Эмбеддинги языковой модели RuBERT в задаче многоклассовой классификации постов пользователей в социальной сети

В. Д. Олисеенко
Санкт-Петербургский федеральный исследовательский центр РАН
vdo@dscs.pro

М. В. Абрамов
Санкт-Петербургский федеральный исследовательский центр РАН
mva@dscs.pro

***Аннотация*. Работа посвящена результату построения моделей для решения существующей задачи многоклассовой классификации постов пользователей в социальной сети. Полученные модели основываются на эмбеддингах, извлеченных из постов посредством языковой модели RuBert, и надстроенной над ними полносвязной нейронной сети. Также проведено сравнение полученных моделей с классическими моделями нейронных сетей на основе архитектуры долгой краткосрочной памяти (LSTM). Полученные результаты позволят улучшить автоматизацию части процесса оценки степени выраженности психологических особенностей пользователей по их постам в социальной сети.**

Ключевые слова: многоклассовая классификация, посты в социальных сетях, RuBert, long-short term memory, нейронные сети, машинное обучение

I. Введение

Задача оценки выраженности психологических особенностей пользователей в социальных сетях является актуальной, так как ее устоявшееся решение еще не предложено, а результаты такой оценки могли бы быть применены в различных областях жизнедеятельности человека: психологии и социологии [1], маркетинге [3], защите пользователей от социоинженерных атак [4] и т. д. Кроме того, одним из преимуществ получения такой оценки (в сравнении с прохождением, например, психологического исследования) является возможность её автоматизации посредством анализа контента, публикуемого пользователем на своей странице в социальной сети. В рамках данной гипотезы в работе [5] была предложена схема для классификации текстовых постов пользователей и выявлена её связь с результатами психологических тестов, пройденных этими пользователями. В рамках данной классификации предлагалось использовать двухуровневую схему, где к верхнему уровню классификации относились посты: информационного, эмоционального и побудительно-деятельностного характера, а к нижнему – их подклассы.

Для информационного класса подклассами выступают формальные, событийные, личные, интеллектуально-рассудительные, ссылочные, кулинарные посты; эмоционального – позитивные, негативные и поздравительные; побудительно-деятельностного – благотворительные, продающие, побудительные к действию. В данной схеме классы верхнего уровня имели пересечения, а их подклассы – нет. Таким образом, каждый пост мог относиться к одному, двум или трём классам верхнего уровня одновременно и только к одному из подклассов класса, к которому относится. В работах [6][8] были предложены некоторые подходы для автоматизации классификации постов, однако они имели ряд проблем, которые будут рассмотрены в разделе II.

Теоретическая значимость работы состоит в комбинировании методов и подходов для повышения точности классификации постов по ранее разработанным критериям. Практическая значимость заключается в доработке существующей системы классификации постов, ее автоматизации и реализации в прототипе комплекса программ, который может быть использован в качестве инструментария для оценки защищенности пользователей информационных систем от социоинженерных атак.

II. Релевантные работы

В ранее представленных работах [6][8] был предложен подход для автоматизации классификации постов посредством построения двухуровневой иерархический модели, где на первом уровне определялись главные классы (информационные, эмоциональные и побудительно-деятельностные) при помощи одной нейронная сети с двумя слоями архитектуры долгой-краткосрочной памяти (англ. long-short term memory), решающей задачу многозначной классификации [6], а на втором – три модели нейронной сети с той же архитектурой (по числу главных классов) для определения всех подклассов (задача многоклассовой классификации) [8]. Основная проблема данного подхода заключалась в сложности подсчёта общей ошибки двухуровневой модели классификации из-за разных решаемых задач (многоклассовой и многозначной классификации) и малым набором данных для обучения. Первая обозначенная проблема решается построением трёх многоклассовых моделей с числом классов по количеству подклассов плюс один (для обозначения отсутствия принадлежности поста одному из подклассов). Таким образом, для информационных постов необходима модель с 7 классами, для эмоциональных и побудительно-деятельностных постов с 4 классами. Для решения второй проблемы – недостатка набора данных – предлагается использовать векторные представления предложений, полученных при

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2022-0003, при финансовой поддержке РФФИ проект №20-07-00839, при финансовой поддержке гранта Президента МК-5237.2022.1.6.

помощи предобученной языковой модели, построенной на основе архитектуры трансформера – RuBERT (обученная на русскоязычном корпусе слов модель BERT) [9], [10].

Предобученные на большом количестве неразмеченных текстовых данных такие нейронные сети как ELMo [11], OpenAI GPT [12] и BERT [9] являются эталонами в решении современных задач обработки естественных языков (классификации текстов, предсказании слов, создании чат-ботов и т. д.). Особо выделяется среди них многослойнйый двунаправленный трансформер (англ. Bidirectional Encoder Representations from Transformers) или BERT, который показывает наилучшие результаты во многих задачах [13]. Кроме того, BERT позволяет получать эмбеддинги (или векторные представления) слов/предложений с учётом контекста, в котором они употребляются, что может сильно улучшить качество классификации в особенности на малом наборе данных [14]. В качестве классифицирующей модели для отнесения того или иного эмбеддинга к одному из классов достаточно построить простую полносвязную трёхслойную нейронную сеть.

III. Постановка задачи

Задачу классификации постов пользователей в социальных сетях, в соответствии со схемой критериев, можно свести к трём задачам мультиклассовой (англ. multi-class) классификации: пусть X – множество постов пользователей, а $y_1 \in \{0,...,6\}$ – метка информационного подкласса (не относящегося к информационному посту, формального, событийного, личного, интеллектуально-рассудительного, ссылочного, кулинарного), $y_2 \in \{0,1,2,3\}$ – метка эмоционального подкласса (не относящегося к эмоциональному посту, позитивного, негативного, поздравительного,), $y_3 \in \{0,1,2,3\}$ – метка побудительно-деятельностного подкласса (не относящегося к побудительно-деятельностному посту, благотворительного, продающего, побудительного к действию). Тогда функции $F_1(X, y_1)$, $F_2(X, y_2)$, $F_3(X, y_3)$ должны ставить каждому посту в соответствии по одной метки (рис. 1).

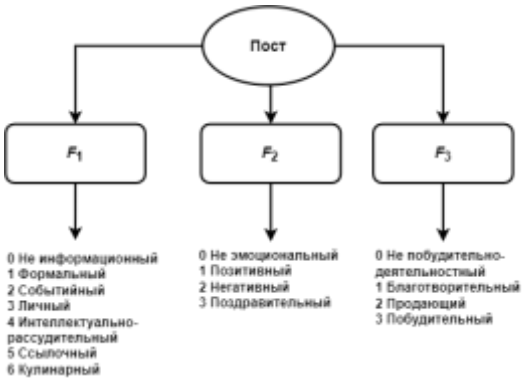

Рис. 1. Модели классификации

Например, для поста «Завтра у меня зачёт. Не люблю зачёты…» функции F_1 поставит метку 1 (событийный подкласс), F_2 метку 1 (негативный подкласс), а F_3 – метку 3 (не относящейся к побудительно-деятельностному подклассу).

В качестве элементов множества X выступает текст постов пользователей. Процесс подготовки текста описан в следующем разделе.

IV. Описание решения задачи

Набор данных (более 2-х тысяч постов), рассматриваемый в данном исследовании, был собран в пилотном исследовании [5] при помощи анкетирования респондентов и сбору их постов с персональных страниц в социальной сети «ВКонтакте». Кроме того, он был дополнен ещё 300 постов, собранных специально для данного исследования. Все полученные посты были размечены группой экспертов.

Для проведения программного эксперимента был выбран язык Python версии 3.7.13. Для предобработки текста использовались следующие библиотеки:

1. Re – для работы с регулярными выражениями (удаления смайликов, символов языков отличных от русского, специальных символов, знаков пунктуации, цифр и т. д.);
2. Natasha – для лемматизации текста;
3. Tensorflow – получение токенов слов, приведение всех постов к единой длине (512 слов).

После предобработки количество постов составило 1797 штуки.

Для получения эмбедингов постов использовалась библиотека DeepPavlov[1] с реализацией RuBERT [10]. На вход данной библиотеки поступал нетокенезированный текст постов длинной до 512 слов, на выход – эмбеддинг (числовой вектор) размера 768, отражающий векторное представление предложения. В классической реализации модель RuBERT на каждый входной текст выдаёт эмбеддинг размера $n \times 768$, где n – число слов. Для получения эмбеддинга предложений используется реализация, предусмотренная библиотекой DeepPavlov, вычисления среднего и максимального по всем n векторов, на которой и будет строится дальнейший эксперимент. В качестве классифицирующей модели использовалась трёхслойная полносвязная нейронная сеть, где на первых двух слоях (256 и 128 нейронов соответственно) использовалась функция активации сигмоида, а на последнем (нейронов по числу классов) – функция активации софтмакс (рис. 2).

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 256)               196864

 dense_1 (Dense)             (None, 128)               32896

 dense_2 (Dense)             (None, 6)                 774

=================================================================
Total params: 230,534
Trainable params: 230,534
Non-trainable params: 0
```

Рис. 2. Пример нейронной сети для информационного подкласса

На вход такой сети подаётся эмбеддинг предложения, на выход оценка принадлежности к каждому подклассу в промежутке [0, 1].

[1] https://deeppavlov.ai/ — An open source conversational AI framework

В качестве базового результата для сравнения будут использованы оценки, полученные при помощи расширения (на один подкласс) ранее разработанных моделей на основе LSTM нейронов [8]. Их архитектура представляет собой 4 слоя: Embedding, SpatialDropout1D, LSTM и Dense слои (рис. 3).

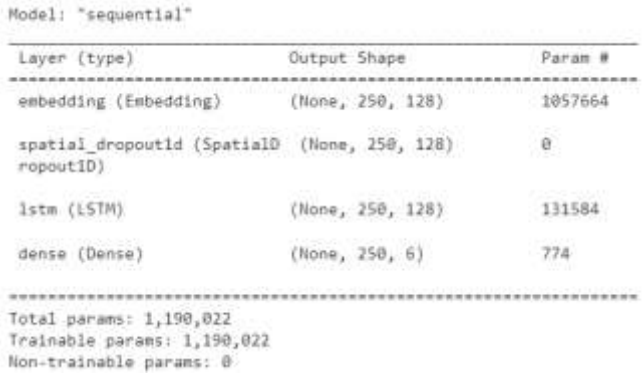

Рис. 3. Пример LSTM сети для информационного подкласса

На вход такой сети поступает токенезированный текст, а на выход – также оценка принадлежности к каждому подклассу в промежутке [0, 1]. Для проверки достоверности получаемых результатов используется процедура скользящего контроля по четырем блокам. Для оценки полученных моделей используются метрики F1-микро, F1-макро и точность (Accuracy) [15].

V. Эксперимент

Результаты эксперимента представлены в таблице. Стоит отметить, что метрики усреднены по результатам процедуры скользящего контроля на четырёх тестовых блоках. В таблице представлены девять моделей: первые шесть используют одинаковые архитектуры – полносвязные нейронные сети с входными эмбеддингами, полученными при помощи RuBERT (по три модели на каждый тип эмбеддинга – усредненный и максимальный), вторые три – полученные на основе LSTM нейронной сети.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

	F1-микро	**F1-макро**	**Точность**
RuBERT (усреднённый эмбеддинг)			
Информационные подклассы	**0.5234**	**0.5156**	**0.5947**
Эмоциональные подклассы	**0.5175**	**0.4912**	**0.5455**
Побудительно-деятельностные подклассы	**0.6873**	**0.5721**	**0.7026**
RuBERT (максимальный эмбеддинг)			
Информационные подклассы	0,4672	0,4390	0,5347
Эмоциональные подклассы	0.4569	0.4489	0.5726
Побудительно-деятельностные подклассы	0.6053	0.4875	0.6223
LSTM нейронная сеть			
Информационные подклассы	0.3587	0.3438	0.3587
Эмоциональные подклассы	0.3387	0.3452	0.3387
Побудительно-деятельностные подклассы	0.3863	0.3401	0.3863

По полученным результатом можно сделать вывод, что наилучшим образом себя показывает полносвязная нейронная сеть, обученная на усреднённом эмбеддинге RuBERT'а. Также стоит отметить фактическую невозможность модели на основе LSTM нейронных сетей различать классы между собой (одинаковые метрики точности и F1-микро). Полученные результаты также свидетельствуют о чрезмерно малой выборке для обучения и сильным дисбалансом в классах, обнаруженном в [6]–[8]. Таким образом, ключевым направлением дальнейших исследований является существенное расширения набора данных.

VI. Заключение

В работе были получены модели, которые основываются на эмбеддингах, извлеченных из постов посредством языковой модели RuBert, и надстроенной над ними полносвязной нейронной сети. Проведено сравнение полученных моделей с классическими моделями нейронных сетей на основе архитектуры долгой краткосрочной памяти (LSTM). Полученные результаты позволят существенно улучшить автоматизацию части процесса оценки степени выраженности психологических особенностей пользователей по их постам в социальной сети (в сравнении с подходом на основе LSTMв том числе, используемом в более ранних работах).

Теоретическая значимость работы состоит в комбинировании методов и подходов для улучшения (повышения точности) автоматизации классификации постов по ранее разработанным критериям. Практическая значимость заключается в доработке существующей системы классификации постов. Дальнейшими направлениями исследований является существенное расширения набора данных, в т.ч. за счёт краудсорсинговых систем, и использования других языковых моделей (ELMo [11], OpenAI GPT [12]).

Список литературы

[1] Graham S., Depp C., Lee E.E., Nebeker C., Tu X., Kim H.-C., Jeste D.V. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview // Current Psychiatry Reports. 2019. 21 (11). № 116. Doi: 10.1007/s11920-019-1094-0

[2] Khlobystova A., Korepanova A., Maksimov A., Tulupyeva T. An Approach to Quantification of Relationship Types Between Users Based on the Frequency of Combinations of Non-numeric Evaluations // Advances in Intelligent Systems and Computing. 2020. 1156 AISC, pp. 206–213. Doi: 10.1007/978-3-030-50097-9_21

[3] Lăzăroiu G., Neguriţă O., Grecu I., Grecu G., Mitran P.C. Consumers' Decision-Making Process on Social Commerce Platforms: Online Trust, Perceived Risk, and Purchase Intentions // Frontiers in Psychology. 2020. 11. № 890. Doi: 10.3389/fpsyg.2020.00890

[4] Khlobystova A.O., Tulupyev A.L. Approaches to modeling development scenarios of multistep social engineering attacks // Proceedings of 2021 4th International Conference on Control in Technical Systems, CTS 2021. 2021. P. 100–102. Doi: 10.1109/CTS53513.2021.9562746

[5] Тулупьева Т.В., Тафинцева А.С., Тулупьев А.Л. Подход к анализу отражения особенностей личности в цифровых следах // Вестн. психотерапии. 2016. № 60 (65). С. 124–137.

[6] Oliseenko V.D., Tulupyeva T.V. Neural Network Approach in the Task of Multi-label Classification of User Posts in Online Social Networks // 2021 XXIV International Conference on Soft Computing and Measurements (SCM). 2021. P. 46–48. Doi:10.1109/SCM52931.2021.9507148

[7] Oliseenko V.D., Tulupyeva T.V., Abramov M.V. Online Social Network Post Classification: A Multiclass approach. In: Kovalev S., Tarassov V., Snasel V., Sukhanov A. (eds) Proceedings of the Fifth International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'21). IITI 2021. 2022. Lecture Notes in Networks and Systems, vol 330. Springer, Cham. Doi:10.1007/978-3-030-87178-9_21

[8] Олисеенко В.Д., Абрамов М.В., Тулупьев А.Л. Нейронные сети lstm и gru в приложении к задаче многоклассовой классификации текстовых постов пользователей социальных сетей. Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2021. № 4. С. 130–141. Doi: 10.17308/sait.2021.4/3803

[9] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 2019. P. 4171–4186

[10] Kuratov Yu., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. 2019-May. 18. P. 333–339.

[11] Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)

[12] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018) [Электронный ресурс]. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/languageunderstanding paper.pdf

[13] Acheampong F.A., Nunoo-Mensah H., Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches // Artificial Intelligence Review, 54 (8), P. 5789–5829. Doi: 10.1007/s10462-021-09958-2

[14] Ezen-Can A. A Comparison of LSTM and BERT for Small Corpus. arXiv preprint arXiv:2009.05451 (2020)

[15] Grandini, M. Metrics for Multi-Class Classifi-cation: an Overview / M. Grandini, E. Bagli, G. Visani // 2020. arXiv preprint arXiv:2008.05756.