Interpretation of the Results of Additional Training of Language Models (Transformers)

Tuesta Llaja Juan Carlos

May 20, 2025

Motivation

- Classifying literary texts in Russian is a challenge in NLP.
- Transformer models such as BERT and GPT have revolutionized natural language processing.
- Evaluating how data modification affects model performance.

Research Objectives

- Classify Russian literary texts into two categories: "definition" and "description."
- Evaluate the impact of additional training on the performance of Transformer models.
- Explore strategies to improve learning through preprocessing techniques.

Problem Statement

Binary Supervised Classification

We analyze the performance of Transformer models in the classification task of Russian literary texts.

- **Input**: Tokenized text datasets.
- **Output**: Binary label assigned by the model.
- **Models Evaluated**: RuBERT, BERT-multilingual, OPT/Facebook, RuGPT3, among others.
- **Evaluation Metrics**: F1-score, precision, recall, and confusion matrices.

Evaluated Models

Model	**Name**	**Type**
DeepPavlov/rubert-base-cased	RuBERT	BERT (Russian)
bert-base-multilingual-cased	BERT-multilingual	BERT (Multilingual)
distilbert-base-multilingual-cased	DistilBERT	Lightweight BERT (Multilin
roberta-base	RoBERTa	BERT (English)
xlnet-base-cased	XLNet	BERT (Multilingual)
xlm-roberta-base	XLM-RoBERTa	BERT (Multilingual)
gpt2	GPT-2	GPT
facebook/opt-125m	OPT-125M	GPT
sberbank-ai/rugpt3small based on gpt2	RuGPT3Small	GPT (Russian)

Table: Evaluated models for text classification in Russian

Data Modifications

[] Кустарник до 1.5 метров высотой. Ветви темно-коричневые. Годовалые побеги утолщенные, короткие, темно-бурые, при сушке чернеющие, серовато-шерстисто-опушенные. Почки мономорфные, 4.5-9 мм длиной, яйцевидные или округло-яйцевидные, со слегка уплощенным отогнутым носиком, со стороны побега немного уплощенные, по бокам килеватые, темно- или красновато-коричневые, шерстисто-опушенные или почти голые. Прилистники 1.5-12 мм длиной, полусердцевидные, по краю железистопильчатые, голые или опушенные. Черешки 0.6-2 см длиной, желобчатые, буроватые, сверху опушенные. Листья 2-9 см длиной и 1.5-5 см шириной, овальные, продолговатоовальные или продолговато-обратноланцетные, острые или короткозаостренные, в основании ширококлиновидные или округлые, пильчатые, с расставленными мелкими согнутыми зубцами или почти цельнокрайные, слегка подвернутые, плотные (жестковатые в сухом состоянии), от рассеянно-волосистых до более или менее густо опушенных, сверху ярко-зеленые, лоснящиеся, снизу сизые, со слабо выступающими боковыми жилками. Генеративные побеги до 2 см длиной, с 2-6 мелкими листьями; оси соцветий густо-волосисто-опушенные. Прицветники 1.6-3 мм длиной, обратнояйцевидные или продолговатые, на верхушке тупые или закругленные, бурые, с обеих сторон опушенные длинными волосками (опушение изнутри более густое). Мужские сережки 2-2.5 см длиной. Нектарники парные: адаксиальный двулопастный, 1-1.2 мм длиной, абаксиальный цельный, более короткий. Тычинок 2, со свободными, волосисто опушенными нитями около 5 мм длиной. Пыльники овальные, 0.7 мм длиной, желтые. Женские сережки при цветках 2-2.5 см длиной, при плодах 3-5 см длиной. Нектарники одиночные, 1-1.2 мм длиной, яйцевидные, цельные или частью двулопастные. Завязи войлочно-опушенные. Столбик 0.5-1.5 мм длиной, цельный или немного расщепленный. Рыльца 0.3-0.7 мм длиной, двулопастные. Плоды 6-8 мм длиной, яйцевидно-веретеновидные, светло-коричневые, рыхло войлочно опушенные, на опушенных ножках 0.5-1 мм длиной (рис. 1, смотри вклейку). [Бакланов] Клетки очень маленькие, 8-11 мкм длиной, 4-6 мкм шириной, обычно эллипсоидные, с апикальными щетинками и воротником вокруг проксимального конца жгутика. Панцирь состоит из 4 типов чешуек (рис. 4, 5). Чешуйки воротника с куполом и асимметричные 2-2.5 мкм длиной, 1.1-1.4 мкм шириной. Купол почти округлый, с коротким пиком, одна сторона его поверхности с ретикулом из угловатых ячей; каждая ячея окаймляет одну крупную пору. Щит чешуйки с папиллами, которые (совместно) с ребрами формируют ретикулум. В чешуйках воротника ретикулум занимает более половины поверхности щита и состоит из угловатых ячей, каждая ячея включает одну пору. Дорзальное субмаргиналыюе ребро хорошо развито и пересекается штрихами, между штрихами и в нижнем вентральном углу чешуйки располагаются поры. Вблизи субмаргинального ребра ретикулум отсутствует, и поверхность щита выглядит гладкой. Кромка узкая, ниже купола дистальный конец кромки образует плечо. Проксимальный край чешуйки асимметричный. Чешуйки тела клетки 1.8-2.3 мкм длиной, 1.2-1.4 мкм шириной, ориентированы поперек, без купола, ромбические и с закругленными углами; щит чешуйки с ретикулом из угловатых ячей только в центральной части. Многочисленные, регулярно расположенные поперечные штрихи отходят билатерально от передних субмаргинальных ребер. Обычно каждая пара штрихов включает одну пору. Проксимальный край широкий и гладкий, задняя кромка узкая. Транзитные (переходные) чешуйки с редуцированным ретикулум, 1.2-1.4 мкм длиной, 1.1-1.2. мкм шириной. Задние чешуйки асимметричные, с очень коротким шипом. Щетинки короткие, 3-4 мкм длиной. Стоматоцисты неизвестны.

Original dataset before modifications

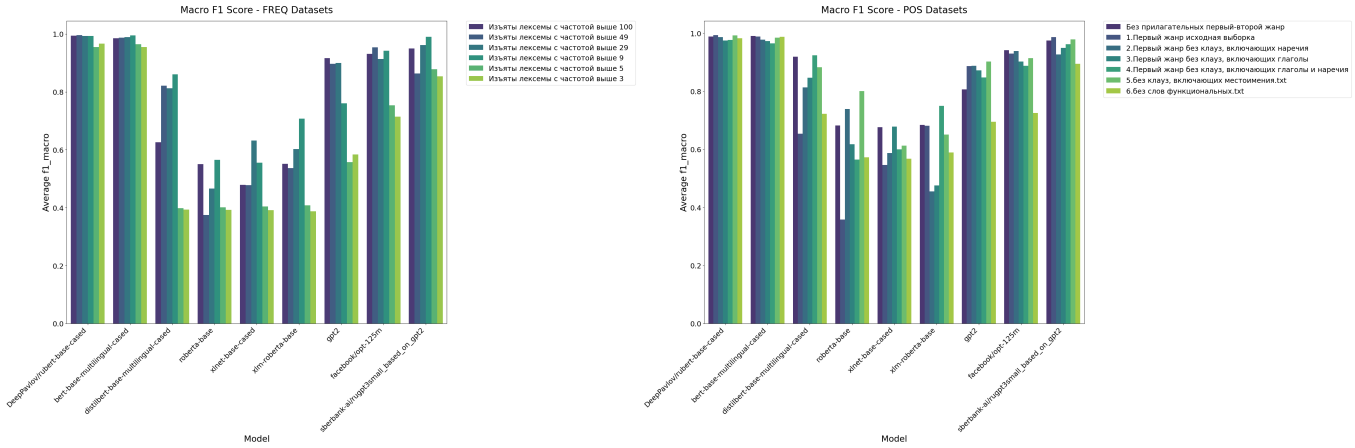
Data Modifications

[] Кустарник Ветви темно-коричневые. Годовалые побеги утолщенные, короткие, темно-бурые, при сушке чернеющие, серовато-шерстисто-опушенные. Почки мономорфные, яйцевидные или округло-яйцевидные, по бокам килеватые, темно- или красновато-коричневые, шерстисто-опушенные. Прилистники полусердцевидные, по краю железистопильчатые, голые или опушенные. Черешки желобчатые, буроватые. Листья овальные, продолговатоовальные или продолговато-обратноланцетные, острые или короткозаостренные, в основании ширококлиновидные или округлые, пильчатые, с расставленными мелкими согнутыми зубцами, плотные (жестковатые в сухом состоянии). Генеративные побеги с 2-6 мелкими листьями; оси соцветий густо-волосисто-опушенные. Прицветники обратнояйцевидные или продолговатые, на верхушке тупые или закругленные, бурые, с обеих сторон опушенные длинными волосками. Мужские сережки. Нектарники парные: адаксиальный двулопастный, абаксиальный цельный. Тычинки со свободными, волосисто опушенными нитями. Пыльники овальные, желтые. Женские сережки при цветках, при плодах. Нектарники одиночные, яйцевидные, цельные или частью двулопастные. Завязи войлочно-опушенные. Столбик цельный или расщепленный. Рыльца двулопастные. Плоды яйцевидно-веретеновидные, светло-коричневые, на опушенных ножках. Клетки с апикальными щетинками и воротником вокруг проксимального конца жгутика. Чешуйки воротника с куполом и асимметричные. Купол с коротким пиком, одна сторона его поверхности с ретикулом из угловатых ячей; каждая ячея окаймляет одну крупную пору. Проксимальный край чешуйки асимметричный. Чешуйки тела клетки ориентированы поперек, без купола, ромбические и с закругленными углами; щит чешуйки с ретикулом из угловатых ячей только в центральной части. Проксимальный край широкий и гладкий, задняя кромка узкая. Транзитные (переходные) чешуйки с редуцированным ретикулум. Задние чешуйки асимметричные. Щетинки короткие. Стоматоцисты неизвестны.

Example of data modification (frequency filtering)

- Removal of words by frequency ($>$100, $>$49, $>$9, $>$3).
- Removal of parts of speech (verbs, adverbs, pronouns, function words).
- Asymmetric approach: Modifications in descriptions while definitions remain unchanged.

Experimental Results



F1-score performance across different models and datasets

- **Best performance** with word removal at frequency >9.
- **Lowest performance** when words with frequency >3 were removed.
- **Positive impact** observed when removing verbs in descriptions.
- **RuBERT, BERT-multilingual, OPT/Facebook, and RuGPT3** showed the highest classification accuracy.

Conclusions

- Strategic word removal improves text classification.
- RuBERT and BERT-multilingual outperform general models.
- Structural modifications influence classification capability.