ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА "МИФИ", 2020, том 9, № 2, с. 189–196

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.032.26

ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ МОДЕЛИ ОПРЕДЕЛЕНИЯ ТИПА ИМИТАЦИИ ВОЗРАСТА В ТЕКСТЕ[1]

© 2020 г. А. Г. Сбоев[1,2,*], И. А. Молошников[1], Р. Б. Рыбка[1], А. В. Наумов[1]

[1] *Национальный исследовательский центр "Курчатовский институт", Москва, 123182, Россия*

[2] *Национальный исследовательский ядерный университет "МИФИ", Москва, 115409, Россия*

**e-mail: sag111@mail.ru*

Поступила в редакцию 19.12.2019 г.
После доработки 19.12.2019 г.
Принята к публикации 10.03.2020 г.

В данной работе рассмотрен метод интерпретации результатов определения типа имитации возраста в тексте для модели на основе управляемого рекуррентного блока (Gated Recurrent Unit, GRU), с дополнительным слоем сети, отражающим активности скрытого слоя для каждого слова в тексте. При этом использовался собранный с помощью краудсорсинговой платформы корпус для задачи определения типа имитации возраста в тексте. Корпус содержит три типа текстов: тексты, написанные в естественном стиле автора, тексты с имитацией стиля младшего возраста, тексты с имитацией стиля старшего возраста. Тексты были представлены в сегментированном на слова и предложения виде, а также произведен их морфологический разбор и лемматизация с использованием программы UDPipe. Топология сети включает: внутренний двунаправленный слой GRU размерностью 32, выходом слоя являются активности для каждого слова документа, передаваемые в полносвязанный слой с активационной функцией ReLU размерностью 32 и в еще один полносвязанный слой с активационной функцией гиперболический тангенс, размерностью 3, по числу классов имитации возраста. Дополнительный интерпретирующий слой возвращает коэффициенты, определяющие к какому классу относится текст. По результатам проведенного анализа экспериментов выявлено, что характерными признаками для определения типа имитации возраста в тексте является то, как человек начинает этот текст и какое приветствие использует.

Ключевые слова: интерпретация результатов, искусственные нейронные сети, обработка естественного языка, классификация текстов, авторское профилирование, имитация стиля другого возраста в тексте

DOI: 10.1134/S2304487X20020121

1. ВВЕДЕНИЕ

Вопрос интерпретируемости результатов нейросетевых моделей в последнее время стал носить все более острый характер, в особой степени благодаря более широкому использованию сетей глубокого обучения, которые отличаются высокой эффективностью, но слабой интерпретируемостью получаемых результатов.

Поиски компромисса между точностью модели и интерпретируемостью данных, полученных с ее помощью приводят зачастую к выбору менее эффективного, но более интерпретируемого метода.

Интерперетация результатов классификации достаточно распространена в задачах анализа изображений.

В работе [4] авторы предлагают методику создания "визуальных объяснений" (выделения областей на изображении значимых для определенного класса) на основе сверточных нейронных сетей (CNN).

Этот подход использует градиенты целевого класса, вычисляемые в последнем сверточном слое для получения грубой карты локализации, выделяющей важные области в изображении для прогнозирования этого класса.

Другой подход для интерпретации решений, предложенный в работе [2] основан на работе с данными и уже готовыми обученными моделями.

Авторы предлагают итеративно вносить изменения в оригинальные данные и, анализируя предсказания модели, смотреть, как эти измене-

[1] Работа была выполнена с использованием оборудования центра коллективного пользования "Комплекс моделирования и обработки данных исследовательских установок мега-класса" НИЦ "Курчатовский институт", http://ckp.nrcki.ru/

ния влияют на правильность предсказания класса.

По сути это решения задачи "черного ящика": подавая части данных исходной модели на ввод и получая от нее ответ, можно понять, какие именно части отвечают за конечный результат, и за счет этого получить интерпретацию решения.

Данный подход дает хорошие результаты для интерпретации данных фиксированной длины – это могут быть задачи классификации изображений или текстов.

При этом тексты должны быть представлены как фиксированное множество слов, что снижает качество интерпретируемости, так как одни и те же слова в документе в разных контекстах могут отражать разные классы.

Второй проблемой данного метода является необходимость большого числа запусков модели на измененных данных, что накладывает дополнительные ограничения на вычислительные мощности.

В данной работе предлагается подход для интерпретации результатов задачи классификации текста с возможностью указать вклад каждого слова или последовательности слов произвольной длины по отдельному классу.

В отличие от вышеприведенных методов, данный подход позволяет интерпретировать результаты классификации за счет модификации архитектуры нейронной сети.

Анализируя результаты интерпретации легче выделить закономерности, отраженные в данных по конкретной задаче.

В работе приводится пример анализа задачи классификации текстов по трем классам: тексты, написанные в естественном стиле автора, попытка имитировать стиль младшего возраста, попытка имитировать стиль старшего возраста.

Анализ с использованием предложенного метода показал, что характерной чертой, определяющей класс имитации, является приветствие в тексте.

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Краткое описание метода

В данной работе использовалась модель на основе слоев GRU (Управляемый рекуррентный блок, Gated Recurrent Unit [1]). Для классификации текст подается в виде последовательности слов с различными морфологическими признаками. Классификация производится по сумме активностей скрытого слоя за каждый класс, размерность выхода скрытого слоя равна числу классов.

GRU – это одна из разновидностей рекуррентных сетей с использованием механизмов вентилей (gate). В отличие от аналогичного механизма в сетях с долгосрочной кратковременной памятью (LSTM), данная сеть имеет меньше параметров для обучения и у нее отсутствует отдельное скрытое состояние (ячейка памяти как в LSTM).

То есть активности сети для соседних слов будут более взаимозависимы, чем в LSTM, это улучшает возможности сети по интерпретации результатов.

2.2. Предварительная обработка данных

Тексты были сегментированы на слова и предложения, а также произведен морфологический разбор и лемматизация с использованием программы UDPipe [5].

В модель текст подается в виде последовательности слов.

Для кодирования слов используется посимвольное представление словоформы, уникальное ID леммы, ID словоформы, часть речи, полный морфологический тэг.

Для кодирования символов слова, леммы, словоформы и часть речи используется уникальный ID, который представлен в виде уникального порядкового номера для каждого значения (в модели эти номера передаются в слой вложения – Embedding), полный морфологический тэг представлен в виде конкатенации бинарных представлений для каждой группы (род, число, падеж, и т.д.) (OneHotEncoding).

2.3. Топология сети

Топологию сети можно разделить на 3 уровня анализа:

- Вход сети – отвечает за анализ каждого слова в документе;
- Внутренний слой (скрытый слой) – возвращает активности на каждое слово для каждого класса;
- Выходной слой – производит суммирование и нормализацию активностей всех слов в документе для предсказания типа имитации возраста.

Схема топологии представлена на рисунке 1.

ВХОД СЕТИ

Для обработки посимвольного представления словоформы каждый символ слова подается в слой embedding размерностью 5 (embedding-слой представляет собой совмещение бинарного кодирования (OneHotEncoding) и линейного полносвязного слоя, что позволяет хранить исходные данные в сжатом формате индекса). После embedding-слоя используется двунаправленный слой GRU, размерностью 5 нейронов для каждого

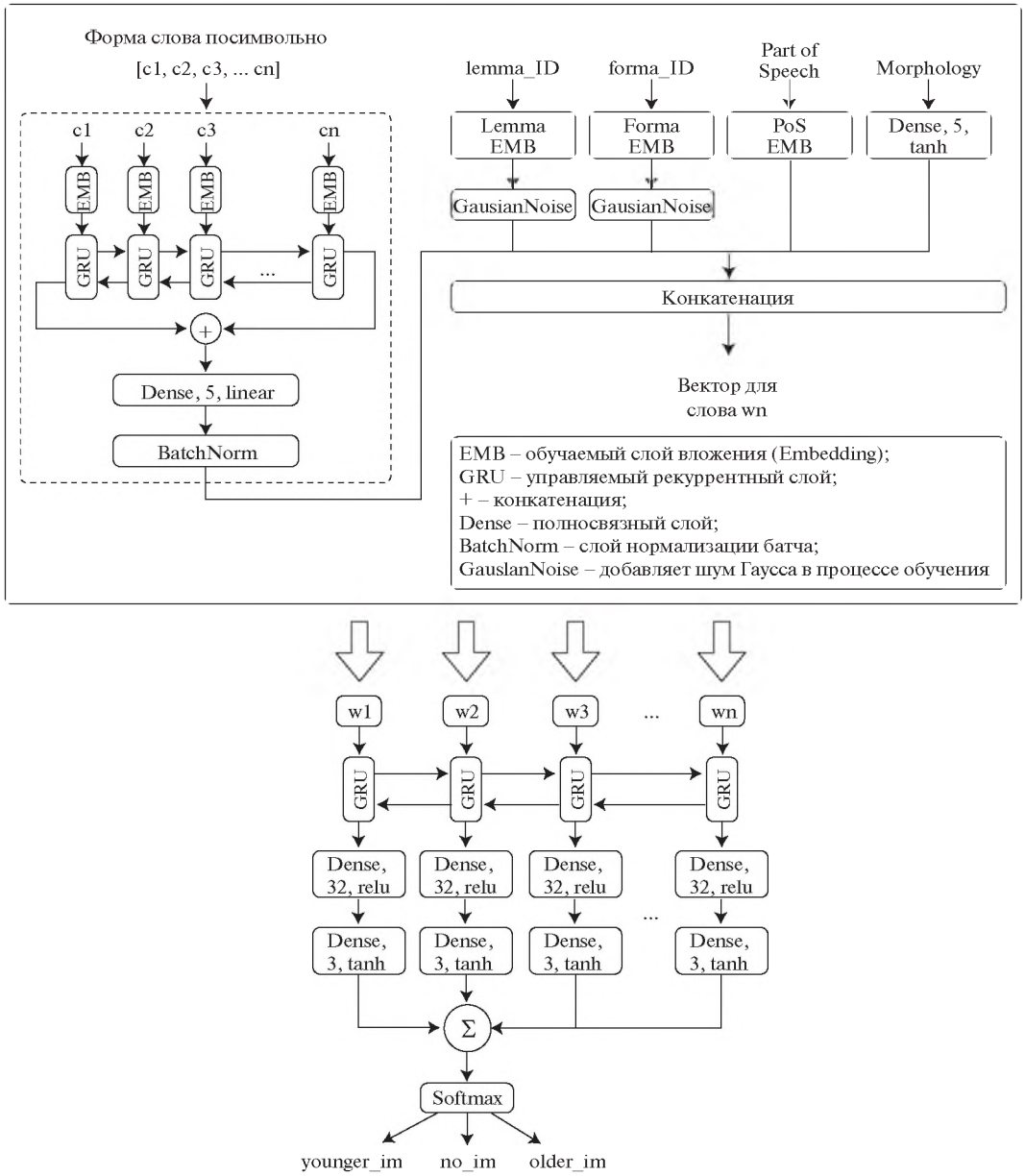

Рис. 1. Топология нейронной сети.

направления. Выходом слоя является вектор для слова.

Далее этот вектор подается в полносвязанный слой размерностью 5 с линейной функцией активации. К выходу этого слоя применяется нормализация батча (BatchNormalization – нормализует активности, таким образом, чтобы активность слоя в батче имела среднее значение 0 со стандартным отклонением, равным 1).

Индексы словоформы и леммы подаются в соответствующие им embedding-слои (на рисунке 1 это входы forma_ID и lemma_ID).

Далее в процессе обучения к закодированным представлениям для лемм и словоформ добавляется гауссов шум (additive zero-centered Gaussian noise) со значением стандартного отклонения, равным 0.8. Использование шума позволяет сделать так, чтобы сеть опиралась на дополнительные признаки, такие как морфология и посимвольное представление, а не “учила” словарь слов.

Дополнительно при кодировании слова используется часть речи, закодированная с использованием слоя embdedding размерностью 5 и полный морфологический тэг. Для кодирования морфологического тэга используется бинарное кодирование (OneHotEncoding) в рамках каждой группы (род, число, падеж и т.д.), далее закодированный морфологический тэг подается в полно-

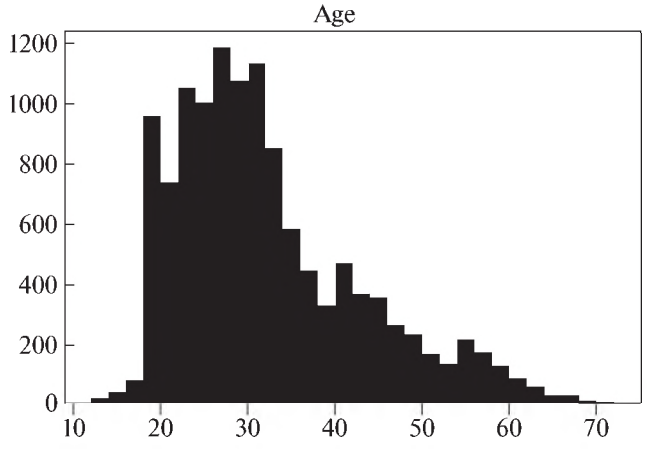

Рис. 2. Гистограмма распределения возрастов авторов по корпусу текстов.

связанный слой с активационной функцией гиперболический тангенс (tanh) размерностью 5.

В итоге все вектора (посимвольное кодирование, embedding для словоформы и леммы, часть речи, морфология) конкатенируются – это и есть закодированное представление слова.

ВНУТРЕННИЙ СЛОЙ (СКРЫТЫЙ СЛОЙ)

Далее для каждого слова необходимо получить активности по каждому из 3-х классов. Для этого закодированное представление подается в двунаправленный слой GRU размерностью 32, выходом слоя являются активности для каждого слова документа.

Далее каждая активность передается в полносвязанный слой с активационной функцией ReLU размерностью 32 и в еще один полносвязанный слой (word_hidden на схеме) с активационной функцией гиперболический тангенс (tanh), размерностью 3, по числу классов имитации возраста.

ВЫХОДНОЙ СЛОЙ

Для интерпретации результатов в сеть был добавлен дополнительный слой “word_hidden”, отражающий активности скрытого слоя для каждого слова. Слой “word_hidden” возвращает коэффициенты для каждого слова в тексте относительно трех классов (без имитации, имитация старшего возраста, имитация младшего возраста), после суммирования всех коэффициентов слов для каждого класса определяется, к какому классу относится текст. Тем самым, получая коэфициент для каждого слова, мы можем понять, почему модель отнесла тот или иной текст к определенному классу.

ПАРАМЕТРЫ ОБУЧЕНИЯ

При обучении использовался метод оптимизации Adam [2] со значением learning rate, равным 0.003, и функцией ошибки categorical crossentropy, с использованием раннего останова по валидационному множеству (ранний останов после 32 эпох с момента, когда ошибка на валидационном множестве стала расти).

После раннего останова загружаются веса на момент наименьшей ошибки на валидационном множестве, размер батча 8.

3. РЕЗУЛЬТАТЫ

3.1. Описание корпуса

В работе использовался корпус для задачи определения типа имитации возраста в тексте. Для сбора корпуса использовалась краудсорсинговая платформа.

Каждый респондент мог выполнить одно или более заданий, задание для респондента выбиралось случайным образом:

- “Расскажите о себе или не о себе, но главное попытайтесь понравиться своему собеседнику и добиться его расположения”;
- “Расскажите о запомнившемся событии/приобретении/слухах или еще что-то интересное, так чтобы понравилось вашему собеседнику”;
- “Попробуйте уговорить вашего собеседника (это может быть, кто угодно) встретиться на вашей территории (где угодно)”.

Распределение возрастов авторов по корпусу приведено на рисунке 2.

В каждом задании необходимо было написать 3 текста: от своей возрастной группы, попытаться имитировать стиль человека старшего возраста, попытаться имитировать стиль человека младшего возраста. Исследования проводились на корпусе размером 10014 документов.

Корпус в равном количестве для каждого класса состоит из текстов без имитации возраста, с попыткой имитации более старшего возраста и более младшего.

Для оценки классификации множество было разбито на 3 части: тренировочное (60% 6027 текстов), валидационное (~20%, 1926 текстов) и тестовое (~20%, 2061 текстов). Разбиение проводилось по авторам.

3.2. Точность классификации

Результаты исследования представлены в таблице 1.

Таблица 1. Результаты, оценка определения типа имитации на тестовом множестве

	Precision	Recall	F-score
Класс “no_im”	0.74	0.64	0.68
Класс “older”	0.73	0.89	0.80
Класс “younger”	0.89	0.82	0.85
В среднем	0.78	0.78	0.78

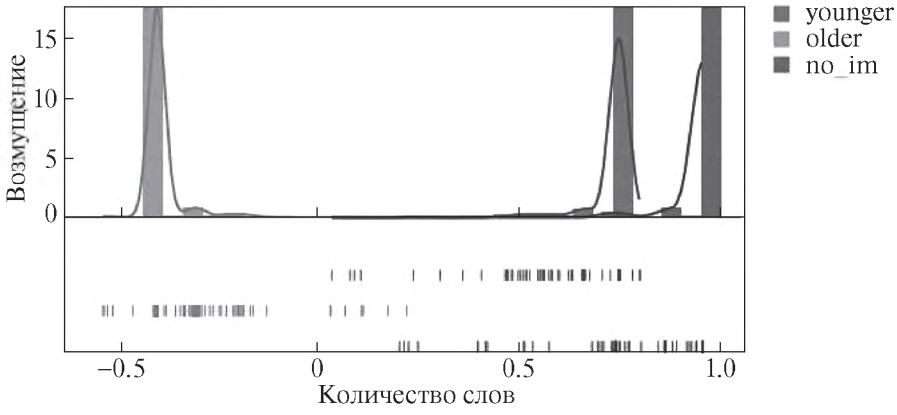

Рис. 3. Распределение коэффициентов для леммы “привет”.

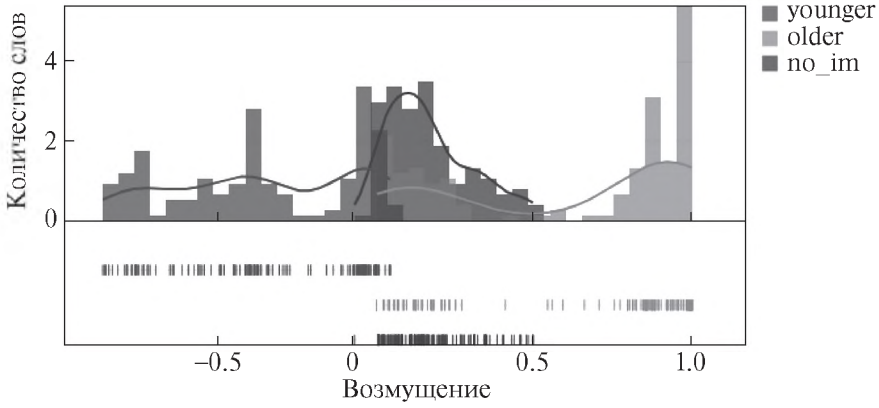

Рис. 4. Распределение коэффициентов для леммы “дорогой”.

Средняя оценка по F1-метрике составляет 0.78, лучший результат сеть показала в определении имитации более молодого возраста в тексте – 0.85.

3.3. Интерпретация

Для интерпретации используется активность скрытого слоя для каждого класса. Активность скрытого слоя используется для интерпретации отношения каждого слова к заданным классам. Для каждого слова в рамках тестового множества был получен список его активностей в каждом документе.

По данному списку рассчитаны статистические характеристики, такие как среднее, медиана и стандартное отклонение и построена гистограмма активности за тот или иной класс.

Как видно из графиков на рис. 3 и 4 одна и таже лемма может получать разные значения коэффициентов в зависимости от контекста.

Однако, несмотря на это, можно выделить некоторые слова, которые обычно дают наибольший вклад в сторону того или иного класса.

Например, слова “здравствуйте”, “дорогой”, “добрый” были определены сетью как слова, свойственные текстам, имитирующим старший возраст, то есть имеющим больший коэффициент для класса “older” и меньший для “younger”; а такие слова как “хай”, “прикольный”, “бро” вносят больший вклад в сторону класса “younger”.

Стоит отметить, что паттерны сформировавшиеся у сети, в целом, верные: очевидно, что пытаясь казаться старше, люди используют более формальную речь, в свою очередь, пытаясь имитировать молодой возраст, люди чаще используют неформальную речь и сленг.

В табл. 2 представлен список из 20 характерных слов для каждого класса и приведена средняя по всему тестовому множеству активность нейрона соответствующего класса (no_im, younger, older). В таблицу отбирались леммы, встречающиеся 10 и более раз в тексте и имеющие максимальную активность для класса (положительную или отрицательную).

Активность отражает значимость слова для данного класса. Положительная активность по-

Таблица 2. Слова, имеющие максимальный и минимальный вклад для классификации текстов за каждый класс

Лемма	“no_im”	Лемма	“younger”	Лемма	“older”
привет	0.93	привет	0.73	здравствовать	0.91
привать	0.69	Приветик	0.68	приветствовать	0.89
прить	0.65	приветик	0.60	добрый	0.80
приветствовать	0.35	Хай	0.59	дорогой	0.62
дмитрий	0.31	Хать	0.57	Николаевна	0.60
здравствовать	0.37	Здаров	0.48	иванович	0.58
юля	0.24	прить	0.48	сутки	0.58
кофе	0.27	здоров	0.41	милый	0.57
прекрасный	0.25	прикинь	0.39	уважаемый	0.54
ольга	0.25	привать	0.34	здоровье	0.50
крутый	–0.43	юный	–0.13	крутый	–0.47
родитель	–0.45	Николаевна	–0.18	крутой	–0.48
крутой	–0.45	дорогий	–0.23	Машка	–0.50
бро	–0.51	милый	–0.24	предок	–0.55
делишки	–0.53	здоровье	–0.32	делишки	–0.55
прикинь	–0.54	сутки	–0.33	прикинь	–0.61
приветик	–0.55	дорогой	–0.35	Хать	–0.65
предок	–0.56	добрый	–0.49	Хай	–0.66
Хай	–0.64	приветствовать	–0.50	приветик	–0.73
Хать	–0.67	здравствовать	–0.54	Приветик	–0.85

Таблица 3. Список словосочетаний различной длины со средними активностями для класса без имитации возраста

Комбинация слов	“no_im”
Привет .	0.91
Привет ! Сегодня	0.84
Привет ! Мы	0.84
Привет Всем	0.82
Привет ! Я	0.81
родители ?	–0.53
Хай , бро . Давно не	–0.53
Хай ! Как твои	–0.53
Хай , бро ! Тыщщу	–0.54
Хай бро , короче есть	–0.57

Таблица 4. Список словосочетаний различной длины со средними активностями для класса younger

Комбинация слов	“younger”
Привет мой	0.70
Привет Маша я	0.64
Привет .	0.61
Привет ,	0.60
Привет ! Я	0.57
Здравствуй , дорогой мой	–0.61
Здравствуй дорогая !	–0.62
Здравствуй , дорогая моя	–0.63
Здравствуй мой дорогой товарищ	–0.63
Здравствуй моя дорогая подруга	–0.66

казывает, что слово хорошо характеризует класс, отрицательная говорит, что слово максимально не соответствует классу.

Можно увидеть, что активность для одного и того же слова может быть разной для каждого класса.

Пример: слово “привет” имеет положительную активность 0.93 за класс “без имитации” и 0.73 за класс “younger” – это говорит, что данное слово характерно для обоих классов. Пример: слово “Хай” имеет отрицательную активность за класс “без имитации” (–0.67) и older (–0.66) и положительную за класс “younger” – это говорит о том, что данное слово является характерным только для одного класс “younger”.

Помимо отдельных слов можно определить значимость последовательностей слов различной длины. Для формирования такого словосочетания берутся слова, идущие последовательно и имеющие активность выше или ниже среднего на значение среднеквадратичного отклонения. При-

Таблица 5. Список словосочетаний различной длины со средними активностями для класса older

Комбинация слов	"older"
Здравствуй , доброго дня	0.93
Здравствуй читатель	0.91
Здравствуй , мой неизвестный друг	0.91
Здравствуй Петровна .	0.91
Добрый времени суток	0.90
Приветик . Ты просила	–0.63
Приветик ! Очень	–0.63
Приветики , как твои	–0.64
приветик ! Что	–0.70
Приветик !	–0.71

меры таких словосочетаний приведены в таблицах 3, 4, 5.

5. ЗАКЛЮЧЕНИЕ

В работе представлена нейронная сеть со специальной топологией, позволяющая легко интерпретировать результаты классификации. С помощью данного метода проведен анализ результатов интерпретации текстов для задачи определения типа имитации возраста. В ходе анализа выявлено, что самой характерной чертой является то, как человек начинает текст и какое приветствие использует.

СПИСОК ЛИТЕРАТУРЫ

1. *Chung J., Gulcehre C., Cho K.H., Bengio Y.* Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
2. *Diederik P., Kingma D.P., Ba J.* Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980, 2014.
3. *Guestrin C., Ribeiro M.T., Singh S.* "why should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938, 2016.
4. *Ramakrishna A.D., Parikh V.D., Ramprasaath D.B., Selvaraju R., Cogswell M.* Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv:1610.02391, 2017.
5. *Straka M., Strakova J.* Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. I*n Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Association for Computational Linguistics. pages 88–99, Vancouver, Canada, August 2017.

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2020, vol. 9, no. 2, pp. 190–197

Neural Net Model to Identify Author Age Imitation with Easy Interpret Results[2]

A. G. Sboev[a,b,#], I. A. Moloshnikov[a], R. B. Rybka[a], and A. V. Naumov[a]

[a] *National Research Center Kurchatov Institute, Moscow, 123182 Russia*

[b] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*

[#] *e-mail: sag111@mail.ru*

Received December 19, 2019; revised December 19, 2019; accepted March 10, 2020

Abstract—A method for interpretation of results of identification age style imitation in the text based on a gated recurrent unit (GRU) with an additional network layer to map the activity of the hidden layer for each word in the text has been proposed. At the same time, the special corpus has been collected for the task of age imitation identification. The corpus contains three types of texts: texts written in the author's natural style, texts with imitation of a younger person style, and texts with imitation of an older person style. The texts have been presented in a segmented form, as words and sentences, and their morphological analysis and lemmatization have been performed using the UDPipe program. The network topology includes: an internal bi-directional GRU layer of 32 neurons providing activity for each word of the document, which is an input of a fully-connected layer with the ReLU activation function and size of 32, which connected to another fully-connected layer with the hyperbolic tangent activation function and 3 neurons (just as the number of age imitation classes). An additional interpretive layer returns the coefficients determining the class to which the text belongs. The analysis of the experiments has revealed that the characteristic features for determining the age imitation type in the text are the beginning and greeting used by a person in the text.

Keywords: result interpretation, artificial neural networks, natural language processing, text classification, author profiling, age style imitation

DOI: 10.1134/S2304487X20020121

[2] This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/

REFERENCES

1. Chung J., Gulcehre C., Cho K.H., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
2. Diederik P., Kingma D.P., Ba J. Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980, 2014.
3. Guestrin C., Ribeiro M.T., Singh S. "why should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938, 2016.
4. Ramakrishna A.D., Parikh V.D., Ramprasaath D.B., Selvaraju R., Cogswell M. Grad-cam: Visual explanations from deep networks via gradient-based localization. arXiv:1610.02391, 2017.
5. Straka M., Strakova J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics. pages 88–99, Vancouver, Canada, August 2017.