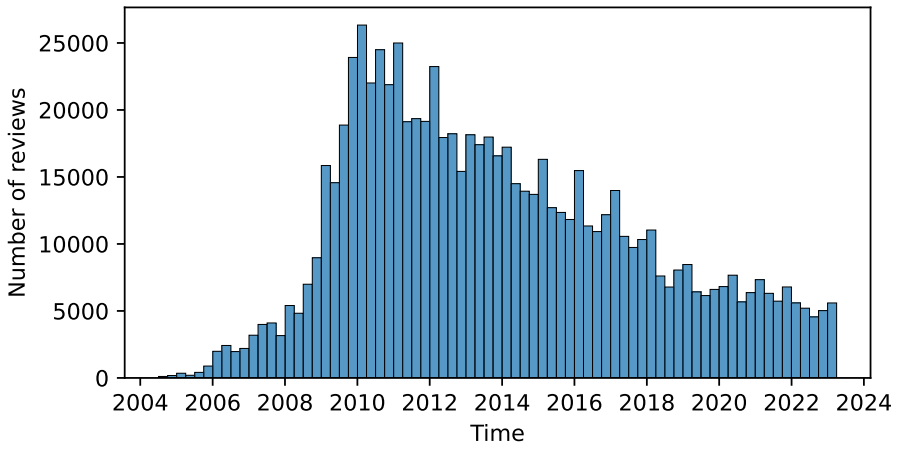

Figure 1: Distribution of reviews by time.

Understanding the distribution of words and characters in a dataset is crucial for analyzing and modeling textual data effectively. This section presents an in-depth examination of the distribution of words and characters in the collected movie reviews dataset. First, let's look at the distribution of characters:

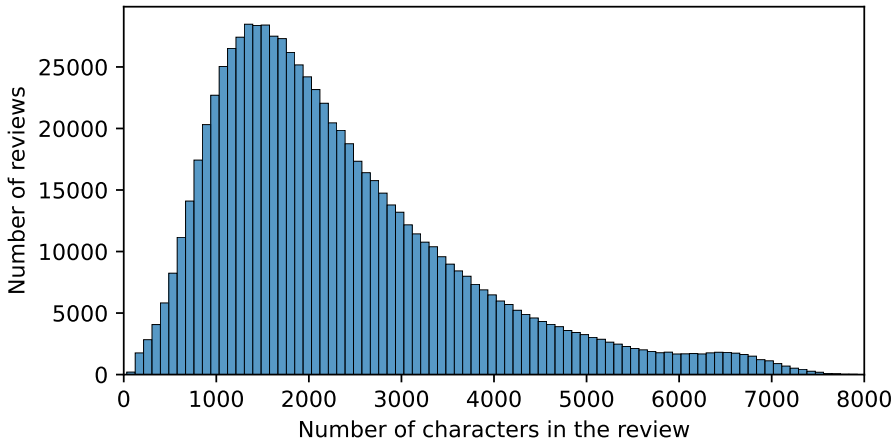

Figure 2: Distribution of the number of reviews by the number of characters

The distribution of characters in the dataset follows a pattern resembling a normal distribution. From the center towards the right tail, the frequency of characters gradually decreases. This indicates that the majority of reviews have a moderate length, with fewer reviews being significantly longer. Now let's look at the distribution of words:

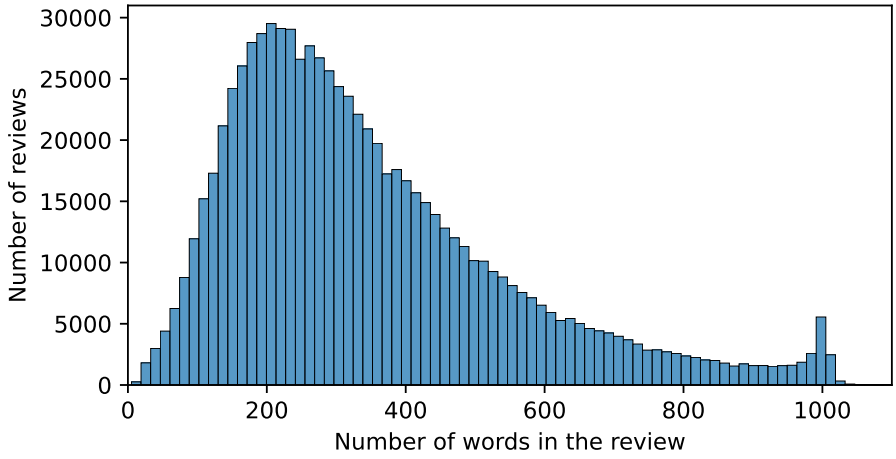

Figure 3: Distribution of the number of reviews by the number of words

Similar to the distribution of characters, the distribution of words in the dataset also exhibits characteristics of a normal distribution. However, there is a notable spike in the number of reviews with approximately 1000 words. This suggests that a substantial number of reviews in the dataset are relatively

lengthy, potentially containing more detailed opinions and analysis.

Figure 4: Word cloud for our dataset. Stop words were removed and word lemmatiztion are done by pymorphy2 package.

4.5 Test dataset

The test dataset is binary sentiment classification dataset containing substantially more data than previous benchmark datasets. Dataset provide a set of 50000 highly polar movie reviews from IMDB for training and testing. The train/val/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order. The sentiment labels are:

0 - negative (25000 samples)
1 - positive (25000 samples)

4.6 Conclusion

The Kinopoisk Movie Reviews Dataset offers a substantial collection of user reviews from the kinopoisk.ru website, providing valuable insights into user sentiments towards various movies. The dataset's comprehensive structure, coupled with the applied preprocessing techniques, ensures its usability for diverse research purposes, including sentiment analysis, opinion mining, and text classification tasks. Researchers and practitioners can leverage this dataset to develop and evaluate innovative algorithms and models in the field of natural language processing.

5 Experiments

5.1 Metrics

Accuracy was used as a metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

5.2 Experiment Setup

Parameters for domain-adoptation: 120680 negative reviews and random sampling 120680 positive reviews, sklearn train/val split with seed 42, test size = 0.2, 5 epoch, batch size - 16, AdamW optimizer, learning rate = 1e-6

Parameters for fine-tuning: sklearn stratify train/val/test split with seed 42, validation size = 0.14(7000 samples), test size = 0.3(15000 samples), 5 epoch, batch size - 16, AdamW optimizer, learning rate = 5e-6

5.3 Baselines

Since the test dataset was taken from the kaggle platform, there are many baselines there:

Solution using bag of words and tfidf with Naive Bayes classfier. Accuracy 75% both

Solution using decision trees with bag of words representation. Accuracy - 66%

6 Results

The results showed that the domain adapted model is better at classifying specific reviews after finetuning than the model without domain adaptation. The

	With domain adoptation	Without domain adoptation
Accuracy	82%	78%

Table 3: Metrics of models with and without domain adaptation.

difference in accuracy is visible even with not the most careful selection of hyperparameters.

7 Conclusion

In the course of the work, a large dataset of movie reviews was collected, with the help of which domain adaptation of the transformer model was carried out, the results of the fine-tuned model with and without domain adaptation were

compared, and the hypothesis of quality improvement after domain adaptation was confirmed

References

[Araci, 2019] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

[Barbieri et al., 2020] Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification.

[von Boguszewski et al., 2021] von Boguszewski, N., Moin, S., Bhowmick, A., Yimam, S. M., and Biemann, C. (2021). How hateful are movies? a study and prediction on movie subtitles.