Вестник Евразийской науки / The Eurasian Scientific Journal https://esj.today

2018, №3, Том 10 / 2018, No 3, Vol 10 https://esj.today/issue-3-2018.html

URL статьи: https://esj.today/PDF/43ITVN318.pdf

Статья поступила в редакцию 27.04.2018; опубликована 23.06.2018

Ссылка для цитирования этой статьи:
Краснов Ф.В. Анализ тональности текста научно-практических статей по нефтегазовой тематике с помощью искусственных нейронных сетей // Вестник Евразийской науки, 2018 №3, https://esj.today/PDF/43ITVN318.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

For citation:
Krasnov F.V. (2018). Analysis of the tone of the text of scientific and practical papers on oil and gas topics using artificial neural networks. *The Eurasian Scientific Journal*, [online] 3(10). Available at: https://esj.today/PDF/43ITVN318.pdf (in Russian)

УДК 316.452

Краснов Федор Владимирович
ООО «Газпромнефть НТЦ», Санкт-Петербург, Россия
Эксперт
Кандидат технических наук
E-mail: Krasnov.FV@Gazprom-Neft.ru
ORCID: http://orcid.org/0000-0002-9881-7371
РИНЦ: https://elibrary.ru/author_profile.asp?id=855886
Researcher ID: http://www.researcherid.com/rid/C-6518-2018

Анализ тональности текста научно-практических статей по нефтегазовой тематике с помощью искусственных нейронных сетей

Аннотация. Проблема поиска отрицательного практического опыта от применения новых технологий остро стоит в нефтегазовой отрасли. Риск повторения чужих ошибок высок в связи с разнообразием технологий и неопределенностями границ их применения. Наиболее полный источник накопленного опыта – это материалы научно-практических конференций. А наиболее точный метод поиска отрицательного практического опыта – это внимательно вчитываться в текст каждой презентации и выискивать положительный и отрицательный опыт от применения технологий. Но такой метод сложно применять на практике так как объемы текстов огромны. В данной работе авторы использовали подход на основе анализа эмоциональной окрашенности для выявления научно-практических статей, содержащих негативные тональности.

Авторами рассмотрены различные подходы к анализу текстов с помощью методов машинного обучения для решения поставленной задачи. Применение рекуррентной нейронной сети позволило авторам осуществить классификацию научно-практических статьей по тональностям с точностью 88 % (метрика accuracy). Разработанная авторами архитектура рекуррентной сети создает модель для извлечения из корпуса текстов научно-практических статей документов, содержащих негативные тональности применительно к заданной технологии. Так же авторами сделан анализ наиболее неуспешных и наиболее успешных применений технологий по всему корпусу текстов.

Ключевые слова: рекуррентные нейронные сети; анализ тональности текста; автоматизированный анализ текстов; наукометрия; векторная модель текста; бизнес разведка

Введение

Любая разведка в интересах бизнеса заинтересована использовать наиболее технологичные и эффективные инструменты. Бытовым примером, подтверждающим этот тезис, служат книги и фильмы про «Агента 007», который всегда пользуется современными достижениями технологий для выполнения заданий. Интерес к высокотехнологичным средствам анализа есть у многих компаний, но не все обладают достаточными ресурсами.

Использование искусственных нейронных сетей для анализа текстов получило развитие в середине 90-х годов в работах [1, 2, 3]. Но из-за высоких требований к вычислительным ресурсам для обучения нейронных сетей оставалось академической дисциплиной.

Ускорение исследований в этом направление связано с ростом скорости вычислений и с появлением таких новых архитектур искусственных нейронных сетей как сверточные нейронные сети [4] и рекуррентные нейронные сети [5].

Для обучения нейронных сетей всегда были нужны значительные массивы размеченных данных. А с появлением большего количества слоев с нейронами потребность в размеченных данных выросла в разы. Для примера, чтобы обучить искусственную нейронную сеть со ста тысячью коэффициентами нужны десятки тысяч размеченных текстов. А для архитектуры глубоких нейронных сетей количество обучаемых коэффициентов составляет миллионы [6].

Поэтому обучение искусственной нейронной сети на собственных данных означает выделение определенного времени и ресурсов на разметку. Другими словами, каждый классифицируемый объект человек должен отнести к одному из классов «вручную».

Не так давно появились размеченные корпусы текстов с открытым доступом, например, и UMICH SI650[1], TreeBank[2], Twitter Sentiments[3], MPQA Opinion Corpus[4] и работы по их анализу [7, 8, 9].

Анализ тональности текста предназначен для выявления в текстах эмоционально окрашенной лексики. Иногда исследователи выделяют термин *Opinion mining* [10], подчеркивая тем самым задачу поиска в текстах оценочных суждений. Кроме академического изучения тональности текста как одного из разделов компьютерной лингвистики [11] существует ряд прикладных исследований, направленных на улучшение процесса принятия управленческих решений [12, 13].

Применение рекуррентных и сверточных нейронных сетей для анализа тональностей [14, 15] позволило достичь значительно большей точности по сравнению с моделями, основанными логистической регрессией.

С учетом всего вышеописанного авторами выдвинута следующая исследовательская гипотеза.

Исследовательская гипотеза: Существует оптимальная архитектура и набор гиперпараметров нейронной сети, позволяющие обучить классификационную модель на публичном наборе данных, содержащем оценочные суждения, и затем предсказать фрагменты текста из научно-практических статей, содержащие оценочные суждения с заданной степенью точности.

---

[1] https://www.kaggle.com/c/si650winter11.

[2] http://nlp.stanford.edu/sentiment/treebank.html.

[3] http://www.sananalytics.com/lab/twitter-sentiment/.

[4] http://mpqa.cs.pitt.edu.

Данное исследование состоит из введения, методического описания подходов авторов к решению задачи классификации текстов по тональности и цифрового эксперимента с разработанным авторами набором данных научно-практических статей.

Методика

Примененные авторами методические подходы могут быть представлены в следующем методическом каркасе исследования (рисунок 1). Рассмотрим более подробно каждый из элементов методического каркаса.

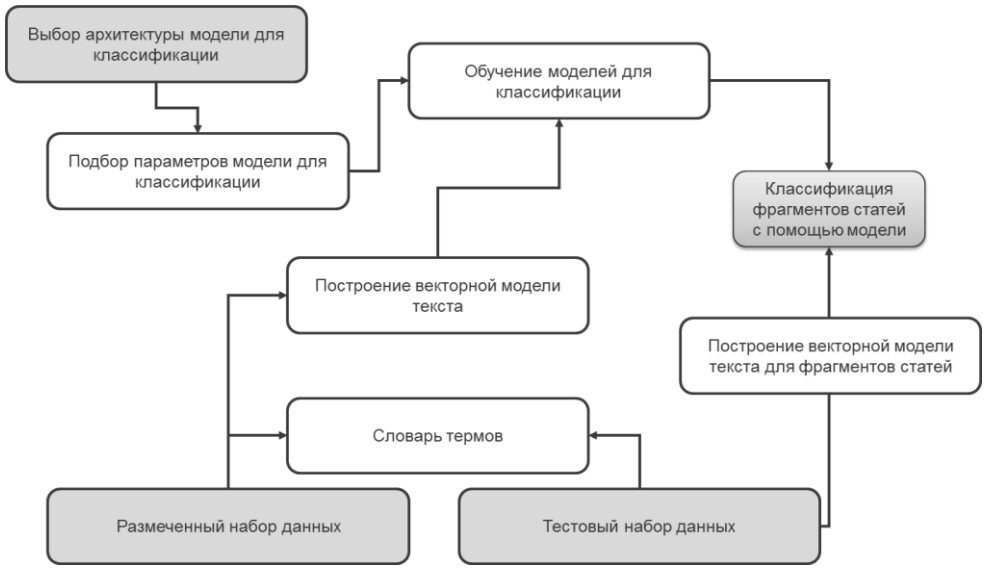

Рисунок 1. *Методический каркас исследования (разработан авторами)*

Наборы данных и словарь

В качестве размеченного набора данных авторами был выбраны отзывы о кинофильмах [7]. В этом наборе данных содержится 25 тысяч положительных и 25 тысяч отрицательных отзывов. Набор данных таким образом сбалансирован для обучения и валидации модели классификации. Длинна отзывов варьируется от 5 до 977 слов и отображена на рисунке (рисунок 2).

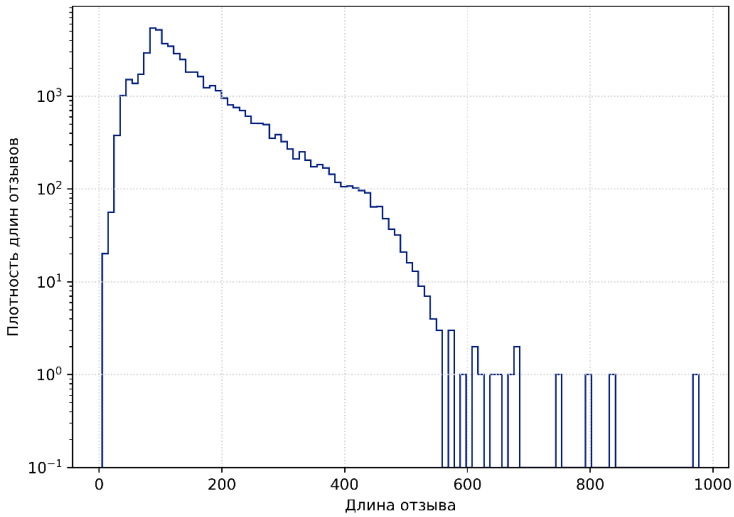

Рисунок 2. *Распределение длин отзывов (разработан авторами)*

43ITVN318
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

При составлении словаря по отзывам были отброшены низкочастотные слова, то есть слова, встречающиеся в документах редко. Распределение частот слов по документам отображено на диаграмме (рисунок 3).

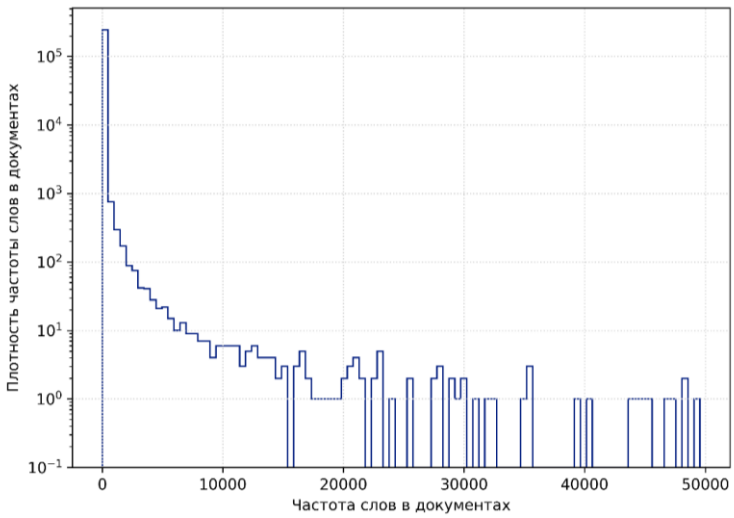

Рисунок 3. *Распределение частот слов по документам (разработан авторами)*

В качестве набора данных для тестирования были выбраны 1696 научно-практических статей с портала OnePetro.org.

Векторная модель текста

Для построения векторной модели текста была использована обученная модель GloVe [16]. Были использованы вектора с размерностью 100 и 300. Преимущества от использования обученной векторной модели текста состоит в существенном сокращении объема вычислений. Количество обучаемых параметров для создания векторной модели текста в разы превосходит количество параметров для выбранных авторами архитектур моделей классификации.

Выбор архитектуры модели для классификации

Автор ограничил себя классом моделей, построенных на основе искусственных нейронных сетей. Среди архитектур искусственных нейронных сетей используемых для классификации текстов можно выделить CNN-LSTM [17, 18] и Stacked LSTM [19, 20].

Авторами были выбраны следующие три варианта архитектур моделей для классификации с использованием искусственных нейронных сетей.

1. Рекуррентная нейронная сеть из одного слоя LSTM. Далее будем называть эту архитектуру RNN и отдельно указывать количество элементов в LSTM слое.
2. Сверточная нейронная сеть из одного слоя Dropout-Conv1D-Conv1D-MaxPooling и рекуррентная нейронная сеть из одного слоя элементов LSTM. Далее будем называть эту архитектуру CNN-LSTM и отдельно указывать количество элементов и параметры сверточных слоев.
3. Рекуррентная нейронная сеть из двух слоев LSTM. Далее будем называть эту архитектуру RNN-2 и отдельно указывать количество элементов в LSTM слое.

Поиск оптимальных параметров модели для классификации

Для рассматриваемых архитектур моделей классификации автор выбрал следующие существенные гиперпараметры:

1. Тип модели классификации: RNN, CNN-LSTM, RNN-2.
2. Размерность словаря. В зависимости от фильтров низкочастотных слов размерность словаря изменялась от 2000 до 200000 слов.
3. Размерность векторной модели текста: 100 и 300.
4. Длинна фрагмента текста: 80, 128 и 196.

Обучение моделей для классификации

Обучение моделей классификации производилось параллельно на нескольких серверах. Набор данных содержал равное количество положительный и отрицательных отзывов поэтому для оценки качества обучения была выбрана метрика *Accuracy*. Оптимизация параметров модели для классификации производилась на основании функции перекрестной энтропии (*Cross Entropy*).

Для ускорения обучения авторами был применен метод ранней остановки обучения на основании метрики *Accuracy* по валидационному набору данных. Кривые обучения для модели классификации типа RNN отображены на рисунках (рисунок 4, рисунок 5).

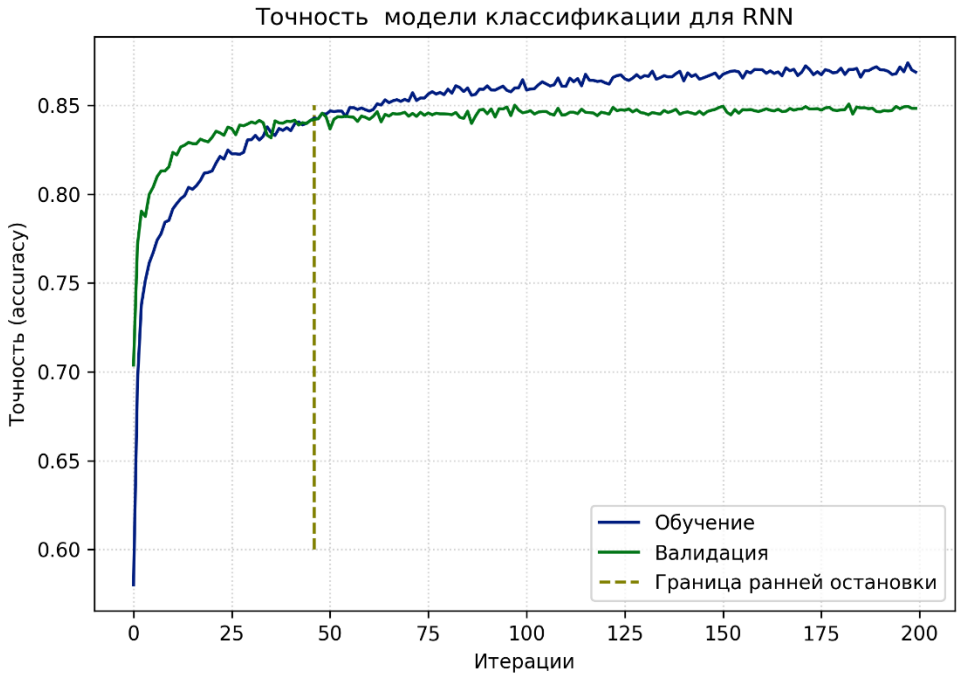

Рисунок 4. *Кривая обучения метрики Accuracy для модели классификации типа RNN (разработан авторами)*

Из зависимости метрик *Accuracy* для тренировочного и валидационного набора данных (рисунок 4) видно, что в районе 42 итерации обучения метрика *Accuracy* перестает увеличиваться для валидационного набора данных. Это явление означает, что модель начинает переучиваться по метрике *Accuracy* и обучение следует остановить. Данная архитектура модели для классификации не позволяет повышать точность на этом наборе данных.

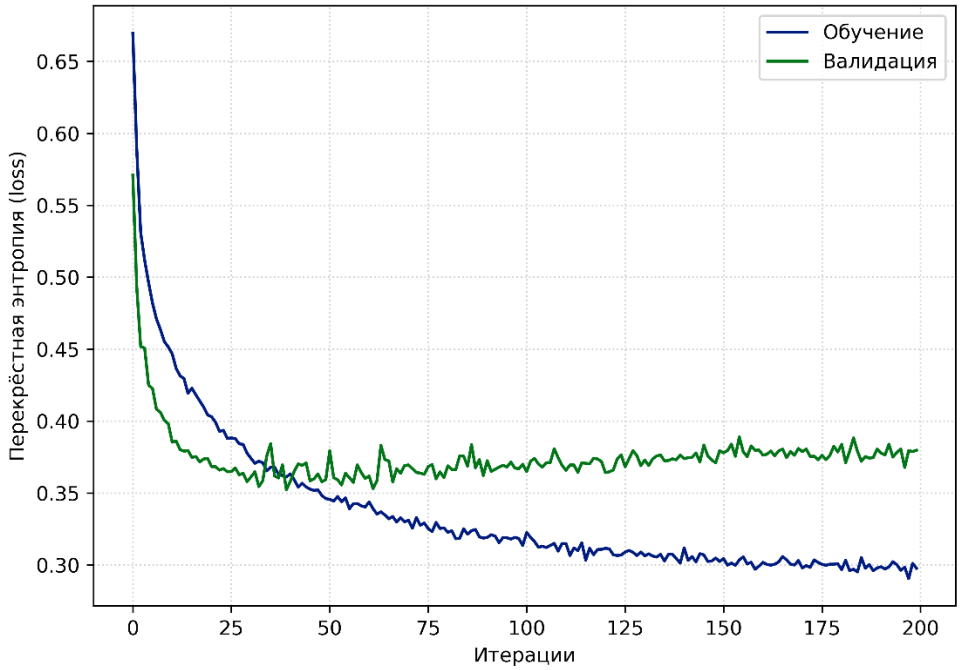

Рисунок 5. *Кривая функции потерь для модели классификации типа RNN (разработан авторами)*

Отметим, что из зависимости, отображенной на рисунок 5 видно, что значение перекрестной энтропии на валидационном наборе данных начинает не убывать в районе 37 итерации. То есть, немногим раньше, чем начинает деградировать метрика *Accuracy*.

Эксперимент

На основании изложенных выше методик было проведено обучение моделей классификации с различными архитектурами и гиперпараметрами. Результаты обучения приведены в таблице (таблица 1).

Таблица 1

Результаты обучения моделей классификации с различными гиперпараметрами

Архитектура модели	Количество параметров, тыс.	Длинна фрагмента текста	Размерность векторного пространства текста	Словарь, количество слов	Точность валидации
CNN+RNN	63	128	100	2 300	0,85
CNN+RNN	63	196	100	2 300	0,87
CNN+RNN	63	196	100	23 400	0,86
RNN	69	128	100	2 300	0,87
RNN	722	196	300	23 400	0,88
RNN	81	128	100	47 969	0,85
RNN-2	161	128	100	2 300	0,86
RNN-2	161	196	100	2 300	0,87
RNN-2	1 443	196	300	23 000	0,87
RNN-2	1 443	196	300	248 739	0,85
RNN-2	1 443	80	300	248 739	0,85

Разработано авторами

Лучшее значение метрики *Accuracy* на валидационном наборе данных показала модель RNN со словарем из 23 тысяч слов и размерностью векторной модели текста равной 300.

Отметим, что на тестовом наборе данных значение метрики *Accuracy* для данной модели составило 88 %.

Полученная модель была использована для предсказания тональности научных статей портала OnePetro.org. Каждая научная статья разбивалась на фрагменты длинной 196 слов для оценки эмоциональной окраски. Затем фрагменты статей собирались обратно для получения эмоциональной карты всей статьи. Таким образом, можно было определить фрагменты статьи, обладающие аномальными эмоциональными окрасками, такими как разочарование и удовлетворенность.

Данное исследование не принимает в расчет семантику текста, поэтому предмет эмоциональной окраски автоматически не определялся. Выбранные фрагменты статьи необходимо проанализировать с помощью эксперта. Но такой подход к аннотированию статьи позволил найти сложно обнаруживаемые фрагменты. В таблица 2 приведены примеры эмоциональных фрагментов статей.

Таблица 2

Эмоциональные фрагменты статей

the results from pilot tests which were using as injectant are disappointing and the results from pilot tests which were using natural gases are encouraging’
to sum up diffusion mechanism for in pilot tests had not been well recognized which in turn did not enhance oil production rate in those wells
the outstanding result from this study
using the other forward model result dramatically bad

Так же автор разработал цветовое представление эмоциональной окраски статей в зависимости от вероятности отнесения фрагмента текста к положительной или отрицательной эмоциональной окраске.

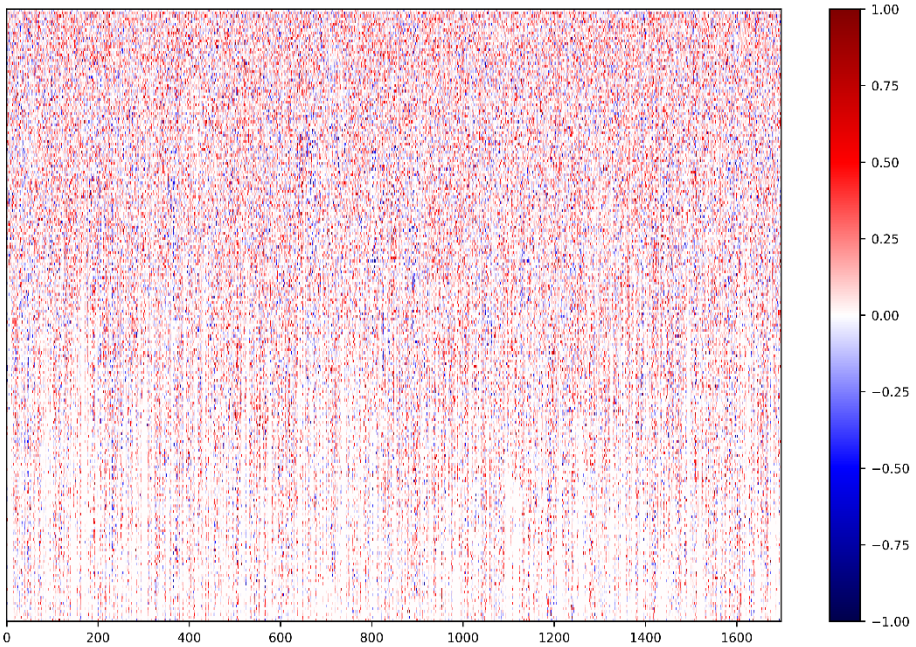

Рисунок 6. *Карта полярности эмоциональной окраски статей 1696 статей. По оси x отложен порядковый номер статьи, по оси y эмоциональная окраска фрагментов статьи. На цветовой шкале отображена цифровая характеристика эмоциональности: негативная (-1), позитивная (+1) (разработан авторами)*

43ITVN318
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

На рисунке (рисунок 6) эмоциональность фрагментов статей отображена в виде карты. Для каждой статьи на оси X цветом отображена эмоциональность каждого фрагмента последовательно по оси Y.

Заключение

Полученные результаты подтверждают гипотезу о возможности выделения эмоционально-окрашенных фрагментов текста из научных статей. Научные статьи используют академическую лексику и ожидать в них градус эмоций сравнимый с отзывами на кинофильмы было бы наивно. Но современные концепции обработки текста, основанные на анализе контексте, позволяют выделять и классифицировать изменения эмоциональности достаточно точно для того, чтобы обрабатывать даже научные статьи. Автор считает, что проведенное исследование открывает возможности по созданию дополнительных инструментов для аннотации и классификации научных текстов.

Наилучшее оценку по качеству выделения эмоционально окрашенных фрагментов текста показали рекуррентные нейронные сети. Точность по метрике *Accuracy для* них составила 88 %. Важно отметить, что по скорости обучения рекуррентные сети существенно проигрывают сверточным сетям. Автор видит объяснение разности в производительности в том, что для обучения сверточных нейронных сетей возможна более высокая степень параллельных вычислений. Тогда как для рекуррентных нейронных сетей необходимо поддерживать последовательность предыдущих состояний нейронов.

В дальнейших исследованиях автор планирует исследовать применимость эмоционально окрашенных фрагментов текста для задач классификации текстов в качестве признаков. Так же на взгляд авторов, научный интерес представляет анализ синтаксиса эмоционально окрашенных фрагментов текста.

ЛИТЕРАТУРА

1. Ng H.T., Goh W.B., Low K.L. Feature selection, perceptron learning, and a usability case study for text categorization // ACM SIGIR Forum. – ACM, 1997. – Т. 31. – №. SI. – С. 67-73.
2. Lam S.L.Y., Lee D.L. Feature reduction for neural network based text categorization // Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on. – IEEE, 1999. – С. 195-202.
3. Lam W., Ruiz M., Srinivasan P. Automatic text categorization and its application to text retrieval // IEEE Transactions on Knowledge and Data engineering. – 1999. – Т. 11. – №. 6. – С. 865-879.
4. Kim Y. Convolutional neural networks for sentence classification // arXiv preprint arXiv:1408.5882. – 2014.
5. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. Recurrent neural network based language model // Eleventh Annual Conference of the International Speech Communication Association. – 2010.
6. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. – 2012. – С. 1097-1105.

43ITVN318
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

7. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C. Learning word vectors for sentiment analysis // Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. – Association for Computational Linguistics, 2011. – С. 142-150.
8. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank // Proceedings of the 2013 conference on empirical methods in natural language processing. – 2013. – С. 1631-1642.
9. Akkaya C., Wiebe J., Mihalcea R. Subjectivity word sense disambiguation // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. – Association for Computational Linguistics, 2009. – С. 190-199.
10. Zhao J., Liu K., Xu L. Sentiment analysis: mining opinions, sentiments, and emotions. – 2016.
11. Shanahan J.G., Qu Y., Wiebe J. (ed.). Computing attitude and affect in text: Theory and applications. – Dordrecht: Springer, 2006. – Т. 20.
12. Hu N., Pavlou P.A., Zhang J. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication // Proceedings of the 7th ACM conference on Electronic commerce. – ACM, 2006. – С. 324-330.
13. Archak N., Ghose A., Ipeirotis P.G. Show me the money!: deriving the pricing power of product features by mining consumer reviews // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2007. – С. 56-65.
14. Chen, T., Xu, R., He, Y., Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN // Expert Systems with Applications. – 2017. – Т. 72. – С. 221-230.
15. Tai K.S., Socher R., Manning C.D. Improved semantic representations from tree-structured long short-term memory networks // arXiv preprint arXiv:1503.00075. – 2015.
16. Pennington J., Socher R., Manning C. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
17. Wang, J., Yu, L.C., Lai, K.R., Zhang, X. Dimensional sentiment analysis using a regional CNN-LSTM model // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2016. – Т. 2. – С. 225-230.
18. Tang D., Qin B., Liu T. Document modeling with gated recurrent neural network for sentiment classification // Proceedings of the 2015 conference on empirical methods in natural language processing. – 2015. – С. 1422-1432.
19. Ruder S., Ghaffari P., Breslin J.G. A hierarchical model of reviews for aspect-based sentiment analysis // arXiv preprint arXiv:1609.02745. – 2016.
20. Cheng J., Dong L., Lapata M. Long short-term memory-networks for machine reading // arXiv preprint arXiv:1601.06733. – 2016.

Krasnov Fedor Vladimirovich
LLC "Gazpromneft NTC", Saint-Petersburg, Russia
E-mail: Krasnov.FV@Gazprom-Neft.ru

Analysis of the tone of the text of scientific and practical papers on oil and gas topics using artificial neural networks

Abstract. The problem of finding negative practical experience from the use of new technologies is acute in the oil and gas industry. The risk of repeating the mistakes of others is high due to the variety of technologies and the uncertainty of their application limits. The most complete source of experience is the materials of scientific and practical conferences. Moreover, the most accurate method of finding a negative practical experience is to carefully read the text of each presentation and seek out the positive and negative experience from the use of technology. However, this method is difficult to apply in practice, as the volume of texts is huge. In this paper, the authors used an approach based on the analysis of emotional color to identify scientific and practical articles containing negative tonalities.

The authors consider various approaches to the text analysis using machine learning methods to solve the problem. The use of recurrent neural networks allowed the authors to carry out the classification of scientific and practical article on the keys with accuracy of 88 % (metric accuracy). The architecture of the recurrent network developed by the authors creates a model for extracting documents containing negative tonalities in relation to a given technology from the body of scientific and practical articles. In addition, the authors analyzed the most unsuccessful and most successful applications of technologies throughout the body of texts.

Keywords: recurrent neural networks; sentiment analysis; automated text analysis; scientometrics; vector text model; business intelligence