ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА “МИФИ”, 2019, том 8, № 6, с. 569–576

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.032.26

МОДЕЛЬ НЕЙРОННОЙ СЕТИ ДЛЯ ВКЛЮЧЕНИЯ СИНТАКСИЧЕСКОЙ СТРУКТУРЫ ПРЕДЛОЖЕНИЯ В ЗАДАЧУ КЛАССИФИКАЦИИ ПОЛА АВТОРА РУССКОГО ТЕКСТА

© 2019 г. А. Г. Сбоев[1,2,*], А. А. Селиванов[1], Р. Б. Рыбка[1], И. А. Молошников[1], Д. С. Богачев[1,3]

[1] Национальный исследовательский центр “Курчатовский институт”, Москва, 123098, Россия

[2] Национальный исследовательский ядерный университет “МИФИ”, Москва, 115409, Россия

[3] Московский физико-технический институт (Национальный исследовательский университет), Москва, 141701, Россия

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

Поступила в редакцию 07.10.2019 г.
После доработки 11.10.2019 г.
Принята к публикации 15.10.2019 г.

В исследовании предлагаются нейросетевые методы на базе глубоких нейронных сетей с слоями долговременной краткосрочной памяти (LSTM) для включения синтаксической структуры в решение задачи классификации русских текстов. Разработаны и опробованы два подхода обработки синтаксических структур. В первом синтаксическая структура предложения преобразуется в последовательность путей с использованием разработанного алгоритма обхода графа. Второй подход основан на графовой сверточной сети. Для получения синтаксических признаков предложений сравнивались два различных синтаксических парсера. Для проверки моделей была выбрана задача профилирования автора по полу с использованием двух корпусов размеченных текстов, собранных на основе метода краудсорсинга и очного опроса. Было рассмотрено несколько вариантов наборов признаков исходных данных для кодирования слов текстов: морфологические бинарные векторы, векторные представления слов на основе модели FastText, а также их комбинации. В результате показано, что включение синтаксической структуры предложения в пространство входных признаков для задачи классификации пола автора текста позволяет улучшить известные точности в среднем на 4% для корпусов RusPersonality и 7% для корпуса Gender Imitation Crowdsource “a”. Итоговая точность по метрике f1 составляет 84% и 83% соответственно.

Ключевые слова: машинное обучение, искусственные нейронные сети, обработка естественного языка, автоматизированный анализ текстов, графовые нейронные сети, авторское профилирование, определение пола автора текста

DOI: 10.1134/S2304487X19060130

1. ВВЕДЕНИЕ

Область анализа текстов на естественном языке включает в себя широкий спектр задач, решение которых имеет высокую научную и практическую значимость. Результаты автоматической обработки текстов могут быть использованы в маркетинге, юриспруденции, медицине и управлении. Сложность таких задач заключается в том, что не существует оптимального представления текста в виде математической сущности без потерь информации, так как текст, как целое, не сводится к отдельным составляющим – предложениям, словам и символам.

В процессе развития данной области были созданы различные подходы интерпретации текстов для последующей машинной обработки: частотные (дистрибутивные модели word to vec [1]), модели последовательной обработки (рекуррентные модели RNN, LSTM [2], Convolutional LSTM [3]) и, наконец, последние наработки – модели, которые учитывают структуру текста (Tree LSTM [4], Attention Tree LSTM [5]).

Как показывают результаты экспериментов из упомянутых выше статей, использование информации о структуре предложения позволяет решать задачи определения эмотивности и семантической близости предложений с более высокой точностью, чем раньше, что приводит к выводу об информационной значимости структуры предложения в рамках данных задач.

Также стоит принять во внимание возможность выделения структуры текстов с высокой точностью при помощью автоматических средств “парсеров” [6–8], что позволяет использовать

любые выборки текстов без предварительно выделенной структуры, а значит расширяет область применения таких моделей.

Наконец, не для всех задач существуют выборки текстов значительной величины – зачастую требуется ручная разметка в рамках классификационной задачи с привлечением экспертов в определенной области. В данном случае необходимо создать такое признаковое пространство, которое бы характеризовало тексты, их отличительные особенности, достаточным образом для абстрагирования общей информации при помощи методов машинного обучения.

Таким образом, актуальной представляется цель данного исследования – разработка топологии искусственной нейронной сети для анализа текстов, представленных в виде деревьев зависимостей.

В качестве практической задачи, на которой будет исследована разработанная топология, выступает задача определения пола автора русскоязычного текста. Данная задача многообразна с точки зрения использования различных категорий признаков для описания текста и позволит подробно оценить вклад добавления различных категорий параметров с учетом структуры русскоязычного текста, выраженной через деревья зависимости отдельных предложений.

2. ОБЗОР ЛИТЕРАТУРЫ

В настоящее время учет структуры графа в задачах машинного обучения возможен на основе двух классов подходов: использование векторного представления графа (подграфа) в качестве входных данных модели машинного обучения или использование особых топологий нейронных сетей, позволяющих учитывать структуру графа в процессе обучения.

К первому классу методов относятся, например, node2vec [10], graph2vec [11]. Ко второму классу методов относятся TreeLSTM [4], TreeLSTM с механизмом внимания [5], графовые конволюционные сети [12].

Node2vec. Алгоритм node2vec позволяет создать векторное представление для узлов графа таким образом, чтобы в низкоразмерном пространстве признаков максимизировать вероятность сохранения информации о связях между соседними узлами графа. Алгоритм основан на процедуре смещенного случайного обхода, которая позволяет совместить преимущества от стратегий поиска в глубину и в ширину. В оригинальной статье также определяют такое понятие, как случайный обход второго порядка, построение пути в данном случае определяется двумя параметрами: p и q, соотношение которых определяет насколько быстро алгоритм покинет окружение корневого узла u.

Для создания векторного представления узла используется алгоритм skip-gram [1], при этом последовательность узлов в пути рассматривается аналогичным образом, как рассматривается последовательность слов в word2vec.

Принципы обхода дерева, которые лежат в основе node2vec, используются в разработанном подходе, описанном в разделе 3.2.

С точки зрения рассматриваемой в настоящей статье задачи (классификация текстов на основе информации о синтаксической структуре предложений), также можно заметить, что методы рассчитаны на графы с богатым числом связей и узлов, в пределах которых происходит “поиск” схожих закономерностей (структурной идентичности, кластеров узлов), построение на их основе пространства, которое позволяет получить векторное представление для узлов графа, а синтаксические графы представляют собой деревья, где у каждого узла число связей значительно меньше, чем в рассматриваемых в публикациях наборах данных.

TreeLSTM, TreeLSTM с механизмом внимания. TreeLSTM – это древовидная структура нейронной сети, которая является обобщением классической LSTM.

Описанная архитектура реализует возможность для каждой ячейки LSTM получать информацию от нескольких ячеек-”потомков”, таким образом обеспечивая рекурсивный обход дерева, при этом обновление внутреннего состояния блока зависит от состояния всех дочерних блоков. Это позволяет TreeLSTM блоку взвешенно объединять информацию от дочерних блоков и анализировать древовидную структуру, в то время, как обычный слой LSTM может быть использован только для анализа линейных последовательностей.

Недостатком предложенного подхода является необходимость наличия разметки на уровне фраз (т.е. метки для каждого узла графа), что сильно сужает спектр возможных задач до тех, где такая разметка присутствует в наборах данных – в описываемом исследовании используется Stanford Sentiment Treebank.

Чтобы создать возможность для решения задач с разметкой на уровне предложений, в статье [5] был предложен подход с использованием механизма внимания и словарей полярных слов. Показано, что полярные (эмотивные) словари вносят больший вклад в прирост точности решения задачи сентимент-анализа предложений на японском языке, чем использование механизма внимания. При этом авторы указывают, что данные словари используются как своеобразный аналог разметки на уровне фраз. Также в исследовании

приводится точность работы метода для Stanford Sentiment Treebank в случае, если учитывается только метка всего предложения – можно заметить, что отсутствие разметки на уровне фраз приводит к снижению точности решения задачи на 8%.

Графовые сверточные сети. Графовые сверточные сети (ГСС) явно используют структуру входных данных в виде графа в рамках нейросетевых топологий. При заданном графе $G = (V, E)$ (V – множество вершин, E – множество ребер), ГСС принимает на вход: матрицу признаков X размерности $N \times F$, где N это количество узлов графа, а F – размерность признаков каждого узла; матрицу смежности A графа размера $N \times N$, которая несет информацию о структуре графа.

Внутренние слои ГСС можно представить в виде $H_i = f(H_{i-1}, A)$, где f – нелинейная функция, H_{i-1} выход предыдущего слоя ($H_0 = X$), а H_i матрица $N \times F_i$ является представлением каждого узла графа в новом пространстве признаков. Таким образом, ГСС позволяет классифицировать узлы графа, либо же получать для них различные векторные представления, учитывающие графовые признаки.

Таким образом, в качестве основы для метода классификации русскоязычных текстов с использованием синтаксической структуры предложений могут быть использованы методы графовых сверточных сетей и совмещение концепций, представленных в методах на основе векторного представления графа и использования рекуррентных нейронных сетей (LSTM), поскольку, с одной стороны, позволяют в процессе обучения нейронной сети учесть как структурные особенности графа, так и векторное представление его узлов, с другой, частично компенсируют описанные у представленных методов недостатки, такие как, например, необходимость разметки каждого узла классифицируемого графа, специфичная форма представления зависимостей слов в предложении, отсутствие адаптации модели к конкретной задаче.

3. МЕТОДЫ И ПОДХОДЫ

3.1. Элементы нейросетевой топологии

Слой долгосрочной кратковременной памяти [2]. LSTM – это топология искусственных нейронных сетей, используемая для обработки последовательностей. Отличительной ее особенностью является наличие “памяти” и механизмов ее контроля, которые позволяют решить проблему “затухания градиента” при длинных последовательностях данных.

Глобальная операция объединения (global max-pooling) [20]. Операция, основанная на последовательном просмотре многомерной матрицы “окном”, смещаемым так, чтобы в конечном счете покрыть всю матрицу. При этом для каждого положения окна в матрице находится максимальное значение среди элементов в окне. Глобальный пулинг – вариант пулинга, при котором матричный вектор многомерных данных преобразуется в вектор максимальных значений по каждой матрице $m \times n$, т.е. матричный вектор $m \times n \times k$ преобразуется в $1 \times 1 \times k$, где k – элементы последовательности данных.

Полносвязный слой. Полносвязный слой является базовым элементом искусственных нейронных сетей. Каждый слой состоит из нейронов, которые принимают на вход произведение активностей прошлого слоя на матрицу весов, добавляя к результату “смещение” (bias). Далее применяется функция активации, а полученные результаты являются выходными активностями слоя. Соединение подобных слоев позволяет обеспечить нелинейное преобразование данных, что лежит в основе глубокого обучения.

В работе использовано 3 типа функций активации: усеченное линейное преобразование (ReLU), сигмоидальная функция, многомерная логистическая функция.

Dropout [21]. Техника регуляризации в нейронных сетях, которая позволяет препятствовать явлению “переобучения” за счет случайного “отключения” части связей нейронов в выбранном слое нейронной сети.

Графовый сверточный слой (ГСС) [12]. Графовые сверточные сети являются инструментом на основе нейронных сетей, разработанным специально для работы с графами и явно использующим их структурную информацию. При заданном графе $G = (V, E)$ (V – множество вершин, E – множество ребер), ГСС принимает на вход: матрицу признаков X размерности $N \times F$, где N это количество узлов графа, а F – размерность признаков каждого узла; матрицу смежности A графа размера $N \times N$, которая несет информацию о структуре графа. ГСС позволяет классифицировать узлы графа, либо же получать для них различные векторные представления.

3.2. Разработанный подход на основе анализа синтаксического окружения слов и нейросетевой топологии на базе LSTM слоев

Разработанная топология (см. рис. 1) содержит 1 LSTM-слой, 2 слоя GlobalMaxPooling с индексами 1 и 2, 1 полносвязный слой с функцией активации ReLU, 1 полносвязный слой с функцией активации Sigmoid, а также 1 полносвязный слой с функцией активации Softmax. Подход (далее Syntactic_LSTM) состоит из нескольких этапов:

1. Кодирование морфологических признаков слов текста. Каждое слово i текста j характеризуется вектором морфологических признаков, за-

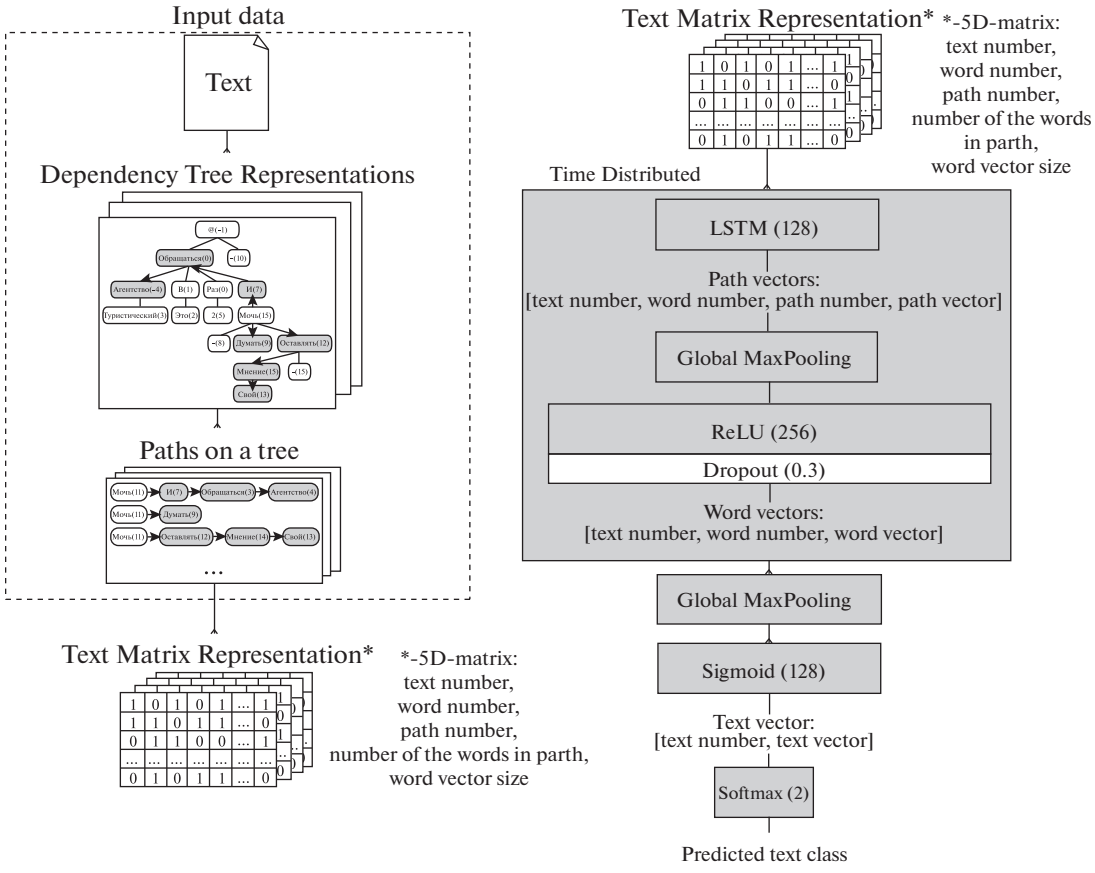

Рис. 1. Схема разработанной нейросетевой топологии на основе LSTM.

кодированных бинарно, образующих вектор размерности d_0.

2. Кодирование синтаксических признаков слов текста. Формируется набор путей $D^{(J, n)}$ для каждого слова i текста k, где J – количество путей, а n – количество слов в пути. Каждый путь $j \in J$ характеризуется матрицей $X^{(nd0)}$. Матрица X обрабатывается слоем LSTM размерности d_1, в результате чего образуется $\mathbf{v}_j$ вектор пути j. Все закодированные пути слова i текста k образуют матрицу $V^{(J, d1)}$. Далее матрица $V_{(i, k)}$ обрабатывается последовательно слоями GlobalMaxPooling, ReLU_1, и результатом обработки является вектор закодированных морфо-синтаксических признаков слова i текста k: $w_{(i, k)}$ размерности d_2 (количество нейронов в слое ReLU). Таким образом, весь k текст из m слов представляется в виде матрицы $W^{(m, d2)}$.

3. Кодирование текста. Для преобразования k текста, закодированного матрицей $W^{(m, d\_2)}$ используется последовательная обработка слоями $\text{GlobalMaxPooling}_2$ и Sigmoid_1. В результате образуется матрица $M^{(K, d3)}$, где K – количество текстов в анализируемых данных, d_3 – размерность слоя Sigmoid_1.

4. Классификация текста по полу автора. Итоговый пол автора текста определяется после обработки матрицы $M^{(K, d3)}$ полносвязным слоем с функцией активации Softmax.

3.3. Разработанный подход на основе графовых сверточных сетей

Архитектура нейронной сети состоит из двухслойной ГСС и двунаправленного LSTM-слоя (BiLSTM) (далее ГСС_BiLSTM). Данная архитектура позволяет классифицировать объекты, каждый экземпляр которых представляет собой набор графов. В данной статье такими объектами являются тексты, состоящие из графов предложений. Схематически данная архитектура представлена на рис. 2.

Сначала каждое предложение текста в виде графа подается на вход ГСС, состоящей из двух слоев со 128-ю нейронами. Граф предложения представляет собой набор узлов-слов, каждому узлу соответствует вектор признаков и ребер-связей между ними. Граф в виде матрицы смежности и матрицы с признаками узлов подается в ГСС, на выходе получаются новые векторы признаков для каждого узла. Затем производится усреднение

векторов всех узлов графа (предложения) для получения векторного представления предложения. Полученная последовательность векторов предложений подается на вход двунаправленному LSTM-слою (BiLSTM) с размерностью 128, полученное в результате данной процедуры представление текста поступает на вход в полносвязный слой размерности 2 (по числу целевых классов) и функцией активации SoftMax, на основе выходных значений которого определяется предсказанный для примера класс.

3.4. Корпуса данных

Разработанные подходы апробировались с использованием двух корпусов примеров, содержащих тексты с указанием пола его автора:

1. **RusPersonality**. Это представительный и валидированный лингвистами набор текстов с разметкой пола, возраста, стиля и других автороведческих параметров. Корпус содержит 1549 текстов-эссе по двум темам: “письмо другу” и “описание картины”. Из них 575 текстов, где авторы мужского пола, и 974 – женского. Тексты RusPersonality были предварительно сбалансированы по классам, итоговый размер выборки составил 1150 текстов.

2. **Gender imitation crowdsource “a” (GI cs “a”) –** корпус текстов, содержащий различную информацию о авторах. Собран средствами краудсорсинга с использованием заданий, составленных. Общее число текстов в корпусе GI cs – 5150. В данной работе мы использовали часть “a” из 1716 текстов с информацией о поле авторов. Тексты GI cs “a” были предварительно сбалансированы по классам, итоговый размер выборки составил 1664 текста.

Для получения синтаксических деревьев предложений было использовано два различных парсера: UDPipe [18] и парсер, описанный в [19]. Результаты приводятся для данных, обработанных каждым автоматическим синтаксическим разборщиком. В качестве признаков сравниваются: вектора морфологических признаков слов (далее “Морфо”), вектора, полученные с использованием FastText модели векторного представления слов (далее “FastText”), а также их сочетание (далее “Гибрид”). В качестве baseline оценки используются алгоритм [6], показывающий текущий State-of-the-art уровень решения задачи определения пола для данных корпусов.

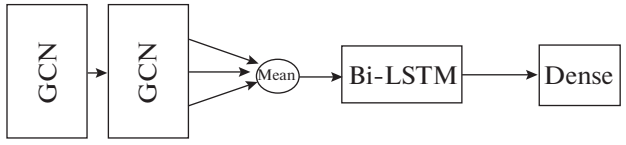

Рис. 2. Схема топологии на основе графовых сверточных сетей и BiLSTM.

4. ЭКСПЕРИМЕНТЫ

Результаты вычислительных экспериментов были получены с использованием стратифицированной кросс-валидации (Stratified Shuffle Split) с разделением исходной выборки на пять частей. На каждой итерации кросс-валидации одна часть принималась за тестировочную выборку, а оставшиеся четыре разделялись на тренировочную и валидационную. Таким образом, на каждой итерации кросс-валидации было получено следующее разбиение: 72% тренировочная выборка, 8% валидационная выборка, 20% тестировочная выборка. При этом тексты тренировочной и тестировочной выборок принадлежали разным авторам. Для оценки точности используется метрика F1-score. Полученные результаты на двух корпусах представлены в таблицах 1 и 2.

Параметры обучения для разработанной нейронной сети подобраны следующими: функция ошибки среднеквадратичная ошибка; функция оптимизации adam с коэффициентом обучения (learning rate), равным 0.001, beta 1 = 0.9, beta 2 = = 0.999. Количество текстов, обрабатываемых за 1 цикл работы сети 4, размеры слоев: d_0 – 48 для GI cs “a”, 52 для RusPersonality, d_1 – 128, d_2 – 256, d_3 – 128. В рамках обучения использовалась техника раннего останова с последующим восстановлением весов модели, установленных на эпохе с наименьшей ошибкой, рассчитанной на валидационной выборке. Ранний останов осуществлялся в случае, когда в течение 15-ти эпох не было снижения валидационной ошибки. Так-

Таблица 1. Результаты определения пола автора с использованием синтаксических признаков (синтаксический разборщик: [19])

	RusPersonality			GI cs “a”		
Модель/признаки	Морфо	FastText	Гибрид	Морфо	FastText	Гибрид
Syntactic_LSTM	82 ± 1	80 ± 2	82 ± 2	81 ± 2	83 ± 1	82 ± 2
ГСС_BiLSTM	84 ± 1	69 ± 2	82 ± 4	82 ± 2	60 ± 2	81 ± 2
[6]	79 ± 3	75 ± 4	76 ± 5	73 ± 8	73 ± 9	75 ± 4

Таблица 2. Результаты определения пола автора с использованием синтаксических признаков (синтаксический разборщик: UDPipe)

	RusPersonality			GI cs “a”		
Модель/признаки	Морфо	FastText	Гибрид	Морфо	FastText	Гибрид
Syntactic_LSTM	83 ± 2	82 ± 2	82 ± 2	82 ± 1	82 ± 1	83 ± 2
ГСС_BiLSTM	84 ± 1	68 ± 2	81 ± 2	81 ± 2	59 ± 2	81 ± 2
[6]	78 ± 2	81 ± 4	77 ± 3	71 ± 8	71 ± 7	77 ± 4

же был использован механизм циклического learning rate [22] с верхней границей, равной 0.01.

5. ЗАКЛЮЧЕНИЕ

На примере задачи авторского профилирования в части определения пола показано, что учет синтаксической структуры повышает точность классификации текстов.

Подход, основанный на рекуррентных и полносвязных слоях нейронной сети с построением набора путей на синтаксическом дереве для каждого слова, позволяет получить средний прирост f1-меры равный 3% в сравнении с результатом, опубликованным для корпуса RusPersonality и 7% для корпуса Gender Imitation Crowdsource “a”. Подход, основанный на графовой сверточной сети обеспечивает средний прирост f1-меры 4% для корпуса RusPersonality, 6% для корпуса Gender Imitation Crowdsource “a”.

БЛАГОДАРНОСТИ

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/.

ФИНАНСИРОВАНИЕ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-10084 “мк”.

СПИСОК ЛИТЕРАТУРЫ

1. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. MIT Press. 2013. V. 2. P. 3111–3119.
2. *Greff K., Srivastava R.K., Koutnık J., Steunebrink B.R., Bas R., Schmidhuber J.* LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems. IEEE. 2016. V. 28. № 10. P. 2222–2232.
3. *Hassan A., Mahmood A.* Deep learning approach for sentiment analysis of short texts. Proceedings of 2017 3rd international conference on control, automation and robotics (ICCAR). IEEE. 2017. P. 705–710.
4. *Tai K.S., Socher R., Manning C.D.* Improved semantic representations from tree-structured long short-term memory networks. In: arXiv preprint arXiv:1503.00075. 2015.
5. *Miyazaki R., Komachi M.* Japanese Sentiment Classification using a Tree-Structured Long Short-Term Memory with Attention. In: arXiv preprint arXiv:1704.00924. 2017.
6. *Sboev A., Moloshnikov I., Gudovskikh D., Rybka R.* A comparison of Data Driven models of solving the task of gender identification of author in Russian language texts for cases without and with the gender deception. Journal of Physics: Conference Series. IOP Publishing. 2017. V. 937. № 1. P. 012046.
7. *Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T.* Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. Procedia computer science. 2018. № 123. P. 417–423.
8. *Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T.* Deep Learning neural nets versus traditional machine learning in gender identification of authors of RusProfiling texts. Procedia computer science. 2018. № 123. P. 424–431.
9. *LeCun Y., Bengio Y.* Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. 1995. № 3361 (10).
10. *Grover A., Leskovec J.* node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016. ACM. 2016. P. 855–864.
11. *Narayanan A., Chandramohan M., Venkatesan R., Chen L., Liu Y., Jaiswal S.* graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:1707.05005. 2017.
12. *Kipf T.N., Welling M.* Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016.
13. *Veličković P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y.* Graph attention networks. arXiv preprint arXiv:1710.10903. 2017.
14. *Xinyi Z., Chen L.* Capsule graph neural network, 2018.
15. *Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J.* Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 2013. P. 3111–3119.
16. *Shervashidze, N., Schweitzer, P., Jan van Leeuwen E., Mehlhorn K., Borgwardt K.M.* Weisfeiler-lehman graph kernels. Journal of Machine Learning Research. 2011. P. 2539–2561.

17. *Goldberg Y., Levy O.* Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722. 2014.
18. *Straka M., Straková J.* Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics. Vancouver, Canada. 2017. P. 88–99.
19. *Rybka R., Sboev A., Moloshnikov I., Gudovskikh D.* "Morpho-syntactic parsing based on neural networks and corpus data. Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT). St. Petersburg. 2015. P. 89–95.
20. *Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M.* Striving for simplicity: The all convolutional net. 2014. arXiv preprint, arXiv:1412.6806.
21. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014. № 15(1). P. 1929–1958.
22. *Smith L.N.* Cyclical learning rates for training neural networks. IEEE Proceedings of the Winter Conference on Applications of Computer Vision (WACV). IEEE. 2017. P. 464–472.

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta "MIFI", 2019, vol. 8, no. 6, pp. 569–576

Neural Network Model for Classification of Text's Author Gender with Including Sentence Dependency Structure[1,2]

A. G. Sboev[a,b], A. A. Selivanov[a], I. A. Moloshnikov[a], R. B. Rybka[a], and D. S. Bogachev[a,c]

[a] *National Research Center "Kurchatov Institute", Moscow, 123098 Russia*

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

[c] *The Moscow Institute of Physics and Technology (MIPT), Moscow, 141701 Russia*

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

Received October 7, 2019; revised October 11, 2019; accepted October 15, 2019

Abstract—The research proposes the neural network methods to include a textual dependency tree structure in classification tasks of Russian texts. Author profiling task of gender identification was chosen to test the models, and two corpora used in experiments: based on a crowdsource, and in-person polling.
The first approach is based on a long short-term memory (LSTM) layers, and developed graph embedding algorithm. The second one is based on a graph convolution network and LSTM. Two syntactic parsers were used to obtain dependency trees from the texts. Input data was represented in different forms: morphological binary vectors, FastText vectors, and their combination. The developed models result was compared to the state-of-the-art, that is neural network model based on a convolutional and LSTM layers. Finally, we demonstrate that including textual dependency tree structure to input feature space improves f1-score of gender classification task on 4% for the RusPersonality dataset, and 7% for the crowdsource dataset in average. The developed models resulting f1-score is 84% and 83%, respectively.

Key words: machine learning, artificial neural networks, natural language processing, automated text analysis, graph neural networks, author profiling, author gender identification

DOI: 10.1134/S2304487X19060130

REFERENCES

1. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. MIT Press. 2013, vol. 2, pp. 3111–3119.
2. Greff K., Srivastava R.K., Koutnık J., Steunebrink B.R., Bas R., Schmidhuber J. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems. IEEE. 2016, vol. 28, no. 10, pp. 2222–2232.
3. Hassan A., Mahmood A. Deep learning approach for sentiment analysis of short texts. Proceedings of 2017 3rd international conference on control, automation and robotics (ICCAR). IEEE, 2017, pp. 705–710.
4. Tai K.S., Socher R., Manning C.D. Improved semantic representations from tree-structured long

[1] The reported study was funded by RFBR according to the research project *№ 18-29-10084 "мк"*

[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/.

short-term memory networks. In: arXiv preprint arXiv:1503.00075. 2015.

5. Miyazaki R., Komachi M. Japanese Sentiment Classification using a Tree-Structured Long Short-Term Memory with Attention. In: arXiv preprint arXiv:1704.00924. 2017.
6. Sboev A., Moloshnikov I., Gudovskikh D., Rybka R. A comparison of Data Driven models of solving the task of gender identification of author in Russian language texts for cases without and with the gender deception. Journal of Physics: Conference Series. IOP Publishing. 2017, vol. 937, no. 1, p. 012046.
7. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. Procedia computer science. 2018, no. 123, pp. 417–423.
8. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. Deep Learning neural nets versus traditional machine learning in gender identification of authors of RusProfiling texts. Procedia computer science. 2018, no. 123, pp. 424–431.
9. LeCun Y., Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks. 1995, no. 3361 (10).
10. Grover A., Leskovec J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016. ACM. 2016, pp. 855–864.
11. Narayanan A., Chandramohan M., Venkatesan R., Chen L., Liu Y., Jaiswal S. graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:1707.05005. 2017.
12. Kipf T.N., Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016.
13. Veličković P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y. Graph attention networks. arXiv preprint arXiv:1710.10903. 2017.
14. Xinyi, Z., Chen, L. Capsule graph neural network, 2018.
15. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 2013, pp. 3111–3119.
16. Shervashidze N., Schweitzer P., Jan van Leeuwen E., Mehlhorn K., Borgwardt K.M. Weisfeiler-lehman graph kernels. Journal of Machine Learning Research. 2011, pp. 2539–2561.
17. Goldberg Y., Levy O. Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722. 2014.
18. Straka M., Straková J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics. Vancouver, Canada. 2017, pp. 88–99.
19. Rybka R., Sboev A., Moloshnikov I., and Gudovskikh D., "Morpho-syntactic parsing based on neural networks and corpus data. Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT). St. Petersburg. 2015, pp. 89–95.
20. Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M. Striving for simplicity: The all convolutional net. 2014. arXiv preprint, arXiv:1412.6806.
21. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014, no. 15 (1), pp. 1929–1958.
22. Smith L.N. Cyclical learning rates for training neural networks. IEEE Proceedings of the Winter Conference on Applications of Computer Vision (WACV). IEEE. 2017, pp. 464–472.