КЛАССИФИКАЦИЯ ТЕКСТОВ С ПОМОЩЬЮ СЛОВАРЯ СИНОНИМОВ

Гилязова А.А.

Институт проблем управления им. В.А. Трапезникова РАН,
Россия, г. Москва, ул. Профсоюзная, д.65
giliazova@mail.ru

Аннотация: В данной работе приводятся результаты классификации 5063 текстов из блогов 68 авторов с помощью модели XLM Roberta Large Xnli (Facebook AI) и словаря русскоязычных синонимов.

Ключевые слова: XLM Roberta Large Xnli, синонимы, классификация текстов, блоги.

Введение

Задача классификации текстов важна для ряда практических приложений, таких как категоризация новостей, классификация научных статей и т.п. Помимо информации о текстах, она также иногда позволяет получить информацию об их авторах, например, об их предпочтениях или даже о средних школьных отметках [1].

Одной из относительно новых моделей для классификации текстов является XLM Roberta Large Xnli, относящаяся к моделям типа Transformer [2]. Нейронные сети типа Transformer используются в различных задачах, например, для идентификации информативных твитов о Covid-19 [3], для видео-поиска [4], разметки доменов [5], парсинга структуры предложений [6], многоязыковых задач [7].

В данной работе мы используем XLM Roberta Large Xnli для классификации текстов из блогов с помощью слов-синонимов.

1 Эксперимент

1.1 Модель

Мы используем модель XLM Roberta Large Xnli model (Facebook AI) из модуля transformers 3.1.0 для языка Python, доступного по адресу: https://huggingface.co/transformers/. Эта модель была обнародована одновременно со статьёй [2].

Это маскированная языковая модель типа Transformer, обученная на сотне языков, включая русский, используя более двух терабайтов фильтрованных данных CommonCrawl. Модель XLM-R существенно превосходит многоязычную модель BERT (mBERT) на нескольких кросс-языковых бенчмарках, включая +14.6% средней точности на XNLI, +13% среднего значения F1 на MLQA и +2.4% значения F1 на NER. [2] Мы используем режим нулевого выстрела («zero-shot classification»).

1.2 Данные

Мы используем словарь русских синонимов с 956 групп прилагательных, доступный по адресу: https://synonymonline.ru/download.html#workers (2011). Набор классифицируемых текстов состоит из 5063 русскоязычных статей, взятых из блогов 68 авторов, доступных по адресу: https://www.gazeta.ru/comments/column/.

Тексты были разделены на части не более 512 слов (ограничение модели), результаты по этим частям усреднялись.

Расчёты заняли около 19 дней с помощью двух GPU.

2 Результаты

2.1 Статистика по статьям

Ожидаемым поведением для модели было выдать оценочное значение в диапазоне от 0 до 1 для каждого слова, отображающее то, подходит ли это слово для описания текста или нет. Если модель не уверена, то должна выдать 0,5. Мы ожидаем, что большинство слов не подходят и должны получить оценки ниже 0,5, но некоторые слова подходят и должны получить оценки выше 0,5. Эти высокие оценки должны быть в статистике максимальных значений по статьям, а не средних значений по статьям.

Мы используем режим многоклассовой классификации, при котором каждое слово в группе синонимов оценивается независимо, так что сумма оценок в группе не обязана быть равна 1. В качестве результата мы рассматриваем среднее значение и среднеквадратическое отклонение оценок для групп синонимов. Мы анализируем результаты по статьям, по группам слов и в разрезе по авторам.

Для средних значений по статьям (Таблица 1, Рис. 1а) среднее значение составило 0,399, что ниже 0,5. Таким образом, модель сочла большинство слов неподходящими для описания этих текстов, что

было ожидаемо. Но максимальное среднее значение по статьям слишком высокое (0,987), что не было ожидаемым. Есть тексты, для которых большинство слов были сочтены подходящими.

Среднее среднеквадратическое отклонение для средних значений по статьям (Таблица 1, Рис. 1б) составило 0,045, что довольно низко. Таким образом, значения внутри текстов в среднем меняются мало.

Таблица 1. Средние значения и среднеквадратические отклонения по статьям

	Для средних значений		**Для среднеквадратических отклонений**	
	Среднее значение	**Среднеквадратическое отклонение**	**Среднее значение**	**Среднеквадратическое отклонение**
Число	5063	5063	5063	5063
Среднее	0,399231	0,045418	0,031107	0,020438
Ср.-кв. отклонение	0,179596	0,037502	0,024184	0,015204
Минимум	0,000683	0,000027	0,000024	0,000014
25%	0,274873	0,017915	0,013261	0,009099
50%	0,395614	0,031311	0,022090	0,015723
75%	0,524378	0,064758	0,043496	0,028668
Максимум	0,986844	0,276591	0,163177	0,116839

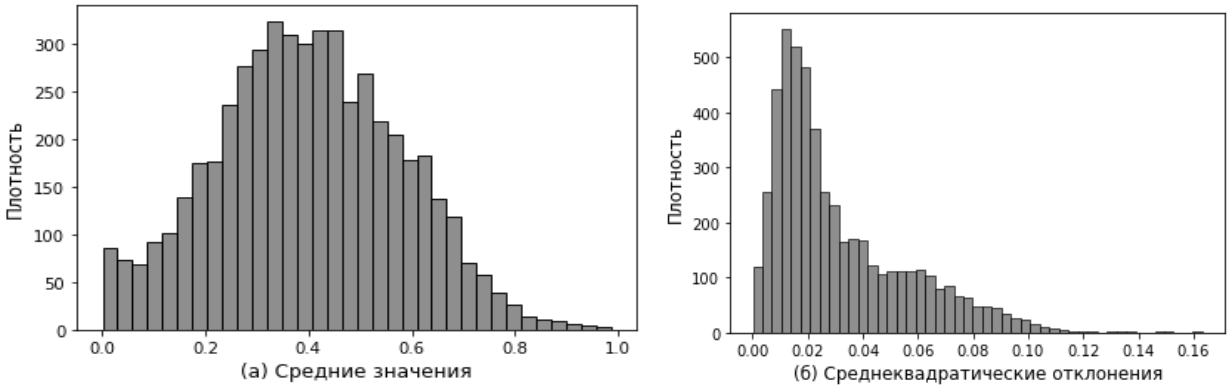

Рис. 1. Средние значения (а) и среднеквадратические отклонения (б) по статьям

Максимальные значения по статьям (Таблица 2, Рис. 2а) могут подниматься до 0,998, что весьма близко к полной уверенности, равной 1. Но среднее максимальное значение всего лишь 0,538, что означает, что модель не уверена насчёт этих групп слов. Таким образом, в среднем по статьям, модели не удалось найти подходящие группы синонимов.

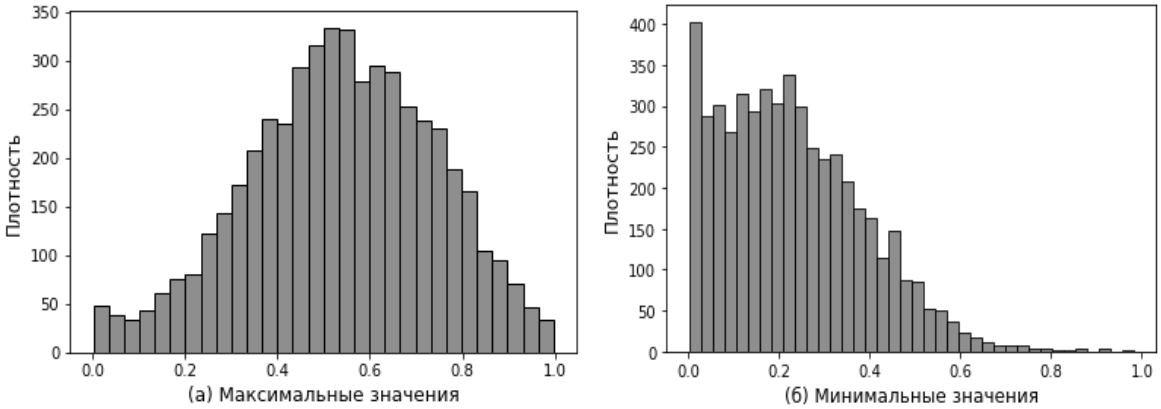

Рис. 2. Максимальные (а) и минимальные (б) значения по статьям

Таблица 2. Максимальные и минимальные значения по статьям

	Для средних значений		**Для среднеквадратических отклонений**	
	Максимальное значение	**Минимальное значение**	**Максимальное значение**	**Минимальное значение**
Число	5063	5063	5063	5063
Среднее	0,537594	0,228721	0,127739	9,302110e-04
Ср.-кв. отклонение	0,203723	0,156610	0,069445	8,452512e-04
Минимум	0,000844	0,000614	0,000132	2,281740e-08
25%	0,399851	0,104550	0,074271	3,235436e-04
50%	0,541491	0,210441	0,118505	7,001684e-04
75%	0,686555	0,330928	0,170576	1,289759e-03
Максимум	0,997850	0,982223	0,468500	8,065244e-03

Среднее минимальное значение для средних значений по группам синонимов (Таблица 2, Рис. 2б) составило 0,229, что ожидаемо, но максимальное значение для этих минимальных значений составило 0,982, что слишком высоко и означает, что в некоторых текстах все группы синонимов были сочтены хорошо подходящими данной моделью для описания этих текстов, что маловероятно.

Таблица 3. Относительные максимальные и минимальные значения по статьям

	Относительный максимум как (максимум – среднее) / среднее	**Относительный минимум как (среднее – минимум) / среднее**
Число	5063	5063
Среднее	0,490339	0,457556
Ср.-кв. отклонение	0,858571	0,226165
Минимум	0,003792	0,004682
25%	0,167724	0,278589
50%	0,315050	0,432373
75%	0,570654	0,612185
Максимум	32,953022	0,998097

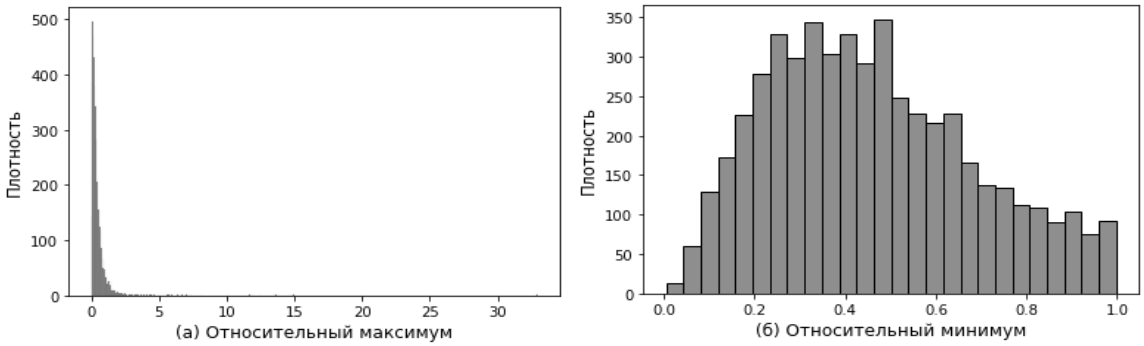

Рис. 3. Относительные максимумы (а) и относительные минимумы (б) по статьям

Под относительным максимумом мы имеем в виду разницу между максимальным и средним значением, поделенную на среднее значение. Под относительным минимумом мы имеем в виду разницу между средним значением и минимальным значением, поделенную на среднее значение. Мы ожидаем, что большинство слов не подходят для описания текстов и получают оценки около 0,5 или ниже, тогда как некоторые группы слов хорошо описывают тексты и получают высокие оценки. Таким образом, относительный максимум и относительный минимум должны быть высокими.

Значения относительных максимумов (Таблица 3, Рис. 3а) и относительных минимумов (Таблица 3, Рис. 3б) близки к 0,5 и не очень высоки.

2.2 Статистика по группам слов

Среднее значение для средних значений по группам слов (Таблица 4, Рис. 4а) составило 0,399, как и для по статьям, но максимальное значение ниже, 0,509. Это означает, что нет особых слов, часто получавших высокие оценки. Среднеквадратические отклонения средних значений (Таблица 4) существенно выше по группам слов, чем по статьям (в среднем 0,186).

Таблица 4. Средние значения и среднеквадратические отклонения по группам слов

	Для средних значений		**Для среднеквадратических отклонений**	
	Среднее значение	**Среднеквадратическое отклонение**	**Среднее значение**	**Среднеквадратическое отклонение**
Число	956	956	956	956
Среднее	0,399231	0,186204	0,031107	0,031734
Ср.-кв. отклонение	0,032022	0,004936	0,011874	0,009238
Минимум	0,254600	0,165225	0,005595	0,007076
25%	0,380141	0,183195	0,022284	0,025441
50%	0,400933	0,186024	0,030105	0,031794
75%	0,420179	0,188962	0,038371	0,037879
Максимум	0,509235	0,207051	0,078700	0,063595

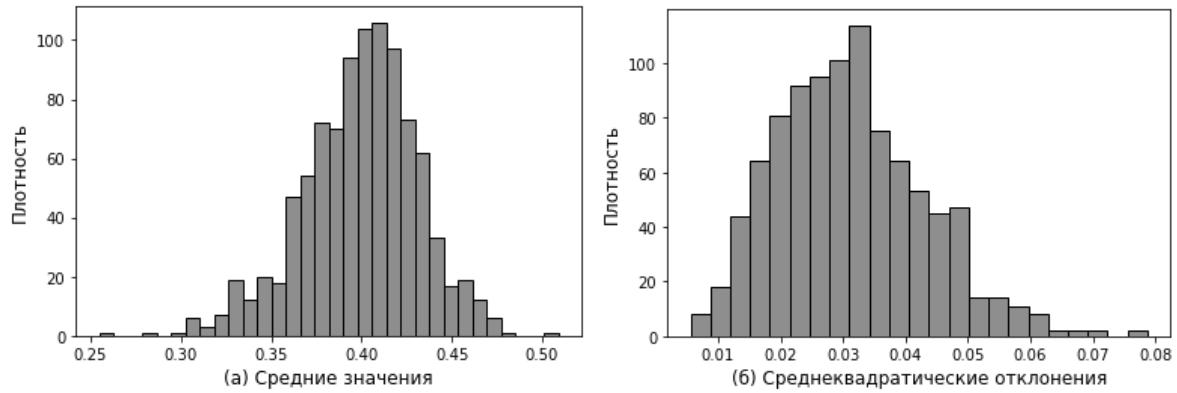

Рис. 4. Средние значения (а) и среднеквадратические отклонения (б) по группам слов

Все группы слов получили уверенные предсказания не ниже 0,982 хотя бы по одному разу (Таблица 5, Рис. 5а), что отличалось от статистики по статьям, где максимальное значение могло составлять 0,001 (Таблица 2).

Таблица 5. Максимальные и минимальные значения по группам слов

	Для средних значений		**Для среднеквадратических отклонений**	
	Максимальное значение	**Минимальное значение**	**Максимальное значение**	**Минимальное значение**
Число	956	956	956	956
Среднее	0,987058	0,000683	0,276598	1,557997e-05
Ср.-кв. отклонение	0,001460	0,000027	0,067417	8,473819e-06
Минимум	0,982223	0,000614	0,090944	2,281740e-08
25%	0,986399	0,000666	0,226242	8,912417e-06
50%	0,986935	0,000680	0,269737	1,556363e-05
75%	0,987506	0,000696	0,327452	2,149978e-05
Максимум	0,997850	0,000809	0,468500	4,591442e-05

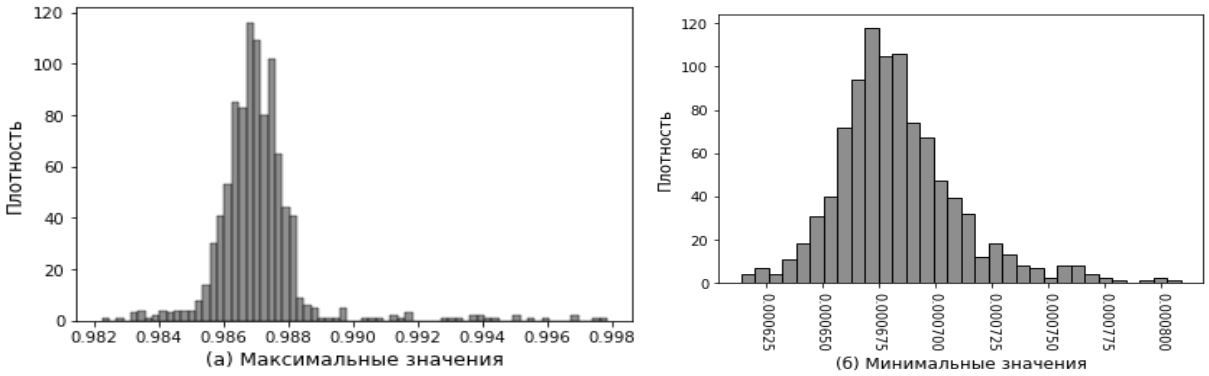

Рис. 5. Максимальные (а) и минимальные (б) значения по группам слов

Относительные максимумы (Таблица 6, Рис. 6а) и относительные минимумы (Таблица 6, Рис. 6б) значительно выше для групп слов, чем для статей (Таблица 3).

Таблица 6. Относительные максимальные и минимальные значения по группам слов

	Относительный максимум как (максимум – среднее) / среднее	**Относительный минимум как (среднее – минимум) / среднее**
Число	956	956
Среднее	1,489149	0,998279
Ср.-кв. отклонение	0,211205	0,000147
Минимум	0,957346	0,997561
25%	1,348424	0,998210
50%	1,460467	0,998310
75%	1,595324	0,998378
Максимум	2,872750	0,998579

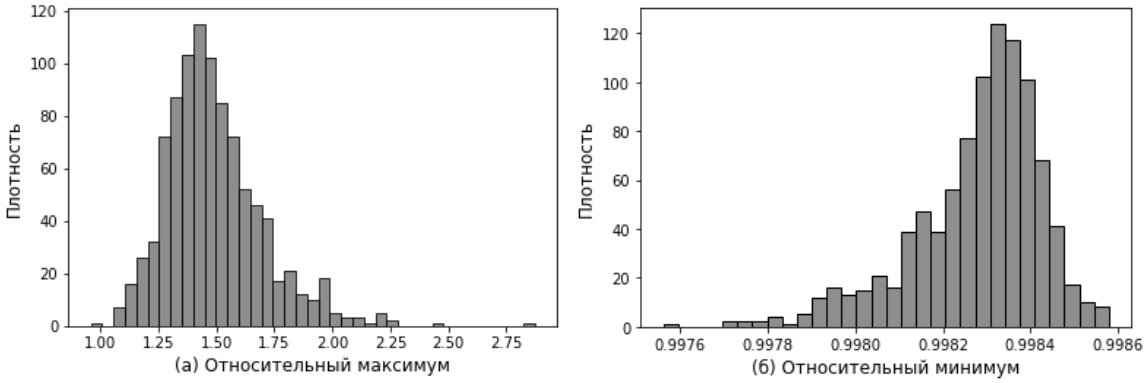

Рис. 6. Относительные максимумы (а) и относительные минимумы (б) по группам слов

2.3 Статистика по авторам

Средние значения по авторам (Таблица 7, Рис. 7а) похожи на средние значения по статьям (Таблица 1). Максимальное среднее значение составило 0,755, что довольно высоко, и означает, что некоторые авторы в среднем получили все оценки групп слов как подходящие с довольно высокой степенью уверенности. Среднеквадратические отклонения по авторам (Таблица 7, Рис. 7б) в среднем ниже, чем по статьям и по группам слов.

Таблица 7. Средние значения и среднеквадратические отклонения по авторам

	Для средних значений		**Для среднеквадратических отклонений**		**Продуктивность (число текстов)**
	Среднее значение	**Среднеквадратическое отклонение**	**Среднее значение**	**Среднеквадратическое отклонение**	
Число	68	68	68	68	68
Среднее	0,390180	0,148209	0,028587	0,028758	74,455882
Ср.-кв. отклонение	0,110397	0,071572	0,013366	0,013885	180,677121

	Для средних значений		Для среднеквадратических отклонений		Продуктивность (число текстов)
	Среднее значение	**Среднеквадратическое отклонение**	**Среднее значение**	**Среднеквадратическое отклонение**	
Минимум	0,055427	0,008227	0,007837	0,003887	1
25%	0,341787	0,110960	0,022248	0,022165	3
50%	0,391491	0,166579	0,027219	0,028923	8
75%	0,439551	0,195476	0,031894	0,035792	38
Максимум	0,754947	0,298349	0,089207	0,084738	976

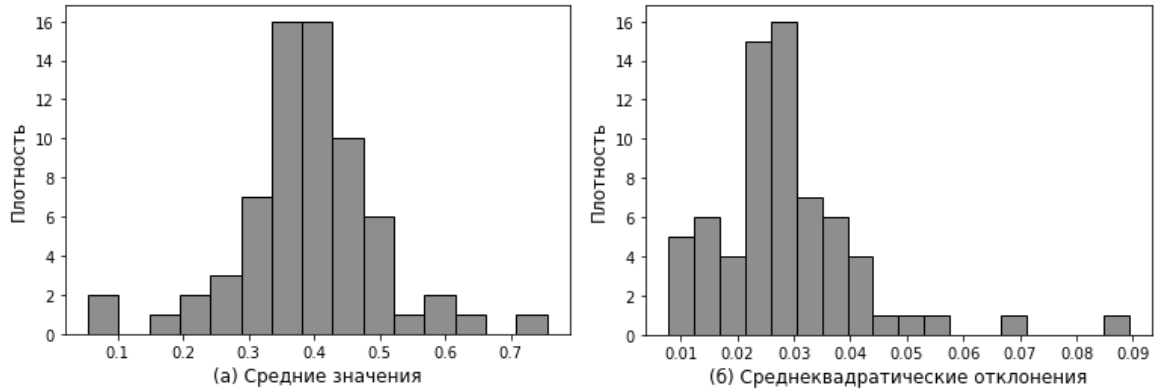

Рис. 7. Средние значения (а) и среднеквадратические отклонения (б) по авторам

В среднем, каждый автор получает максимальную оценку (Таблица 8, Рис. 8а) не ниже 0,762, что является довольно уверенным предсказанием, хотя есть авторы с максимальной оценкой 0,105. Максимальное значение минимальных оценок по авторам (Таблица 8, Рис. 8б) составило 0,545, что довольно много для минимальных оценок.

Таблица 8. Максимальные и минимальные значения по авторам

	Для средних значений		Для среднеквадратических отклонений	
	Максимальное значение	**Минимальное значение**	**Максимальное значение**	**Минимальное значение**
Число	68	68	68	68
Среднее	0,762235	0,074168	0,209269	1,966744e-04
Ср.-кв. отклонение	0,213288	0,111387	0,090553	2,784520e-04
Минимум	0,104501	0,000614	0,035964	2,281740e-08
25%	0,661660	0,004208	0,161512	9,113559e-06
50%	0,823251	0,024662	0,218436	8,025511e-05
75%	0,926780	0,105161	0,253840	2,921584e-04
Максимум	0,997850	0,545188	0,468500	1,167859e-03

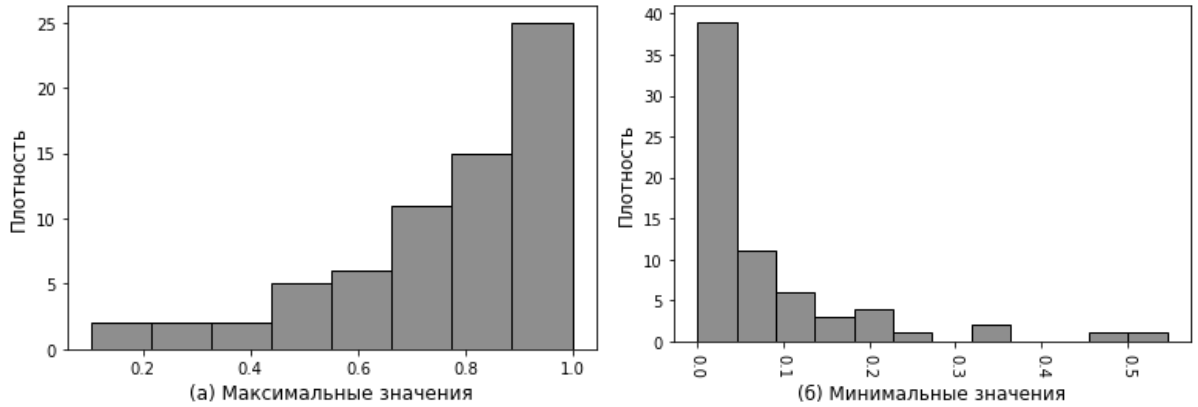

Рис. 8. Максимальные (а) и минимальные (б) значения по авторам

Относительные максимумы (Таблица 9, Рис. 9а) и относительные минимумы (Таблица 9, Рис. 9б) по авторам ниже, чем по группам слов (Таблица 6), но выше, чем по статьям (Таблица 3).

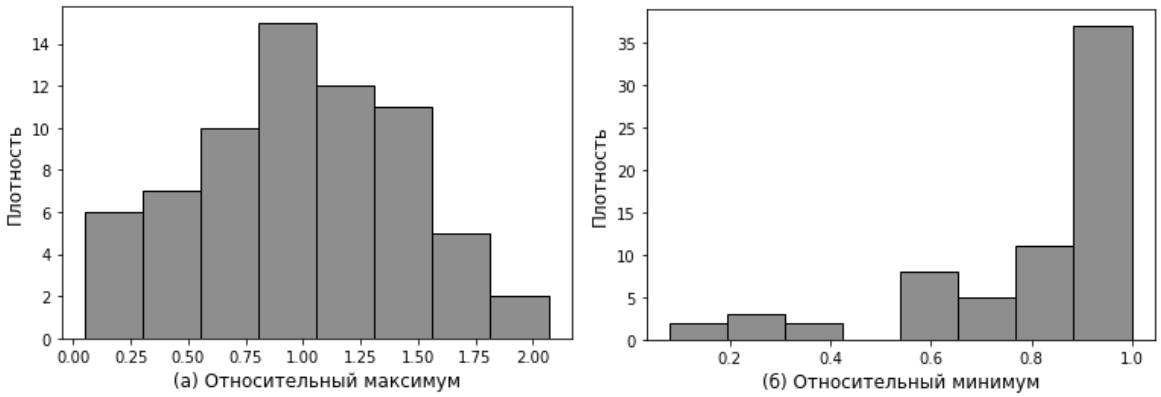

Рис. 9. Относительные максимумы (а) и относительные минимумы (б) по авторам

Таблица 9. Относительные максимальные и минимальные значения по авторам

	Относительный максимум как (максимум – среднее) / среднее	**Относительный минимум как (среднее – минимум) / среднее**
Число	68	68
Среднее	0,993790	0,818008
Ср.-кв. отклонение	0,472324	0,234457
Минимум	0,051299	0,081065
25%	0,712735	0,690448
50%	1,002944	0,915209
75%	1,379996	0,986916
Максимум	2,067960	0,998504

Очень продуктивные авторы (Таблица 7) имеют средние оценки (Рис. 10).

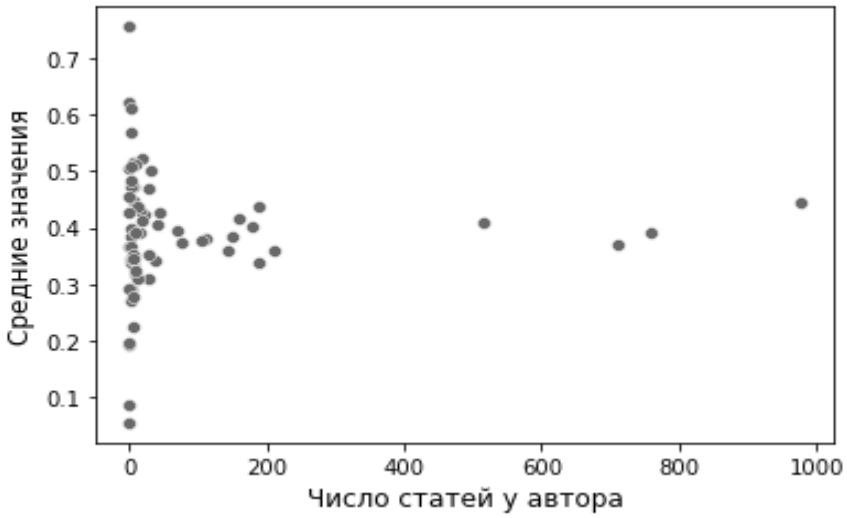

Рис. 10. Средние значения по авторам и их продуктивность

Заключение

Модель получает на вход группы синонимов и мы ожидаем, что модель оценит, насколько хорошо разные слова описывают заданные тексты, но получаем неожиданные статистические результаты. Например, максимальное значение для минимальных значений по статьям составило 0,982, что слишком высоко, и означает, что для некоторых текстов все группы слов были оценены как подходящие. Требуются дальнейшие исследования.

Литература

1. *Smirnov I.* Estimating Educational Outcomes from Students' Short Texts on Social Media // EPJ Data Science, 9, 2020, article number: 27. https://doi.org/10.1140/epjds/s13688-020-00245-8
2. *Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzman F., Grave E., Ott M., Zettlemoyer L., Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // 2019, arXiv:1911.02116v2 [cs.CL]
3. *Siva Sai.* Fine-tuning Transformer Neural Networks for Identification of Informative Covid-19 Tweets // Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), pp. 337–341. https://www.aclweb.org/anthology/2020.wnut-1.45/
4. *Huang Po-Yao, Patrick M., Hu J., Neubig G., Metze F.*, Hauptmann A. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models // 2021, arXiv:2103.08849v2 [cs.CV]
5. *Sainz O., Rigau G.* Ask2Transformers: Zero-Shot Domain labelling with Pre-trained Language Models // 2021, arXiv:2101.02661v2 [cs.CL]
6. *Kim T., Li B., Lee S.* Chart-based Zero-shot Constituency Parsing on Multiple Languages // 2020, arXiv:2004.13805v2 [cs.CL]
7. *Sheth J., Lee Y.-S., Astodillo R. F., Naseem T., Florian R., Roukos S., Ward T.* Bootstrapping Multilingual AMR with Contextual Word Alignments // 2021, arXiv:2102.02189v1 [cs.CL]