Применение больших языковых моделей в стилометрии

Регина Р. Насырова
МГУ имени М. В. Ломоносова
Москва, Россия, regina.nasyrova55@gmail.com
Данил А. Алексеев
МГУ имени М. В. Ломоносова
Москва, Россия, dnl.alksv@gmail.com

1. Стилометрия — прикладная дисциплина, исследующая автоматическое определение различных характеристик текста, в том числе его авторства. Наиболее распространены в данной области статистические методы — в частности, Дельта Бёрроуза [Burrows 2003], которая на русском материале была использована для подтверждения авторства «Тихого Дона» [Великанова, Орехов 2019]. Применяются в стилометрии и нейросетевые методы, начиная с пионерских работ [Matthews, Merriam 1993, 1994].

В последнее время становится актуальным также использование больших языковых моделей (англ. *large language models*, сокр. *LLMs*) Насколько нам известно, на данный существует лишь одна работа-препринт [Patel et. al 2023], где описывается процесс создания стилометрической модели LISA на основе GPT-3 [Brown et al. 2020].

2. Целью настоящего исследования было установить эффективность применения классификаторов на основе предобученной языковой модели BERT [Devlin et al. 2018] в данной задаче. Базовой моделью был выбран наивный байесовский классификатор. Применение этого алгоритма в стилометрии показано, например, в [Howedi, Mohd 2014].

3. В качестве обучающих и тестировочных данных использовались произведения классической русской литературы, взятые из интернет-библиотеки М. Мошкова[1]. Из каждого текста были выбраны предложения длиной от 15 до 25 токенов включительно (см. Таблицу 1).

В обучающую и тестовую выборку не попали предложения из одних и тех же произведений, поэтому исключено корректное предсказание авторства, например, на основе имени, которое встречается в единственном произведении.

[1] lib.ru

Автор	Предложений
Н. В. Гоголь	2129
И. А. Гончаров	4213
Ф. М. Достоевский	3804
А. С. Пушкин	1595
Л. Н. Толстой	7043
И. С. Тургенев	1742
А. П. Чехов	2103

Таблица 1. Количество предложений, подобранных для каждого автора.

4. В экспериментах в качестве базовой модели использовался вариант наивного байесовского классификатора для полиномиально распределенных данных из библиотеки Scikit-learn [Pedregosa et al. 2011]. В (1) отражено, что согласно данной модели вероятность отнесения текста, состоящего из токенов $x_1,\ldots,x_n$ к некоторому классу y пропорциональна произведению вероятности класса на произведение условной вероятности каждого токена при данном классе.

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (1)$$

Модель BERT, выбранная для сравнения с базовой, имеет 12-слойную двунаправленную архитектуру с механизмом внимания и была предобучена на задачах предсказания пропущенного слова и определения следования двух предложений друг за другом, что даёт ей некоторые знания о частотности слов и их совместной встречаемости.

В рамках наших экспериментов мы дообучили для классификации авторства следующие модели на основе BERT: bert-base-cased, DeepPavlov/rubert-base-cased, ai-forever/ruRoberta-large (доступны на huggingface.co). Первая модель была предобучена на англоязычных данных, вторая и третья — на русскоязычных. Третья модель значительно превосходит вторую по количеству данных и параметров при обучении.

Тестирование базовой модели показало, что ее качество значительно ниже случайного, см. Таблицу 2; чем выше представленность автора в выборке, тем лучше модель справляется с его определением, ср. Таблицу 1.

	Чехов	Достоевский	Гоголь	Гончаров	Пушкин	Толстой	Тургенев
F1-мера	18.18	34.54	11.95	31.59	24.70	48.59	13.00
F1-макро	26.08						
Корректность (accuracy)	34.12						

Таблица 2. Результаты применения базовой модели к полному датасету.

Было решено протестировать все модели на выборке со всеми авторами (датасет *classics-large*) и на сокращённом датасете (*classics-small*), в который вошли предложения из произведений трёх наиболее представленных авторов — Достоевского, Гончарова и Толстого. Результаты представлены в Таблице 3.

	classics-large		*classics-small*	
	Корректность	F1-макро	Корректность	F1-макро
naive bayes	34.12	26.08	53.43	49.98
bert-base-cased	32.78	26.50	51.22	45.14
rubert-base-cased	38.45	31.49	61.82	57.05
ruRoberta-large	**44.77**	**36.53**	**62.12**	**57.81**

Таблица 3. Общие результаты экспериментов.

Результаты на *classics-small* оказались значительно выше, чем на *classics-large*, следовательно, большие языковые модели также чувствительны к представленности автора в выборке.

RuRoberta-large предсказуемо продемонстрировала лучшие результаты на обеих выборках, хотя можно заметить, что на *classics-small* метрики rubert-base-cased оказались почти так же высоки. Интересным кажется и то, что на сокращённом датасете метрики наивного байесовского классификатора оказались даже выше, чем у bert-base-cased, которая была предобучена на английском материале.

5. Таким образом, в задаче определения авторства текста результаты больших языковых моделей оказываются лучше наивного байесовского классификатора. Одним из дальнейших путей исследования применения больших языковых моделей в стилометрии может являться их тестирование на текстах других жанров.

Литература

Великанова, Орехов 2019 — Великанова Н.П., Орехов Б.В. Цифровая текстология: атрибуция текста на примере романа М.А. Шолохова «Тихий Дон». Мир Шолохова, 2019. № 1(11). С. 70-82.

Brown et. al 2020 — Brown T., Mann B., Ryder N., Subbiah M., Kaplan D. J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D.. 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. Curran Associates, Inc.

Burrows 2003 — Burrows J. F. Questions of authorship: attribution and beyond. Computers and the Humanities, 37: 5–32.

Devlin et al. 2018 — Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Howedi, Mohd 2014 — Howedi F., M. Mohd. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. Computer Engineering and Intelligent Systems, 5, 4: 48–56

Matthews, Merriam 1993 — Matthews R. A. J., Merriam T. V. N. "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher". *Literary and Linguistic Computing*. **8** (4): 203–209. doi:10.1093/llc/8.4.203.

Matthews, Merriam 1994 — Matthews R. A. J., Merriam T. V. N. "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe". Literary and Linguistic Computing. 9 (1): 1–6. doi:10.1093/llc/9.1.1.

Patel et. al 2023 — Patel A., Rao D., Calisson-Burch C. Learning Interpretable Style Embeddings via Prompting LLMs. https://arxiv.org/abs/2305.12696

Pedregosa, F. et al. 2011 — Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), pp.2825–2830.