Интерпретация матриц внимания языковых моделей при анализе тональности текстов

Д.Э. Пащенко
Вятский государственный университет
Киров, Россия
depashchenko@mail.ru

Е.В. Разова
Вятский государственный университет
Киров, Россия
razova.ev@gmail.com

А.В. Котельникова
Вятский государственный университет
Киров, Россия
kotelnikova.av@gmail.com

С.В. Вычегжанин
Вятский государственный университет
Киров, Россия
vychegzhaninsv@gmail.com

Е.В. Котельников
Вятский государственный университет
Киров, Россия
kotelnikov.ev@gmail.com

***Аннотация*—Глубокие нейросетевые языковые модели демонстрируют в последнее время впечатляющие результаты при обработке естественного языка, в том числе в области анализа тональности текстов. Однако уверенность пользователей в результатах анализа невысока в связи с плохой интерпретируемостью таких моделей. В работе предлагается метод интерпретации языковых моделей при помощи вероятностно-статистического анализа матриц внимания.**

Ключевые слова— анализ тональности, нейронные сети, языковые модели, BERT, интерпретация, внимание, оценочная лексика.

1. Введение

Стремительное развитие в последние 3–4 года глубоких нейросетевых языковых моделей [1], таких как BERT [2], GPT-3 [3], T5 [4], Switch Transformer [5], ERNIE [6] и других, привело к значительному прогрессу в обработке естественного языка, в частности, при анализе тональности текстов. В этой области современные модели демонстрируют точность выше 97% для задач бинарной классификации (позитив/негатив) и около 60% для задач с пятизначной шкалой [7]. Большинство современных языковых моделей использует в качестве основы архитектуру Transformer [8], в которой ключевая информация содержится в т.н. матрицах внутреннего внимания (self-attention), характеризующих степень влияния различных входных сигналов на результат.

Несмотря на существенные успехи, применение глубоких нейросетевых языковых моделей имеет ряд проблем, в частности, необходимость наличия большого количества размеченных данных для обучения, высокую вычислительную трудоемкость процесса обучения, низкую интерпретируемость.

Альтернативой являются словарные методы (или методы на основе правил) – быстрые, не требующие обучения, хорошо интерпретируемые [9]. Основным лингвистическим ресурсом для них являются словари оценочной лексики, большое количество которых было разработано за последние годы, в том числе для русского языка [10]. Однако словарные методы не обеспечивают высокого качества анализа тональности и значительно уступают в этом плане нейросетевым моделям [11]. Причина такой ситуация остается до конца неясной в связи с плохой интерпретируемостью нейронных сетей.

Таким образом, актуальной является задача улучшения интерпретируемости глубоких нейросетевых языковых моделей с целью обеспечения прозрачности процедур классификации текстов по тональности и для потенциального повышения качества работы словарных методов. В настоящей работе указанная задача решается на основе вероятностно-статистического анализа распределения весов в матрицах внимания, формируемых после дообучения модели типа BERT на корпусе текстов, размеченных по тональности.

2. Обзор предыдущих работ

Интерпретация на основе внимания предполагает анализ распределения внимания модели по токенам. Токены с наибольшим вниманием считаются наиболее полезными для предсказания модели. Тан и др. исследовали влияние токенов с наибольшим весом внимания в модели $BERT_{BASE}$ на правильность классификации текстов по тональности [12]. Результаты показали, что первый слой более эффективен для выбора влиятельных токенов, чем последний. Као и др. [13] предложили метод дифференциальной маскировки, позволяющий определить, что «знают» о входных данных разные слои модели и где в разных слоях хранится информация о предсказании. Было установлено, что большинство слоев полагается на строго позитивные или негативные слова. Ву и др. [14] разработали метод послойного отслеживания внимания для анализа структурированных весов внимания в моделях на архитектуре Transformer. В работе была определена доля внимания, уделяемая отдельными головами оценочным словам. Авторы показали, что веса внимания имеют наибольшие значения для сильно позитивных и негативных слов.

3. Метод

Предлагаемый метод интерпретации глубоких нейросетевых языковых моделей с механизмом внутреннего внимания включает четыре этапа.

1) Построение усредненных матриц внимания для каждого текста. Значения в исходных матрицах внимания моделей рассредоточены по отдельным входным токенам, которые часто являются не словом целиком, а его частью. Поэтому возникает задача нахождения усредненных весов внимания для слов с учетом лемматизации.

2) Построение единых матриц внимания по всем текстам анализируемого корпуса. Единые матрицы внимания формируются на основе усреднения матриц, полученных на предыдущем этапе.

3) Анализ статистических характеристик сформированных единых матриц. Анализ осуществляется с учетом различных аспектов: оценочной и нейтральной лексики; эталонной разметки текстов по тональности и разметки, присвоенной моделью. В процессе анализа вычисляются и сравниваются средние, медианные, максимальные и минимальные значения весов внимания.

4) Анализ вероятностных распределений весов внимания сформированных единых матриц. Расчет дивергенции Кульбака-Лейблера между распределениями весов внимания различных множеств слов – позитивных, негативных и нейтральных.

Предложенный метод позволяет проверять различные гипотезы о степени внимания, уделяемого нейросетевыми языковыми моделями лексике различного типа в разных текстах при разных условиях.

4. Результаты

Для экспериментов с предложенным методом в работе использовалась модель ruRoberta-large [15], продемонстрировавшая лучшие результаты среди моделей типа BERT в рейтинге Russian SuperGLUE [16].

Эксперименты проводились на материале корпуса русскоязычных новостей, включающего 1823 текста, размеченных аннотаторами по трехбалльной шкале: «позитивно», «негативно» и «нейтрально».

Словарь оценочной лексики должен обладать высокой точностью и полнотой. Чтобы создать такой словарь были проанализированы девять общедоступных словарей русскоязычной оценочной лексики [10]. В итоговый словарь включались только те слова, которые входят не менее чем в три исходных словаря. В итоговый словарь вошли 4325 слов, в том числе 1521 позитивных (35,2%) и 2804 негативных (64,8%).

Была выдвинута гипотеза о том, что нейросетевая языковая модель ruRoberta-large уделяет в процессе анализа тональности больше внимания оценочной лексике, чем нейтральной. Применение предложенного метода позволило подтвердить данную гипотезу с помощью критерия суммы рангов Уилкоксона (уровень значимости p=0,05) [17]. Также были проанализированы отличия вероятностно-статистических характеристик внимания для правильно определяемых моделью текстов и ошибочных.

5. Заключение

Предложенный в исследовании метод интерпретации матриц внимания глубоких нейросетевых языковых моделей позволяет проверять различные гипотезы относительно распределений значений внимания между словами различных множеств при анализе тональности текстов. Метод может быть использован для повышения уверенности пользователей в результатах работы нейросетевых языковых моделей и для улучшения качества анализа тональности на основе словарных методов.

Литература

[1] Han, X. Pre-Trained Models: Past, Present and Future / X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu // AI Open. – 2021.

[2] Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of 7th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). – 2019. – P. 4171-4186.

[3] Brown, T.B. Language Models are Few-Shot Learners / T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan // Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS). – 2020. – P. 1877-1901.

[4] Raffel, C. Exploring the limits of transfer learning with a unified text-to-text transformer / C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang // Journal of Machine Learning Research. – 2020. – Vol. 21. – P. 1-67.

[5] Fedus, W. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity / W. Fedus, B. Zoph, N. Shazeer // Computing Research Repository. – 2021. – Access mode: http://arxiv.org/abs/2101.03961 (08.02.2022).

[6] Sun, Y. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation / Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang // Computing Research Repository. – 2021. – Access mode: http://arxiv.org/abs/2107.02137 (08.02.2022).

[7] Papers with code: Sentiment Analysis [Electronic resource]. – Access mode: https://paperswithcode.com/task/sentiment-analysis (08.02.2022).

[8] Vaswani, A. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones // Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 6000-6010.

[9] Taboada, M. Sentiment Analysis: An Overview from Linguistics / M. Taboada // Annual Review of Linguistics. – 2020. – Vol. 2. – P. 325-347.

[10] Котельников, Е.В. Современные словари оценочной лексики для анализа мнений на русском и английском языках / Е.В. Котельников, Е.В. Разова, А.В. Котельникова, С.В. Вычегжанин // Научно-техническая информация. Сер. 2. Информационные процессы и системы. – 2020. – № 12. – С. 16-33.

[11] Kotelnikova, A.V. Lexicon-based Methods and BERT Model for Sentiment Analysis of Russian Text Corpora / A.V. Kotelnikova, D.E. Paschenko, E.V. Razova // CEUR Workshop Proceedings. – 2021. – Vol. 2922. – P. 73-81.

[12] Tan, N.O. An explainability analysis of a sentiment prediction task using a transformer-based attention filter / N.O.Tan,, J. Bensemann, D. Benavides-Prado, Y. Chen, M. Gahegan // Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems. – 2021. – P. 1-7.

[13] Cao, N.D. How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking / N.D. Cao, M.S. Schlichtkrull, W. Aziz, I. Titov // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – 2020. – P. 3243-3255.

[14] Wu, Z. Structured Self-Attention Weights Encode Semantics in Sentiment Analysis / Z. Wu, T.-S. Nguyen, D. Ong // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. – 2020. – P. 255-264.

[15] Sberbank-ai/ruRoberta-large [Electronic resource]. – Access mode: https://huggingface.co/sberbank-ai/ruRoberta-large (08.02.2022).

[16] Russian SuperGLUE: Leaderboard [Electronic resource]. – Access mode: https://russiansuperglue.com/leaderboard/2 (08.02.2022).

[17] Wilcoxon, F. Individual comparisons by ranking methods / F. Wilcoxon // Biometrics Bulletin. – 1945. – Vol 6(1). P. 80-83.