ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

П.Н. Беген, Ю.Г. Мисников,
О.Г. Филатова

Использование автоматизированного инструментария для исследования интернет-дискурса: опыт применения рекуррентных нейронных сетей для изучения дискуссий по пенсионной реформе

Рекомендуемая форма библиографической ссылки

Беген П.Н., Мисников Ю.Г., Филатова О.Г. Использование автоматизированного инструментария для исследования интернет-дискурса: опыт применения рекуррентных нейронных сетей для изучения дискуссий по пенсионной реформе // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 119-130. — URL: http://keldysh.ru/abrau/2019/theses/91.pdf doi:10.20948/abrau-2019-91

Размещена также презентация к докладу

Использование автоматизированного инструментария для исследования интернет-дискурса: опыт применения рекуррентных нейронных сетей для изучения дискуссий по пенсионной реформе

П.Н. Беген[1], Ю.Г. Мисников[1], О.Г. Филатова[2]

[1] Университет ИТМО
[2] Санкт-Петербургский государственный университет

Аннотация. В статье представлены концептуальная модель прагматично-морального дискурса, послужившая основой для формирования массива обучающих данных, а также результаты эксперимента по применению рекуррентных нейронных сетей (RNNs) для оценки того, насколько точно они могут определять позицию участников интернет-дискурса по отношению к политике проведения пенсионной реформы в России на основе их сообщений и комментариев, размещенных на 16 дискуссионных площадках по всей стране. Проведенный эксперимент выявляет возможности и ограничения применения искусственных нейронных сетей для более глубокого понимания результатов общественных дискуссий.

Ключевые слова: рекуррентные нейронные сети, машинное обучение, сентимент-анализ, интернет-дискурс, дискурс-анализ, электронное участие, пенсионная реформа

Application of Automated Tools in Researching Internet Discourses: Experience of Using the Recurrent Neural Networks for Studying Discussions on Pension Reform

P.N. Begen[1], Y.G. Misnikov[1], O.G. Filatova[2]

[1] ITMO University
[2] St.Petersburg State University

Abstract. The paper presents a conceptual model of a pragmatic-moral discourse as a basis for assembling a training dataset, as well as the results of an experiment of using such data by the Recurrent Neural Network (RNN) to assess how accurately it can determine the attitude of Internet discourse participants towards the pension

reform in Russia based on their messages and comments posted on 16 discussion platforms across the country. The experiment shows possibilities and limitations of using artificial neural networks for a deeper understanding of the results of public discussions.

Keywords: recurrent neural networks, machine learning, sentiment analysis, Internet discourse, discourse analysis, e-participation, pension reform

1. Введение

Начавшийся в конце 90-ых и начале нулевых интерес к созданию и развитию инструментов электронной демократии и электронного участия, способствующих вовлечению граждан в государственную деятельность и принятие политических решений, по-прежнему остается важным социально-политическим трендом в обществе. В настоящее время в России и за рубежом имеется достаточно много платформ, способствующих проведению публичных дискуссий в интернете по различным общественно значимым проблемам.

В данной статье представлено исследование общественно-политического интернет-дискурса как одной из форм электронного участия. Анализ дискуссий в интернет-пространстве, касающихся важных для общества политических решений, позволяет исследовать инструменты электронного участия граждан в политике и определить их роль в обеспечении взаимодействия между обществом и властью [1].

Значительную роль в контексте электронного участия играет дискурсивный инструментарий, разработанный Ю. Хабермасом в рамках теории этики дискурса и коммуникативной этики, на базе его модели делиберативной демократии (deliberation (англ.) – обсуждения, дискуссия) [2]. Теория Хабермаса основывается на понятии базовых притязаний на нормативную значимость, правоту (basic validity claims), т. е. запросов, заявлений на некоторую тему, высказываемых участниками морально ориентированных дискурсов – нацеленных на поиск некоей «правды» – с ожиданием реакции на них со стороны других участников. Как правило, такую реакцию можно представить в агрегированном виде, выраженном через согласие или несогласие, что представляет собой актуализацию притязаний. Подобные дискурсы являются этически обоснованной формой политической организации общества, основанного на дискурсе и различиях свободных от принуждения граждан – делиберативной демократии. Наше исследование онлайн-дискуссий основывается на этой теории, а также опирается на методику дискурс-анализа, разработанную и описанную Ю. Г. Мисниковым в PhD-диссертации и ряде других работ [1, 3]. Инструментализированная таким образом концепция базовых притязаний Ю. Хабермаса является удобным методологическим средством изучения семантики дискурсивных событий в интернет-среде, в отличие от тонального анализа, который не «привязан» методологичски к какой-либо объясняющей социальной теории, выступая чисто техническим средством скорее лингвистического, чем семантического

анализа текста. Главная проблема такого рода инструментов (как анализ тональности текста) состоит в неспособности учесть делиберативность (дискуссионость) и диалогичность (интерактивность) интернет-дискурса, в то время как актуализация каждого притязания через согласие или несогласие является прямым отражением такой диалогичной интерактивности.

Развитие информационно-коммуникационных технологий, значительное увеличение количества данных и рост вычислительных мощностей обусловило возможности применения алгоритмов глубокого обучения (deep learning) для анализа текста в лексико-лингвистических и семантических исследованиях [4, 5]. Наиболее широкое распространение получило применение методов машинного обучения, в частности искусственных нейронных сетей, в сентимент-анализе (анализе тональности текста) и компьютерной лингвистике. Понимание этих тенденций позволило нам провести пилотное исследование с применением машинного обучения для сентимент-анализа дискуссий граждан на актуальную общественно-политическую тему.

В качестве конкретной проблематики для анализа интернет-дискурса была рассмотрена инициатива власти осуществить пенсионную реформу в России, где основным поводом для дискуссии стала публикация законопроекта о повышении пенсионного возраста. Целью данного исследования стала разработка автоматизированного инструментария для проведения анализа интернет-дискурса, в частности проведение эксперимента по применению искусственных нейронных сетей в определении позиций граждан (сентимент-анализ) в онлайн дискуссиях, касающихся проведения пенсионной реформы.

2. Концептуальная модель дискурса

С практической точки зрения анализ интернет-дискурса состоит в выявлении означенных выше притязаний на нормативную правоту, содержащихся в размещенных комментариях (сообщениях) и их кодировке (разметке) в целях определения как прагматичной составляющей дискурса, например, в форме рациональной аргументации, так и морально-этической, мировоззренческой позиции участников дискурса. Наличие или отсутствие такой позиции является свидетельством того, является ли дискурс морально-ориентированным либо остается на уровне рационально-прагматичного обмена мнениями. Модель перехода от оригинального авторского текста к утверждению нормативной правоты с помощью рациональной аргументации концептуально можно представить в виде пирамиды дискурса (см. рис. 1).

Рис. 1. Концептуальная модель прагматично-морального дискурса

На уровне 1 обученный в области контент-анализа кодировщик (человек, а не машина) анализирует оригинальный авторский текст и сокращает его до минимально возможного без потери смысла, убрав те части лексики, которые избыточны с точки зрения понимания основного смысла текста (предлоги, союзы, другое); однако текст по-прежнему должен оставаться авторским, т.е. должен выглядеть естественным и кратко написанным с использованием лексики автора (не допускается изменение текста).

Уровень 2 фиксирует с помощью использования приемов рациональной аргументации акт согласия или несогласия с ранее высказанной точной зрения, т. е. происходит актуализация притязания на значимость, если такое притязание имело место (что не обязательно).

Выделяется два типа актуализации факта согласия или несогласия: один тип – это целевая актуализация, когда (не)согласие направлено напрямую на конкретный пост либо с использованием таких встроенных функций, как "ответ", "цитата", либо в виде реакции (ответа) на непосредственно предшествующий пост; второй тип является логической актуализацией, когда (не)согласие не адресовано конкретному посту напрямую, но такой пост (как правило это один пост) может быть идентифицирован логически по смыслу или по другим признакам. Идентификация (не)согласий второго типа начинается с анализа содержания наиболее близких по времени размещения записей,

охватывая, как правило, 10 последних по отношению к данной записи постов. Необходимость выявления согласия-несогласия дает возможность не только указать причины (т.е. аргументы) такого согласия-несогласия, но и дать свое видение вопроса в случае несогласия, выдвинув новое притязание и соответственно обосновав его, перейдя тем самым на 3-ий уровень дискурса.

Уровень 4 обобщает притязание на нормативную правоту до уровня моральной позиции, если это уместно и возможно (как правило, это достижимо в диспутах на общественно значимые темы) в формате, например, «Согласен с принятием пенсионной реформы» или «Правильно, что пенсионная реформа будет принята».

Наполнение уровней соответствующим содержанием фактически означает разметку исходного текста для последующего использования в формировании материала для обучения нейросети с выделением ключевых слов и формированием первой пары, связывающей полный авторский текст с притязанием на правоту и моральной позицией. Авторы данной работы провели ряд исследований, которые показали надежность кодирования притязаний на значимость в ручном режиме, т. е. обученными кодировщиками. Например, был проведен анализ высказываний участников различных интернет-дискуссий на предмет отношения к политике уничтожения санкционных продуктов, поступающих из западных стран [1, 6, 7, 8]. Но такие исследования, базирующиеся на ручной кодировке, требуют много времени, внимательности со стороны кодировщиков и относительно небольшой выборки. При этом для каждого нового исследования необходимо проводить новое кодирование. В то же время использование должным образом обученных нейросетей позволяет использовать результаты кодировки многократно на сходном по содержанию материале. Опыт применение принципов машинного обучения для дискурс-анализа представлен ниже.

3. Разработка автоматизированного инструментария для исследования интернет-дискурса

Целью проведения эксперимента является выявление возможностей применения искусственных нейронных сетей для более глубокого понимания результатов общественных дискуссий.

Для проведения эксперимента, основанного на глубоком обучении нейронных сетей с тем, чтобы сеть научилась определять позицию участников дискуссий по отношению к пенсионной реформе («за», «против» или «нейтрально»), необходимо сформировать массивы данных на основе методики анализа дискурса.

Для анализа интернет-обсуждений, касающихся пенсионной реформы, были выбраны интернет-площадки одиннадцати разных по численности населения российских городов. Согласно своду правил Минэкономразвития РФ, города подразделяются на крупнейшие, крупные, большие, средние и малые [9]. Из каждой группы было отобрано по два города, выявлены их

наиболее популярные интернет-площадки и проанализированы онлайн-дискуссии по пенсионной тематике. Анализировались дискуссии в следующих городах: Санкт-Петербург и Волгоград (крупнейшие), Калининград и Севастополь (крупные), Братск и Нальчик (большие), Белореченск и Ханты-Мансийск (средние), Урюпинск и Снежинск (малые); отдельное внимание в исследовании уделялось Москве.

В качестве дополнительных источников данных для машинного обучения было выбрано три максимально разных форума, представляющих самые разные социальные группы: всероссийский женский портал Woman.ru, сайт отзывов otzovik.com и сайт электронного периодического издания KM.ru.

Всего было проанализировано 16 форумов, содержащих 10592 постов, которые разместили 998 человек. Из всех постов: «за» – 304 сообщения (3% от общего числа); «против» – 2510 сообщений (24%); «нейтральное отношение» – 7778 сообщений (73%).

Данные для машинного обучения (посты участников дискуссий) были представлены в Excel-таблицах, а затем экспортированы в формате .csv для удобной работы с ними в программной среде. Все собранные посты были размечены кодировщиками тремя цифрами следующим образом: 0 – категория «против», 1 – «за» и 2 – «нейтральное отношение».

Для реализации автоматизированного инструментария для исследования интернет-дискурса и написания программной части использовался язык программирования Python и сторонние библиотеки для работы с данными и методами машинного обучения.

Данные .csv были загружены в программную часть с помощью библиотеки pandas. Для работы были использованы 2 столбца с данными – «текст сообщения» и «за-1/против-0/нейтрально-2», загруженные как семантически связанные пары X и Y соответственно, где X является набором данных для обучения на входе (текст поста или комментария), а Y – целевым результатом на выходе (позиция участника).

Для работы с данными в реализации алгоритма машинного обучения была произведена предварительная обработка текста с помощью встроенного класса Tokenizer библиотеки Keras, который позволяет удалить ненужные символы, привести слова к нижнему регистру, вычислить частоту встречаемости слов и т. д. Таким образом, были удалены знаки пунктуации, невидимые символы, символы переноса, цифры. В соответствии с международной практикой обработки текста [10, 11] не были взяты в расчет наиболее частые и наименее редкие слова, а был взят набор из 3000 наиболее релевантных слов.

Далее данные были разделены на обучающую и тестовую выборки в соотношении 80/20, т. е. 20% всей совокупности данных использовались для финального тестирования обученной модели. Причем тестовые данные не попадали в обучение модели, соответственно модель «увидела» эти данные впервые при финальном тестировании.

Обработанный текст далее был представлен в виде векторной последовательности слов с размерностью 200. Все слова в обучающем массиве будут представлены в виде векторов с данной размерностью. Изначально такой вектор заполнен случайными числами (чаще всего нулями), а в процессе обучения значения будут изменяться таким образом, чтобы слова, используемые в одном контексте, были максимально близки в векторном пространстве.

В качестве алгоритма машинного обучения была использована рекуррентная нейронная сеть (RNN) с LSTM-блоками. Такой выбор связан с тем, что рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины, а LSTM-блоки (блоки с долгой/краткосрочной памятью) являются разновидностью рекуррентных сетей и очень хорошо справляются с задачами классификации и прогнозирования [12, 13, 14]. Такая концепция позволила нашей модели запоминать в ходе обучения предыдущие значения в последовательности векторов для дальнейшего принятия решения и настройки весов на скрытых слоях нейронной сети.

Изначальное количество итераций, за которое модель должна была обучиться, составило 100 циклов. Для дополнительного предотвращения переобучения сети была использована функция «ранней остановки» (EarlyStopping), проверяющей погрешность между функциями потерь и ошибок сети. Было замечено, что данная функция «ранней остановки» сработала после 10-го цикла обучения.

Для получения результатов мы использовали рекуррентную нейронную модель, состоящую из следующих слоев и блоков:

- входного слоя (т. е. последовательного набора данных) (Input);
- связующего слоя, который переводит весь словарь слов к настраиваемой размерности (200) для дальнейшего обучения (Embedding);
- LSTM-блока, запоминающего информацию с предыдущих последовательностей;
- полносвязного слоя с блоком линейной ректификации (функция активации ReLU), отвечающего за определение значений нейронов и их настройку;
- слоя, не допускающего переобучения сети на данных (посредством исключения нейронов, которые при любых значениях и параметрах возвращают 0) (Dropout);
- выходного слоя с настраиваемым количеством выходов (в нашем случае 2 или 3, с функциями активации sigmoid или softmax).

Подчеркнем, что в рамках реализации машинного обучения было использовано два подхода для построения выходного слоя:

1) Использование бинарной классификации (только категории «за» или «против», категория «нейтральное отношение» в данном случае

не использовалась), соответственно было использовано только 2814 высказываний (1/4 основного объема сообщений).

2) Использование классификации по трем категориям: «за», «против» и «нейтральное отношение». Данные были использованы в максимальном объеме (10592 высказывания).

В результате применения подхода с построением бинарной классификации на выходном слое (2 категории «за» и «против») моделью были получены показатели точности примерно 89% определения. Данный показатель является приемлемым для решения задачи и определения тональности высказывания, однако с увеличением массива данных и появления новых слов, не задействованных в составленном моделью словаре, этот коэффициент будет стремительно снижаться, т. к. значение функции потерь составило 4%.

Во втором случае, при использовании всех трех категорий, показатели точности определения категории составили 78%. Результат оказался хуже, чем при бинарной классификации в силу неравномерности распределения данных по категориям и увеличения массива новыми данными, однако данный показатель также является приемлемым для дальнейшего использования при работе с интернет-дискурсом.

Также было проведено открытое тестирование, т. е. были взяты случайным образом данные из тестовой выборки и с помощью обученной модели были получены цифры категорий, которые сравнили с истинным результатом. Такое тестирование показало, что высказывания, относящиеся к категории «Нейтральное отношение к пенсионной реформе», определяются с наиболее высокой вероятностью (высокой точностью), чем высказывания, относящиеся к категориям «против» и «за» в силу неравномерного распределения данных по категориям (3/4 от всего объема данных приходится на категорию «Нейтральное отношение»). Соответственно, для дальнейшей интерпретации результатов проведенного исследования, необходим новый эксперимент либо с большим количеством равномерно распределенных данных о высказываниях участников, либо с применением других методов машинного обучения и подходов к решению.

4. Заключение

В результате данной работы было рассмотрено концептуальное предложение модели прагматично-морального дискурса, на основе которого был собран обучающий массив данных, содержащий мнения участников касательно проведения пенсионной реформы, и проделан эксперимент по применению искусственных нейронных сетей в определении позиции участника в онлайн-дискуссиях. В дальнейшем также будет продолжена работа с концептуальной моделью дискурса, которая будет совершенствоваться и актуализироваться для решения новых задач.

Был разработан автоматизированный инструментарий для исследования интернет-дискурсов на основе рекуррентных нейронных сетей с LSTM-блоком.

Для бинарной классификации (категории «за» и «против») коэффициент точности составил 89%. Для классификации трех позиций (категории «за», «против», «нейтральное отношение») коэффициент точности составил 78%.

На основе анализа полученных результатов было определено проведение новых экспериментов по улучшению показателей точности классификации. В качестве подходов можно выделить: увеличение или корректировка массива обучающих данных и его балансировка по категориям; использование предобученных словарей терминов (дистрибутивных тезаурусов) [15, 16]; использование слоя «внимания» (Attention layer) [17, 18] для предотвращения эффекта «забвения» у рекуррентной нейронной сети; использование алгоритмов Word2Vec [19], Doc2Vec [120], GloVe [21] для работы с контекстом в тексте и определения семантически близких слов.

Работа выполнена при поддержке РНФ, проект №18-18-00360 «Электронное участие как фактор динамики политического процесса и процесса принятия государственных решений».

Литература

1. Мисников Ю. Г., Филатова О. Г., Чугунов А. В. Электронное взаимодействие власти и общества: направления и методы исследований // Научно-технические ведомости СПбГПУ. Гуманитарные и общественные науки. 2016. №1. С. 52–60.
2. Habermas J. Moral consciousness and communicative action. Cambridge: Polity Press, 1992.
3. Misnikov Y. Public Activism Online in Russia: Citizens' Participation in Webbased Interactive Political Debate in the Context of Civil Society. Development and Transition to Democracy. University of Leeds, 2011.
4. Wang P., Xu J., Xu B., Liu C., Wang H.Z.F., Hao H. Semantic Clustering and Convolutional Neural Network for Short Text Categorization // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, July 26-31, 2015. P. 352–357. URL: http://www.aclweb.org/anthology/P15-2058.
5. Socher R., Perelygin A., Wu J.Y., Chuang J., Manning C.D., Ng A.Y., Potts C. Recursive deep models for semantic compositionality over a sentiment treebank // EMNLP. 2013. P. 1631–1642.
6. Misnikov Y., Chugunov A., Filatova O. Converting the outcomes of citizens' discourses in cyberspace into policy inputs for more democratic and effective government // Alois A. Paulin, Leonidas G. Anthopoulos, and Christopher G. Reddick (eds). Beyond Bureaucracy: Towards Sustainable Governance Informatisation. Springer Science and Business Media, Public Administration and Information Technology Book Series. 2017. P. 259–291.

7. Misnikov Y., Chugunov A., Filatova O. Online Discourse as a Microdemocracy Tool: Towards New Discursive Epistemics for Policy Deliberation // ACM International Conference Proceeding Series. 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016; Montevideo; Uruguay. 2016. P. 40–49.
8. Misnikov Y., Chugunov A., Filatova O. Citizens' deliberation online as will-formation: The impact of media identity on policy discourse outcomes in Russia // Lecture Notes in Computer Science. Springer, 2016. Vol. 9821. P. 67–82.
9. СП 42.13330.2011 Градостроительство. Планировка и застройка городских и сельских поселений. Актуализированная редакция СНиП 2.07.01-89 (с поправкой). М., 2011
10. Ингерсолл Г. С. Обработка неструктурированных текстов. Поиск, организация и манипулирование. / Ингерсолл Г. С., Мортон Т. С., Фэррис Э. Л. // Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 414 с.
11. Барсегян А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.
12. Aggarwal C. C. Data Classification: Algorithms and Applications. Text Classification. Chapman & Hall/CRC, 2014. 705 p.
13. Zhou C., Sun C., Liu Z., Lau F.C.M. A C-LSTM Neural Network for Text Classification. 2015. URL: https://arxiv.org/abs/1511.08630
14. Prasanna P. D. R. Rao L. Text classification using artificial neural networks // International Journal of Engineering & Technology. 2018. Vol. 7 (1.1). P. 603–606.
15. RusVectōrēs. URL: https://rusvectores.org/ru/about/
16. Национальный корпус русского языка. URL: http://ruscorpora.ru
17. Colin R., Ellis D.P.W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. 2015. URL: https://arxiv.org/abs/1512.08756
18. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. Attention Is All You Need. 2017. URL: https://arxiv.org/abs/1706.03762
19. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. URL: https://arxiv.org/abs/1301.3781
20. Mikolov T., Le Q. Distributed Representations of Sentences and Documents. URL: https://cs.stanford.edu/~quocle/paragraph_vector.pdf
21. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation. URL: https://nlp.stanford.edu/pubs/glove.pdf

References

1. Misnikov Y. G., Filatova O. G., Chugunov A. V. Elektronnoe vzaimodejstvie vlasti i obshchestva: napravleniya i metody issledovanij // Nauchno-

tekhnicheskie vedomosti SPbGPU. Gumanitarnye i obshchestvennye nauki. 2016. №1. S. 52–60.
2. Habermas J. Moral consciousness and communicative action. Cambridge: Polity Press, 1992.
3. Misnikov Y. Public Activism Online in Russia: Citizens' Participation in Webbased Interactive Political Debate in the Context of Civil Society. Development and Transition to Democracy. University of Leeds, 2011.
4. Wang P., Xu J., Xu B., Liu C., Wang H.Z.F, Hao H. Semantic Clustering and Convolutional Neural Network for Short Text Categorization // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, July 26-31, 2015. P. 352–357. URL: http://www.aclweb.org/anthology/P15-2058
5. Socher R., Perelygin A., Wu J.Y., Chuang J., Manning C.D., Ng A.Y., and Potts C. Recursive deep models for semantic compositionality over a sentiment treebank // EMNLP. 2013. P. 1631–1642.
6. Misnikov Y., Chugunov A., Filatova O. Converting the outcomes of citizens' discourses in cyberspace into policy inputs for more democratic and effective government // Alois A. Paulin, Leonidas G. Anthopoulos, and Christopher G. Reddick (eds). Beyond Bureaucracy: Towards Sustainable Governance Informatisation. Springer Science and Business Media, Public Administration and Information Technology Book Series. 2017. P. 259–291.
7. Misnikov Y., Chugunov A., Filatova O. Online Discourse as a Microdemocracy Tool: Towards New Discursive Epistemics for Policy Deliberation // ACM International Conference Proceeding Series. 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016; Montevideo; Uruguay. 2016. P. 40–49.
8. Misnikov Y., Chugunov A., Filatova O. Citizens' deliberation online as will-formation: The impact of media identity on policy discourse outcomes in Russia // Lecture Notes in Computer Science. Springer, 2016. Vol. 9821. P. 67–82.
9. SP 42.13330.2011 Gradostroitel'stvo. Planirovka i zastrojka gorodskih i sel'skih poselenij. Aktualizirovannaya redakciya SNiP 2.07.01-89 (s popravkoj). M., 2011
10. Ingersoll G. S. Obrabotka nestrukturirovannyh tekstov. Poisk, organizaciya i manipulirovanie. / Ingersoll G. S., Morton T. S., Ferris E. L. // Per. s angl. Slinkin A. A. – M.: DMK Press, 2015. – 414 s.
11. Barsegyan A. A. Analiz dannyh i processov: ucheb. posobie / A. A. Barsegyan, M. S. Kupriyanov, I. I. Holod, M. D. Tess, S. I. Elizarov – 3-e izd., pererab. i dop. – SPb.: BHV-Peterburg, 2009. – 512 s.: il. + CD-ROM – (Uchebnaya literatura dlya vuzov).
12. Aggarwal C. C. Data Classification: Algorithms and Applications. Text Classification. Chapman & Hall/CRC, 2014. 705 p.

13. Zhou C., Sun C., Liu Z., Francis C.M. Lau. A C-LSTM Neural Network for Text Classification. 2015. URL: https://arxiv.org/abs/1511.08630
14. Prasanna P. D. R. Rao L. Text classification using artificial neural networks // International Journal of Engineering & Technology. 2018. 7 (1.1). P. 603–606.
15. RusVectōrēs. URL: https://rusvectores.org/ru/about/
16. Nacional'nyj korpus russkogo yazyka. URL: http://ruscorpora.ru
17. Colin R., Ellis D.P.W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. 2015. URL: https://arxiv.org/abs/1512.08756
18. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. Attention Is All You Need. 2017. URL: https://arxiv.org/abs/1706.03762
19. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. URL: https://arxiv.org/abs/1301.3781
20. Mikolov T., Le Q. Distributed Representations of Sentences and Documents. URL: https://cs.stanford.edu/~quocle/paragraph_vector.pdf
21. Pennington J., Socher R., Manning C. D. GloVe: Global Vectors for Word Representation. URL: https://nlp.stanford.edu/pubs/glove.pdf