УДК 004.8

НЕЙРОННЫЕ СЕТИ ДЛЯ АНАЛИЗА ТЕКСТОВ: ПОЛНОСВЯЗНЫЕ И РЕККУРЕНТНЫЕ АРХИТЕКТУРЫ В СЕНТИМЕНТ-АНАЛИЗЕ

Вишневская А.А.

Студент

ФГБОУ ВО «Поволжский государственный университет телекоммуникаций и информатики»

Самара, Россия

Лиманова Н.И.

д.т.н., профессор, научный руководитель

ФГБОУ ВО «Поволжский государственный университет телекоммуникаций и информатики», ФГБОУ ВО «Самарский государственный технический университет»

Самара, Россия

Аннотация

В статье рассматривается использование нейронных сетей для анализа текстов в задачах обработки естественного языка и сентимент-анализа. Описаны основные архитектуры, включая полносвязные сети и рекуррентные нейронные сети (RNN), а также их преимущества и ограничения. Особое внимание уделено переходу от обработки текста как набора независимых признаков к учёту последовательности слов и контекста, что позволяет более точно моделировать смысл и эмоциональную окраску текста. Рассматриваются механизмы скрытых состояний, проблемы затухающих и взрывающихся градиентов, а также архитектуры LSTM и GRU для работы с долгосрочными зависимостями.

Ключевые слова: нейронные сети, обработка текстов, полносвязные сети, рекуррентные нейронные сети, LSTM, GRU, NLP, сентимент-анализ, контекст слов, последовательность текста

NEURAL NETWORK FOR TEXT ANALYSI: DENSE AND RECCURENT ARCHITECTS IN NLP AND SENTIMENT ANALYSIS

Vishnevskaya A.A.

Student

Volga Region State University of Telecommunications and Informatics

Samara, Russia

Limanova N.I.

Doctor of Technical Sciences, Professor, Scientific supervisor

Volga Region State University of Telecommunications and Informatics

Samara State Technical University

Samara, Russia

Abstract

This section discusses the use of neural networks for text analysis in natural language processing and sentiment analysis tasks. The main architectures, including fully connected networks and recurrent neural networks (RNN), as well as their advantages and limitations, are described. Special attention is given to the transition from processing text as a set of independent features to considering the sequence and context of words, enabling more accurate modeling of text meaning and sentiment. Hidden state mechanisms, vanishing and exploding gradient problems, as well as LSTM and GRU architectures for long-term dependencies, are also discussed.

Keywords: neural networks, text analysis, fully connected networks, recurrent neural networks, LSTM, GRU, NLP, sentiment analysis, word context, text sequence

Современные методы анализа текстов активно используют нейронные сети, которые способны выявлять сложные зависимости между словами и фразами, учитывать контекст и семантику текста. В задачах сентимент-анализа и обработки естественного языка (Natural Language Processing, NLP) нейросети

позволяют классифицировать тексты, выделять эмоциональную окраску и извлекать значимые характеристики из последовательностей слов [1; 2; 3; 4; 5].

Основная идея применения нейросетей заключается в том, что текст, представленный в цифровом виде (векторизованный), может быть подан на вход сети, которая затем обучается выявлять закономерности и делать предсказания. В рамках данной статьи рассматриваются различные архитектуры нейронных сетей, начиная с полносвязных моделей и переходя к сетям, способным работать с последовательными данными,

Полносвязные нейронные сети (Fully Connected Neural Networks, FCNN), также называемые многослойными перцептронами (MLP), представляют собой класс нейросетевых моделей, в которых каждый нейрон одного слоя соединён со всеми нейронами следующего слоя [6; 7]. Такая архитектура позволяет эффективно моделировать сложные зависимости между признаками входных данных и выдавать на выходе прогнозные значения.

В контексте обработки текстов полносвязные сети обычно применяются после того, как текст был представлен в числовом виде, например с использованием one-hot кодирования или плотных векторных представлений слов (embeddings) [8; 9]. Каждый входной нейрон сети соответствует определённой характеристике текста — будь то конкретное слово, частота его встречаемости или компонент векторного представления.

Главным преимуществом полносвязных сетей является их способность аппроксимировать произвольные функции, что делает их эффективным инструментом для задач классификации и регрессии [10]. В задачах сентимент-анализа FCNN могут применяться для предсказания тональности текстов на основе заранее сформированных признаков, что делает их удобными для

базовых экспериментов и сравнительного анализа с более сложными архитектурами.

Однако полносвязные сети рассматривают входные данные как набор независимых признаков и не учитывают последовательность слов в тексте. Каждое слово или токен воспринимается изолированно, без контекста соседних слов, что существенно ограничивает возможности FCNN при анализе сложных текстов и предложений с синтаксической и семантической зависимостью [11; 12].

Таким образом, полносвязные нейронные сети представляют собой базовую архитектуру для анализа текстов, удобную для экспериментов с векторизованными данными, но имеющую ограничения при обработке последовательной информации. Эти ограничения обосновывают необходимость перехода к моделям, способным учитывать порядок слов и контекст — в частности, к рекуррентным нейронным сетям.

Рассмотрим ограничения, возникающие при работе с текстами:

Полносвязные нейронные сети, несмотря на свою универсальность, обладают рядом ограничений при работе с текстовыми данными. Основная проблема заключается в том, что такие сети рассматривают текст как набор независимых признаков, игнорируя порядок слов и внутренние зависимости между ними [6; 11].

Для задач сентимент-анализа это является критическим недостатком, поскольку смысл предложения и эмоциональная окраска часто зависят от последовательности слов. Например, предложения «Фильм не понравился» и «Фильм понравился не …» содержат схожий набор лексем, но имеют противоположное значение, что полносвязная сеть не способна корректно различить при стандартной векторизации.

Другим ограничением является высокая размерность входных данных. При использовании one-hot кодирования каждая уникальная лексема представляется отдельной размерностью, что приводит к формированию разреженных векторов и резкому росту числа параметров сети. Это увеличивает требования к объёму обучающих данных и вычислительным ресурсам, а также повышает риск переобучения модели [8; 13].

Кроме того, FCNN не обладают механизмом памяти о предыдущих входных данных. Каждое слово обрабатывается изолированно, что делает невозможным моделирование долгосрочных зависимостей внутри текста. В научной литературе подчёркивается, что для анализа сложных текстов необходимы архитектуры, способные учитывать контекст и последовательность, такие как рекуррентные нейронные сети и их модификации — LSTM и GRU [14; 15].

Также следует учитывать чувствительность полносвязных сетей к шуму и редким словам. Появление неизвестных или редко встречающихся токенов может существенно снижать качество предсказаний, поскольку FCNN не обладают механизмами обобщения новых словоформ без дополнительной обработки [12].

Таким образом, ограничения полносвязных нейросетей при работе с текстами включают:

Игнорирование порядка слов и контекста.

Высокая размерность входного пространства при использовании разреженных представлений.

Отсутствие памяти о предыдущих входах, что мешает моделированию долгосрочных зависимостей.

Чувствительность к редким и неизвестным словам.

Таким образом, вышеприведенные ограничения мотивируют исследователей использовать архитектуры, способные учитывать последовательность и контекст слов, что становится предметом последующего изучения рекуррентных и трансформерных моделей.

Перейдём к моделям, учитывающим последовательность.

Ограничения полносвязных нейронных сетей в обработке текстов, рассмотренные в предыдущем разделе, стимулировали развитие архитектур, способных учитывать порядок слов и их взаимозависимости. В отличие от FCNN, такие модели анализируют текст как последовательность элементов, где каждое слово влияет на последующие, а сеть «помнит» контекст предыдущих слов.

Рекуррентные нейронные сети (RNN) стали первым классом моделей, активно использующих концепцию учёта последовательности данных. Основная идея RNN заключается в том, что на каждом временном шаге сеть принимает на вход текущее слово (или его векторное представление) и скрытое состояние, которое агрегирует информацию о предыдущих словах. Таким образом, модель может учитывать порядок слов и передавать контекст вдоль всей цепочки [14; 15].

Переход от полносвязных нейронных сетей к RNN и другим последовательностным моделям обусловлен необходимостью решения нескольких ключевых задач:

Рассмотрим моделирование долгосрочных зависимостей. В тексте значение слов может зависеть от отдалённых элементов предложения или абзаца. Последовательные модели способны сохранять информацию о

предыдущих словах и использовать её для предсказания эмоциональной окраски или общего смысла текста.

Перейдём к рассмотрению контекстуализации слов. В отличие от one-hot кодирования, где каждое слово имеет фиксированный и независимый вектор, рекуррентные сети используют плотные векторные представления(embedding) и скрытые состояния для формирования контекстно-зависимых векторов слов, что существенно повышает точность анализа текста [9; 16].

Обсудим вопрос снижения зависимости от размера словаря. В полносвязных сетях редкие и новые слова сложно обрабатывать без расширения словаря. Последовательные модели, использующие плотный векторный слой(embedding) и обучение на субсловных единицах (subword units), позволяют более эффективно работать с неизвестными токенами и морфологическими вариациями слов [17].

В работе проведён вычислительный эксперимент на выборке 50 тысяч слов, классическом датасете отзывов на фильмы с IMDB, в котором были задействованы рассмотренные архитектуры: dense, RNN, LSTM, GRU.

Сопоставим результаты работы вышеназванных архитектур по метрикам точности, приведённым в Таблице 1, а также на Рис.1-4:

Таблица 1: Метрики качества обучения архитектур нейронных сетей.

	Precision	Recall	F1-мера	AUC
Dense	0.50	0.50	0.50	0.50
SimpleRNN	0.5589	0.5392	0.5489	0.589
LSTM	0.827	0.861	0.8436	0.890
GRU	0.8618	0.8245	0.8427	0.927

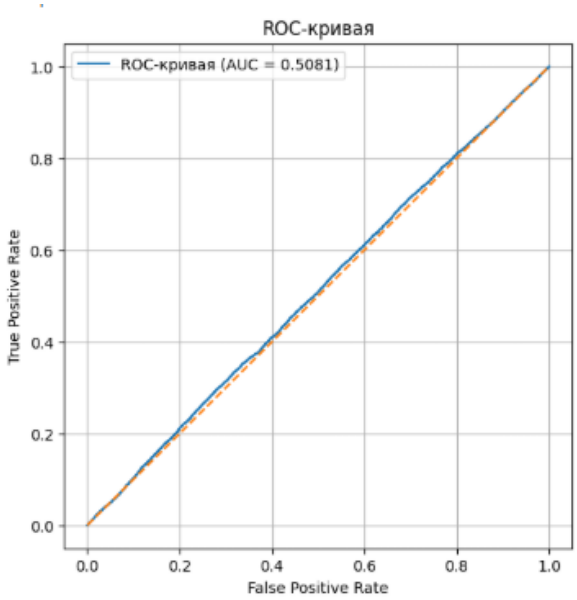

Рис. 1 Dense

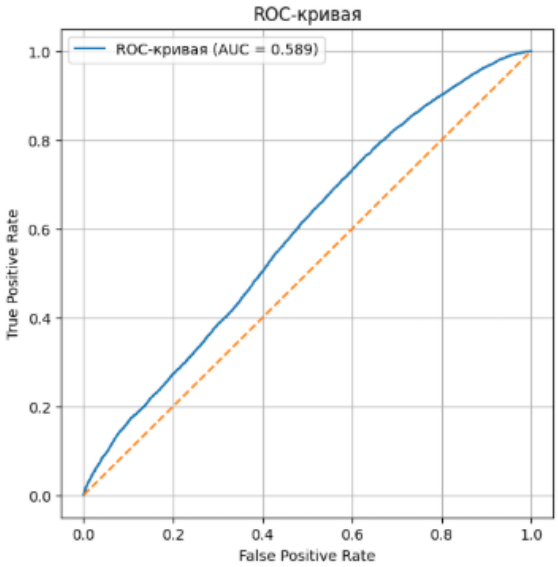

Рис. 2 SimpleRNN

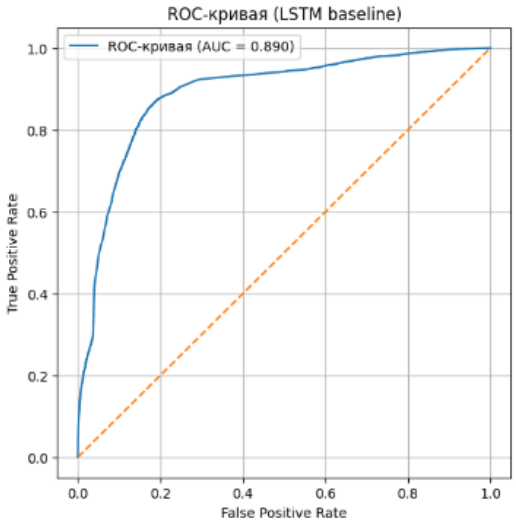

Рис.3 LSTM

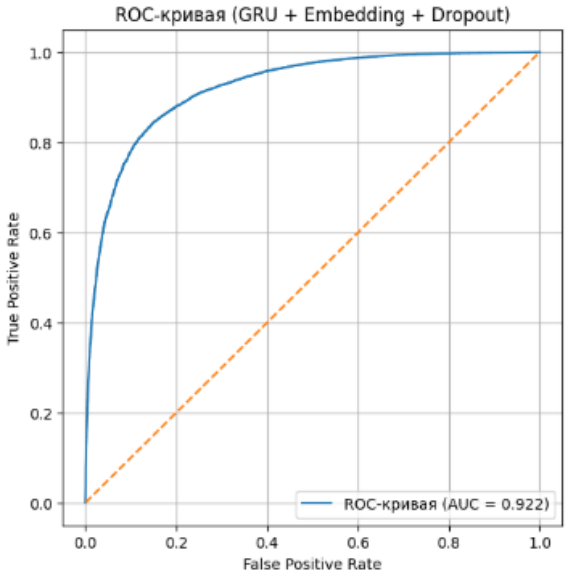

Рис.4 GRU

Говоря о метриках точности, архитектура Dense практически не справляется с определением последовательности слов в тексте, F1-мера = 0.50, а также приближающаяся к диагонали случайного классификатора ROC-кривая (см. рис. 1) свидетельствует о том, что модель запоминает выборку и не способна различать порядок слов, что делает её непригодной для решения задач сентимент-анализа.

Немногим лучшие метрики качества демонстрирует архитектура SimpleRNN. Судя по метрикам Precision и Recall модель находит всего 55 процентов верных ответов и не ошибается в их нахождении в 53 процентах случаев. Слабые показатели демонстрирует и ROC-кривая (Рис.2), небольшой изгиб которой сообщает о слабой способности модели различать классы. В совокупности, все это сообщает об ограниченных возможностях модели к анализу контекста употребления слов, это делает модель все еще слабо пригодной для сентимент-анализа.

Гораздо более высокие метрики демонстрируют LSTM и GRU, F1-мера которых составляет 0.8436 и 0.8427, это говорит о довольно высоком проценте нахождения положительных ответов и точность в этом вопросе. Сильный изгиб ROC-кривых (Рис.3 и Рис.4) и высокий параметр AUC сообщает о высокой способности модели к классификации, а значит и определению контекста слов. Данные модели являются наиболее предпочтительными при выборе архитектуры для сентимент-анализа.

Таким образом, переход к моделям, учитывающим последовательность, является ключевым этапом в эволюции нейросетевых методов анализа текстов. Он позволяет перейти от обработки текста как набора независимых признаков к обработке текста как структурированной информации, где порядок и контекст слов играют решающую роль. Это, в свою очередь, значительно повышает качество задач сентимент-анализа, машинного перевода и других приложений NLP.

Библиографический список

1. Maas, A. L., et al. Learning Word Vectors for Sentiment Analysis // ACL. — 2011.

2. Pang, B., Lee, L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. — 2008.
3. Киссель, А. А., Коваль, М. В. Анализ тональности текстов на русском языке: подходы и методы. — СПб: НИУ ИТМО, 2020.
4. Unicode Consortium. The Unicode Standard. — 2023.
5. Mohammadi, M., Hossein, M. Sentiment Analysis: Concepts, Methods and Applications. — Springer, 2021.
6. Salton, G., Buckley, C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. — 1988.
7. Pennington, J., Socher, R., Manning, C. GloVe: Global Vectors for Word Representation. — 2014.
8. Goldberg, Y. Neural Network Methods in Natural Language Processing. — Morgan & Claypool, 2017.
9. Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. — 1997.
10. Cho, K. et al. Learning Phrase Representations using RNN Encoder–Decoder. — 2014.
11. Баев, А. А. Сентимент-анализ текстов: методы и приложения // Информационные технологии. — 2018. — №4.
12. Тихомиров, Д. М., Коршунов, А. А. Методы автоматического анализа тональности текстов. — М.: Физматлит, 2019.
13. Jurafsky, D., Martin, J. H. Speech and Language Processing. — 3rd Edition. — Prentice Hall, 2023.
14. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. A Neural Probabilistic Language Model // JMLR. — 2003.
15. Mikolov, T. et al. Efficient Estimation of Word Representations in Vector Space. — 2013.
16. Esuli, A., Sebastiani, F. SentiWordNet: A Lexical Resource for Opinion Mining. — 2006.
17. Панченко, А. Методы анализа тональности текстов. — Диалог, 2012.
18. Кузнецов, И. Нейросетевые модели для русского языка // Вестник МГУ. — 2020.
19. Cho, K. et al. Learning Phrase Representations using RNN Encoder–Decoder. — 2014.
20. Лиманова Н.И., Ковтун Д.С. Искусственный интеллект и обработка естественного языка как основа чат-ботов // Бюллетень науки и практики. 2024. Т. 10. № 4. С. 426-429