ИНФОРМАТИКА

УДК 004.852, 004.891, 51-77
MSC 68T50, 91B99

Модификация языковой модели SBERT для выявления ESG-рисков на основе текстовых данных компаний и контрольно-надзорных мероприятий

А. В. Бузмаков[1], *Д. А. Кирпищиков*[1], *Ю. Н. Найденова*[1], *С. Н. Паклина*[1], *П. А. Паршаков*[1], *Р. И. Соломатин*[1], *Н. С. Сотириади*[2]

[1] Национальный исследовательский университет «Высшая школа экономики», Российская Федерация, 614000, Пермь, бул. Гагарина, 37
[2] ПАО «Сбербанк», Российская Федерация, 117312, Москва, ул. Вавилова, 19

Для цитирования: *Бузмаков А. В., Кирпищиков Д. А., Найденова Ю. Н., Паклина С. Н., Паршаков П. А., Соломатин Р. И., Сотириади Н. С.* Модификация языковой модели SBERT для выявления ESG-рисков на основе текстовых данных компаний и контрольно-надзорных мероприятий // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2025. Т. 21. Вып. 1. С. 75–91.
https://doi.org/10.21638/spbu10.2025.106

Разработан подход для выявления рисков, связанных с влиянием компаний на окружающую среду, социальной ответственностью и качеством управления (Environmental, Social and Governance — ESG-рисков), на основе собранной текстовой информации о компании. Для достижения этого предлагается модификация языковой модели SBERT с четко заданной функцией расстояния пространства эмбеддингов. Модель обучена на данных контрольно-надзорных мероприятий и текстов сайтов компаний. Приведен пример интерпретации результатов модели.

Ключевые слова: ESG, модель обработки естественного языка, обучение модели, тематическое моделирование, веб-сайт.

1. Введение. Выявление рисков, связанных с влиянием компаний на окружающую среду, социальной ответственностью и качеством управления (Environmental, Social and Governance — ESG-рисков), — актуальный вопрос с точки зрения прогнозирования устойчивости бизнес-модели компании и определения ее финансового здоровья и доходности [1]. ESG — это трехмерный подход к оценке устойчивости и социальной ответственности компаний. Он основывается на оценке воздействия компании на окружающую среду, общество и способах корпоративного управления, которые могут влиять на бизнес-модель и финансовые результаты компании. Текстовая информация о компании и данные контрольно-надзорных мероприятий (КНМ) могут быть использованы в качестве признаков, указывающих на наличие ESG-риск-факторов

для конкретной компании, и применяться в процессе идентификации и оценки соответствующих рисков.

Учет ESG-рисков помогает компаниям и их инвесторам избежать потенциальных плохих последствий, связанных с негативной реакцией стейкхолдеров и общества, экологическими и климатическими инцидентами, а также может способствовать достижению долгосрочной устойчивости. Однако на данный момент оценка ESG-рисков непубличных компаний проводится экспертно, что в большинстве случаев трудозатратно и субъективно.

Бóльшая часть информации о ESG-факторах, в том числе ESG-рисках, компании представлена в текстовом неструктурированном виде. Для оперативного анализа такой информации целесообразно использовать модели анализа текста, основанные на нейросетях, таких как модель BERT [2]. Для достижения интерпретируемости модель BERT была адаптирована, а именно в ней изменены финальные слои нейросети с целью явного контроля за поведением функции расстояния в пространстве эмбеддингов отдельных предложений текста. Разработанная модификация модели позволяет получать более интерпретируемые результаты анализа текстов, при этом незначительно теряя в точности. Она может применяться для анализа текстов о компании с целью определения подверженности компании ESG-рискам и дает возможность понять, какие элементы текста ассоциируются с повышенным ESG-риском.

Целью статьи является модификация модели BERT для выявления ESG-рисков на основе собранной текстовой информации о компании и данных КНМ и повышения интерпретируемости полученных результатов. Для достижения цели предлагается модификация языковой модели BERT с четко заданной функцией расстояния (рис. 1). Используются данные российских компаний металлургической отрасли (информация с веб-сайтов компаний, отчетные материалы, информация регулятора о проведенных КНМ).

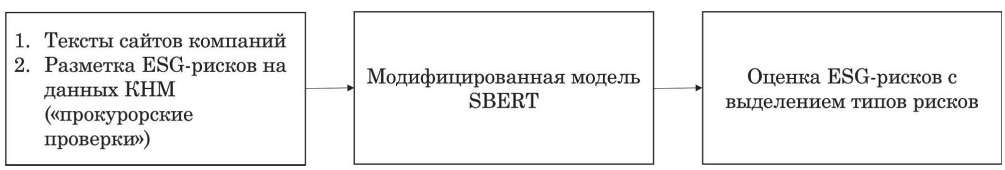

Рис. 1. Постановка и используемые данные

2. Использование подходов обработки естественного языка в области ESG. Последние исследования в области анализа ESG-рисков подчеркивают их растущую значимость в финансовом управлении и инвестиционных стратегиях. Наблюдается экспоненциальный рост публикаций, связанных с ESG, за последние пять лет, с акцентом на устойчивые финансы, управление рисками и формирование инвестиционных портфелей [3]. ESG-факторы все чаще интегрируются в рамки оценки рисков, что оказывает существенное влияние на финансовую результативность и устойчивость бизнеса [4]. Тем не менее остается проблема отсутствия четкой таксономии ESG-рисков, что затрудняет единообразную оценку [5]. Библиометрические анализы выявляют ключевые темы в исследованиях ESG, такие как соотношение риска и доходности, а также определяют влияние ESG-рейтингов на инвестиционные решения [6, 7]. Интеграция ESG-рисков в бизнес-модели варьируется в зависимости от размера компаний и регионов: крупные компании и развитые страны с большей вероятностью внедряют ESG-подходы в управление рисками [8]. В целом анализ ESG-рисков

становится важным инструментом для максимизации скорректированной на риск доходности и создания долгосрочной ценности для заинтересованных сторон [9, 10].

Важную роль в анализе ESG-факторов начинают играть методы обработки естественного языка (NLP). Установлено, что NLP может автоматизировать процесс оценки ESG, а это значительно ускоряет процесс обработки данных и повышает его точность [11]. Например, с помощью NLP создаются ESG-индексы [12], анализируются отчеты об устойчивом развитии [13], а также улучшается оценка документов с использованием таких моделей как BERT [11, 12]. Разработаны методы для выявления специфичных для отраслей ESG-проблем [14] и оценки продолжительности воздействия ESG-событий [15]. NLP также применяется для анализа финансирования ESG [16], оценки коммуникаций, связанных с ESG [17], и анализа литературы по технологии распределенных реестров в контексте ESG [18]. Эти исследования подчеркивают потенциал NLP для повышения эффективности анализа ESG, улучшения процесса принятия решений и сокращения пробелов в измерении и оценке ESG-факторов.

Особое внимание уделяется моделям, основанным на BERT, для анализа ESG-рисков и улучшения ESG-оценок. Выявлено, что настройка BERT на текстовые данные, связанные с ESG, может значительно повысить точность классификации ESG-задач [11, 19, 20]. Такие модели показали высокую эффективность в определении ESG-проблем на разных языках [21] и в извлечении текстовых доказательств для ESG-оценок [22]. Дополнение BERT лингвистическими и семантическими признаками также способствовало улучшению классификации данных ESG [23], а прогнозирование продолжительности воздействия ESG-новостей стало возможным благодаря развитию методов NLP [24]. Эти достижения открывают новые возможности для автоматизации ESG-оценок, создания ESG-портфелей и отслеживания новостного контекста в динамике [11, 25]. В конечном итоге такие подходы способствуют более стандартизированной и эффективной оценке ESG-рисков в финансовой отрасли.

Цель настоящей работы заключается в разработке модели оценки ESG-рисков для русского языка, обладающей двумя ключевыми особенностями. Во-первых, мы отказываемся от привлечения экспертов для разметки данных, что позволяет избежать субъективности в оценке. Во-вторых, предложенная модель включает детализированную классификацию типов ESG-рисков, что способствует более точному и углубленному анализу.

3. Анализ данных. После консультаций со специалистами в предметной области и самостоятельного поиска источников, включающих контролирующие и надзорные органы (в том числе пожарный надзор, потребнадзор, технадзор, трудовой надзор и т. п.), было выяснено, что бóльшая часть ведомств публикует открытые данные в агрегированном виде. Поскольку данные о ESG-рисках необходимы на уровне компаний, было решено сконцентрироваться на данных КНМ, которые включают в себя набор причин проверок и предписаний. Этот набор данных используется в качестве разметки ESG-рисков.

Исходная выборка состоит из компаний отраслей «производство металлургическое» и «производство готовых металлических изделий, кроме машин и оборудования» по общероссийской классификации видов экономической деятельности, данные по которым доступны в системе Спарк-Интерфакс. Общее количество компаний составило 6841, однако в процессе работы выборка была уменьшена из-за недоступности сайтов некоторых компаний, а также отсутствия данных о КНМ для них. Индивидуальные номера налогоплательщиков (ИНН) компаний используется как основной идентификатор во всех приложенных к отчету данных.

3.1. Разметка ESG-рисков по данным КНМ для обучения модели. Для выделения и классификации рисков, которым подвержена компания, на основе данных о КНМ были собраны два набора данных за 2021 и 2022 гг., предоставляемых в открытом доступе генеральной прокуратурой Российской Федерации: Единый реестр проверок (ЕРП) и Единый реестр контрольных (надзорных) мероприятий (ЕРКНМ).

ЕРП содержит результаты проверок компаний, опираясь на федеральные законы (ФЗ) № 294, 131 и 184. ФЗ 294 регулирует отношения в области организации и осуществления государственного контроля (надзора). ФЗ 131 устанавливает общие правовые, территориальные, организационные и экономические принципы организации местного самоуправления в Российской Федерации, определяет государственные гарантии его осуществления. ФЗ 184 регулирует отношения, возникающие при разработке, принятии, применении и исполнении обязательных требований к продукции, оценке ее соответствия. ЕРКНМ содержит результаты проверок компаний, которые были проведены согласно ФЗ 248, регулирующим отношения по организации и осуществлению государственного контроля (надзора), муниципального контроля и устанавливающим гарантии защиты прав граждан и организаций как контролируемых лиц.

Из ЕРП для проведения анализа были собраны данные по вынесенным предостережениям и результатам проведенных проверок для выборки российских металлургических компаний. В исходную выборку вошла 6841 компания. После исключения тех компаний, для которых был указан статус «Реорганизуется» и «Ликвидируется», а также тех, по которым недоступны ключевые финансовые показатели (выручка, стоимость активов или стоимость чистых активов) за период 2019–2021 гг., а также компаний с отрицательной стоимостью чистых активов и выручкой меньше 1 млн рублей, итоговая выборка составила 4415 компаний. Из этой выборки для 1100 компаний были найдены записи по 3519 проверкам. Таким образом, в среднем проверки затронули 25 % металлургических компаний, в среднем на каждую компанию пришлось по 3 проверки. В свою очередь, из ЕРКНМ были собраны данные по выявленным нарушениям и результатам проведенных КНМ. Для металлургической отрасли были найдены записи по 1527 проверкам для 562 компаний из итоговой выборки. Таким образом, КНМ затронули около 13 % компаний, в среднем на каждую компанию пришлось также по 3 проверки. Далее было произведено тематическое моделирование по текстовым данным, представленным в этих двух реестрах. Сначала анализ был проведен для реестра ЕРП, а затем для ЕРКНМ. Текст результата каждой проверки был разбит по словам (токенам). Затем из текста были удалены стоп-слова. Далее была проведена процедура стемминга — нахождение морфологической основы слова, были отфильтрованы слова, встречающиеся реже, чем в 5 текстах проверок, т. е. были исключены специфические и редкие слова. После экспертного анализа часто встречающихся основ слов был создан список «шумовых слов», который включает слова, не несущие смысловой нагрузки в рамках задачи выделения рисков. Далее полученный набор данных был преобразован в Document-Term Matrix, и на основе этой матрицы были удалены слова, входящие менее чем в 1 % текстов проверок.

Следующий шаг заключается в выборе количества тем для моделирования. По результатам тестов (рис. 2), выполненных на языке программирования R при помощи пакета «ldatuning», было решено рассмотреть категоризацию полученных слов по 23 темам при помощи латентного размещения Дирихле (Latent Dirichlet Allocation, или LDA) при анализе обоих наборов данных. В общем модель LDA позволяет объяснять результаты наблюдений с помощью неявных групп, благодаря чему возможно

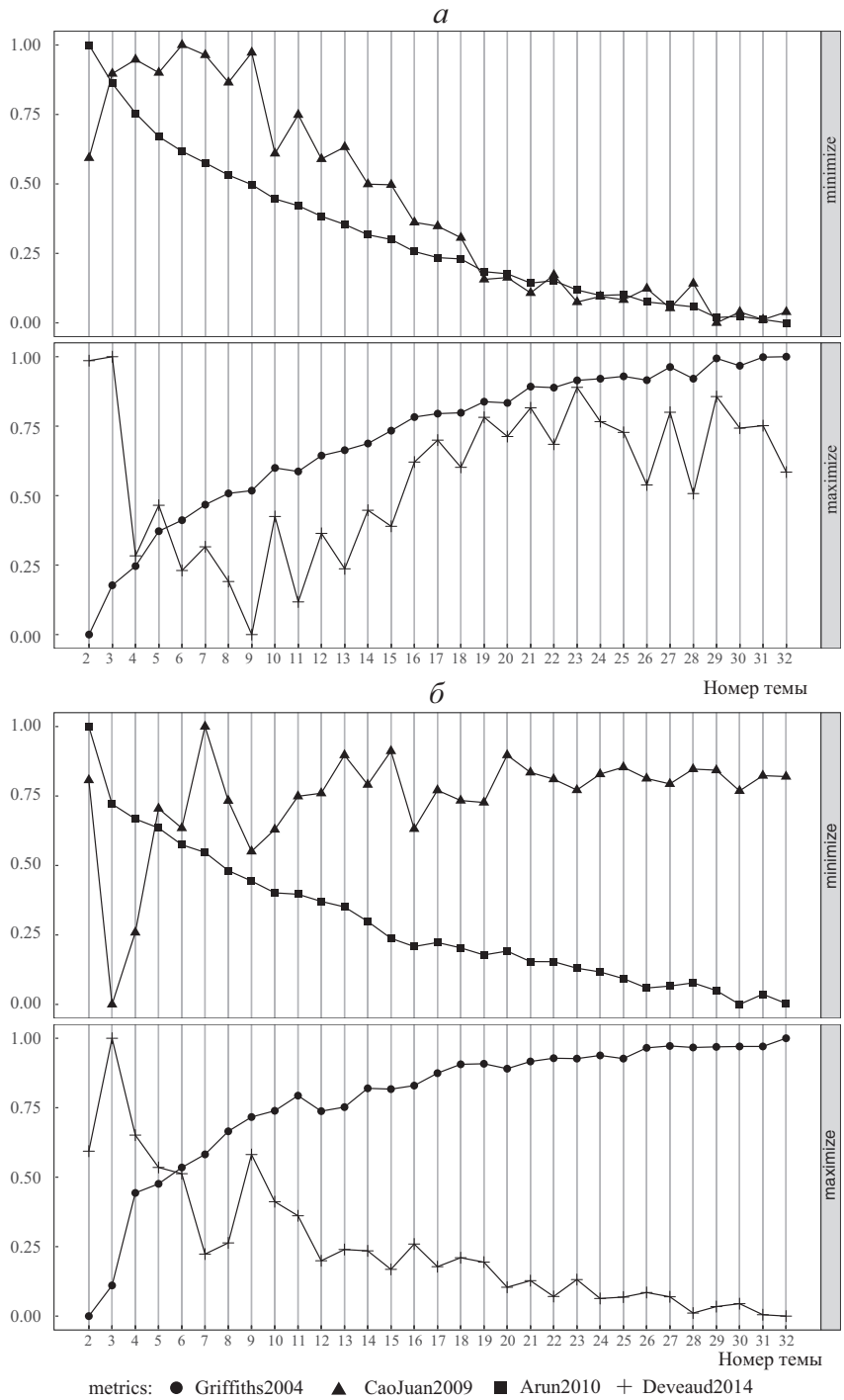

Рис. 2. Результаты определения оптимального количества тем при анализе текстов проверок для реестров ЕРП (*а*) и ЕРКНМ (*б*) при помощи пакета «ldatuning»

выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, то утверждается, что каждый документ представляет собой смесь некоторого количества тем и появление каждого слова связано с одной из тем документа. В табл. 1 представлены результаты проведенного анализа и выделенные группы риска на основе полученных тем при использовании ЕРП, а в табл. 2 — при использовании ЕРКНМ. Эти темы были объединены на основе экспертной оценки того, насколько они связаны с ESG-рисками. В табл. 3 приведен список данных тем с тремя самыми часто встречающимися в них основами слов по результатам анализа текстов ЕРП, а в табл. 4 — текстов ЕРКНМ.

Таблица 1. **Список выявленных ESG-рисков на основе анализа содержания результатов проверок ЕРП**

Категория	Номер темы	Слово 1	Слово 2	Слово 3
Соблюдение Трудового кодекса	1	трудов	сведен	деятельн
	18	социальн	страхован	обязательн
	22	работник	трудов	тк
Соблюдение санитарно-эпидемиологических правил и норм	2	санитарно-эпидемио-логическ	требован	санитаро-защитн
Общая экология	3	экологическ	окружа	сред
Работа с рисками объектов	4	риск	контроль	соответст
Контроль загрязнения атмосферы	5	выброс	воздух	веществ
Безопасность во время исполнения трудовых обязанностей	9	несчастн	производств	случа
Подкарантинная продукция	14	продукц	подкарантин	карантин
Общая тема, результаты проверок	11	требован	обязательн	контроль

Таблица 2. **Список выявленных ESG-рисков на основе анализа содержания результатов мероприятий ЕРКНМ**

Категория	Номер темы	Слово 1	Слово 2	Слово 3
Пожарная безопасность	1	оборудова	автоматическ	пожарн
	4	пожарн	безопасн	этаж
Безопасность на производстве	20	требован	нарушен	безопасн
	6	нарушен	выявл	подписа
Соответствие проектной документации	7	проектн	документац	выполн
	9	работ	документац	проектн
	22	предписан	выполн	устранен
Электробезопасность	11	отсутств	кабельн	соответст
Гражданская оборона	13	сооружен	защитн	гражданск
Эксплуатация техники	16	кран	отсутств	ход
Экология	12	направл	почт	завод

Для каждой темы была найдена его средняя представленность во всех документах. По результатам анализа текстов из ЕРП (табл. 3) темы «Соблюдение Трудового кодекса», «Работа с рисками объектов» и «Общая тема» имеют в среднем наибольшую представленность в проанализированных текстах результатов проверок.

Таблица 3. **Средняя представленность тем в проанализированных документах по результатам анализа текстов ЕРП**

Категория	Номер темы	Среднее значение
Соблюдение Трудового кодекса	1	0.032
	18	0.038
	22	0.042
Соблюдение санитарно-эпидемиологических правил и норм	2	0.031
Общая экология	3	0.025
Работа с рисками объектов	4	0.061
Контроль загрязнения атмосферы	5	0.035
Безопасность во время исполнения трудовых обязанностей	9	0.033
Подкарантинная продукция	14	0.030
Общая тема, результаты проверок	11	0.089

Таблица 4. **Средняя представленность тем в проанализированных документах по результатам анализа текстов ЕРКНМ**

Категория	Номер темы	Среднее значение
Пожарная безопасность	1	0.015
	4	0.014
Безопасность на производстве	20	0.008
Общая тема	6	0.017
Соответствие проектной документации	7	0.013
	9	0.030
	22	0.023
Электробезопасность	11	0.013
Гражданская оборона	13	0.013
Эксплуатация техники	16	0.008
Экология	12	0.019

По результатам анализа текстов ЕРКНМ (табл. 4) темы «Соответствие проектной документации» и «Экология» наиболее распространенные. Затем тексты проверок, которые связаны с ESG-рисками, такие как «Экология», «Безопасность на производстве» и «Соблюдение Трудового кодекса», использовались для обучения модели.

3.2. Текстовые данные сайтов компаний. Предположим, что каждый вид деятельности компании должен быть связан со специфичными рисками. Текстовое описание на сайте содержит наиболее детальную информацию о виде деятельности компании. Поэтому ассоциирование рисков (выявленных, например, в ходе прокурорских проверок) с текстом может выделить риски компании, в том числе те, которые пока еще не проявились.

Для изначальной выборки из 6841 компании из базы данных СПАРК были собраны сайты в сети Интернет. Для каждой компании дан как минимум один сайт. Из указанных сайтов удалось получить доступ к 2963, из которых активными являются около 70 %, что было оценено по времени сбора информации, превышающему 5 страниц. Под активным подразумевается такой сайт, при переходе на который запрашиваемый домен находится в использовании, а под неактивным — такой, домен которого свободен. Итог сбора данных составил 385 984 страниц. Среднее количество страниц на один веб-сайт равно 130, медиана — 37. Наиболее часто (309 раз) встре-

чаются сайты, содержащие только одну страницу. С 50 % сайтов было собрано более 37 страниц, а с 25 % — более 153.

На всех страницах содержится 103 373 013 предложений (с учетом незначащих предложений, в частности, элементов меню). На одной странице в среднем находятся 272 предложения (медиана составила 168, а мода — 3). Мода, принимающая значение 3, обычно сигнализирует, что страница содержит текст о том, что она не существует (например, ошибка 404 или 502). При этом среди собранного текста очень много элементов, состоящих из одного-двух слов. Важно отметить, что данные веб-сайтов достаточно насыщены текстовой информацией: 50 % страниц содержат более 543 слов, 25 % — более 975.

Итоговая выборка состоит из 2963 сайтов, с которых было собрано 385 984 страниц, на которых содержатся 103 373 013 предложений. Исходя из этого, имеется достаточная по объему выборка предложений для обучения модели. Таким образом, собранные данные подтверждают возможность сбора и анализа текстовой информации с сайтов российских металлургических компаний. Сайты значительно различаются по объему представленной информации, количеству страниц и их наполнению.

4. Описание модели и принципов ее обучения. В основе модели лежит языковая модель Bidirectional Encoder Representations from Transformers (BERT), которая позволяет создать эмбеддинги как векторное представление для элементов текстовых данных. BERT [2] — это нейросетевая языковая модель, которая относится к классу трансформеров. Она состоит из 12 «базовых блоков» (слоев) и 768 параметров на каждом слое. На вход модели подается предложение или пара предложений, которые затем разделяются на отдельные слова (токены). После чего в начало последовательности токенов вставляется специальный токен [CLS], обозначающий начало предложения или начало последовательности предложений. Пары предложений группируются в одну последовательность и разделяются с помощью специального токена [SEP], затем к каждому токену добавляется эмбеддинг, показывающий, к какому предложению относится токен. Далее все токены трансформируются в эмбеддинги по механизму, описанному в работе [26], в которой рассматривается возможность модели BERT создавать эмбеддинги для предложений. На вход модели BERT подаются два предложения в текстовом виде. По ним модель BERT строит так называемые эмбеддинги, т. е. численные вектора размерности 768.

Анализ текстовых данных начинается с того, что исходное предложение разбивается на токены, т. е. слова и части слов. Если в тексте встречается неизвестное слово, то оно разбивается на составляющие части, известные модели. Для анализа эмбеддингов предложений важен токен начала предложения (токен CLS), эмбеддинг которого будем считать эмбеддингом всего предложения. Далее модель сопоставляет каждому известному токену некоторый фиксированный эмбеддинг, т. е. уникальный номер токена. Он не имеет семантики и служит исключительно как кодировка для передачи текстовой информации в модель. Далее эти фиксированные эмбеддинги пропускаются через модель BERT, состоящую из 12 слоев-трансформеров. Каждый слой на основании эмбеддинга текущего слова и эмбеддингов других слов пересчитывает эмбеддинг текущего слоя. Таким образом, выходной эмбеддинг модели для каждого слова зависит не только от текущего слова, но и от контекста, в котором оно было употреблено. Также определяется эмбеддинг для токена предложения CLS, который кодирует все предложение целиком. Эмбеддинг этого токена размерности 768 пропускается через один полносвязный слой, получая новый вектор размерности 768, который и является финальным эмбеддингом всего предложения.

Так как рассматриваемая задача сводится к поиску похожих текстов, то было решено выбрать архитектуру Sentence-BERT, которая позволяет приблизить похожие эмбеддинги текста и отдалить непохожие [27].

По эмбеддингам предложений рассчитывается расстояние между ними, которое преобразуется в вероятность встретить пару входных предложений в одном контексте. Расстояние между предложениями определяется с помощью расстояния Эвклида (d), а вероятность рассчитывается по формуле

$$P = \frac{2}{1 + \exp d}.$$

Для контроля за обучением модели и исключения проблемы переобучения выборка данных делится на обучающую и тестовую подвыборки. В данной работе 80 % всех текстовых данных были отнесены к обучающей подвыборке, а остальные 20 % сформировали тестовую подвыборку. Для каждой подвыборки были сформированы наблюдения для обучения и тестирования модели. Каждое наблюдение представляет собой пару предложений. Для задачи анализа расстояния между предложениями рассматриваются пары двух типов: пара предложений из одного контекста и пара предложений из разных контекстов.

Для формирования пар предложений из одного контекста случайным образом выбирается «первое» предложение из всего множества обрабатываемой подвыборки данных. Вторым предложением берется случайное предложение, находящееся рядом с «первым» предложением. Здесь под «рядом» понимается предложение с того же сайта и расположенное в непосредственной близости от выбранного ранее «первого» предложения. Для формирования наблюдений из пар предложений из разных контекстов «первое» предложение выбиралось также случайно. А второе бралось либо с другого сайта, либо с того же сайта, но на расстоянии не менее 10 предложений вперед или назад относительно «первого» предложения.

Обучение нейросетевых моделей произведено группами наблюдений («батчами»). Для повышения скорости и качества обучения батчи сформированы так, чтобы внутри одного батча содержалось равное количество наблюдений «первого» и «второго» типов, т. е. примеров одного предложения из разных контекстов. Размер одного батча был задан в 16 наблюдениях ввиду ограниченности доступной видеопамяти, используемой для обучения модели.

Для ограничения времени обучения количество наблюдений внутри одной эпохи задано константой в 20 тыс. пар предложений для обучающей подвыборки и в 10 тыс. пар предложений для тестовой. Такой размер эпохи позволяет увидеть изменение качества модели, которое считается в конце каждой эпохи как на тестовой, так и на обучающей подвыборке, а также дает возможность сократить время обучения до примерно 5–6 мин. Меньшее количество наблюдений в обучающей подвыборке по сравнению с тестовой приводит к увеличению скорости расчетов. В процессе обучения оптимизация достигается за счет минимизации перекрестной энтропии.

Обучение модели проводилось методом подстройки, начиная с модели DeepPavlov/rubert-base-cased-sentence [28]. Для обучения модели обновлялись веса последних трех слоев исходной модели BERT. При этом на первых 5 эпохах обновлялись веса только последнего слоя, на следующих 5 эпохах — только веса последних двух слоев и, наконец, на последующих эпохах веса последних трех слоев. Важно отметить, что используемая функция потерь (loss function) существенно отличается от функции исходной модели (см. приведенную выше формулу).

Рассмотрим динамику изменения качества модели при увеличении количества эпох. Применялись две метрики качества модели: точность и правдоподобие. Точность определяется как отношение количества правильных предсказаний к количеству всех попыток. Правдоподобие — это усредненная величина логарифма значения правдоподобия для одного примера. С целью повышения интерпретируемости среднее значение приводится обратно к обычному правдоподобию и презентуется среднегеометрическое правдоподобие одного предсказания, усредненное по всем наблюдениям эпохи, которое соответствует среднегеометрической вероятности правильного класса с точки зрения модели.

На рис. 3 показана динамика точности на первых 5 эпохах. На этих эпохах обучался только последний слой модели BERT с одновременным обучением полносвязного слоя, преобразующего эмбеддинг токена CLS. На этом этапе качество модели быстро увеличивается с ростом количества эпох (рис. 4).

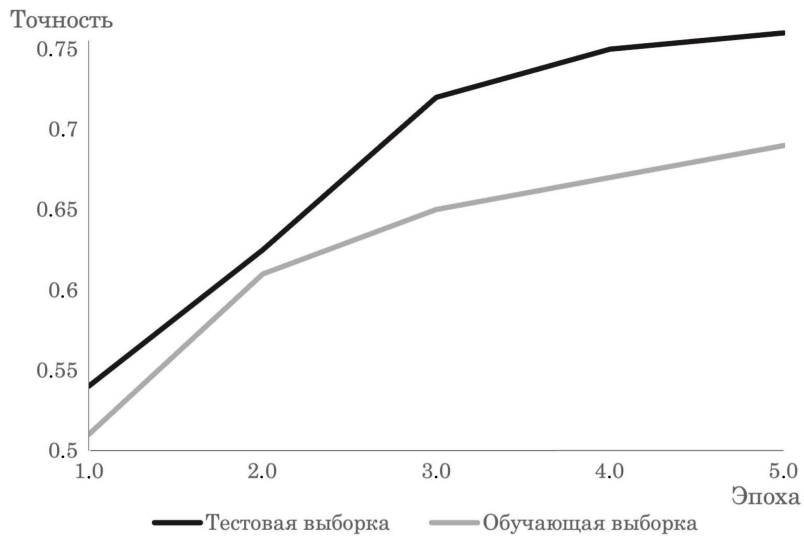

Рис. 3. Динамика точности модели в зависимости от эпохи обучения (1-й этап, один слой)

На следующем этапе обучения обновлялись два последних слоя модели BERT. Скорость роста качества модели замедляется, но он все еще наблюдается. На последнем этапе менялись веса на последних трех слоях модели BERT. Данный этап состоял из 25 эпох. Критерий качества модели представлен на рис. 3. Можно заметить, что качество модели все еще имеет потенциал для дальнейшего обучения. Также отметим, что качество работы модели на тестовом множестве выше, чем на обучающем. Связано это в том числе и с тем, что качество работы модели на обучающем множестве считается в процессе обучения, и, таким образом, качество работы модели на первом батче внутри одной эпохи является ожидаемо ниже, чем на последнем батче внутри той же эпохи. В то же время качество работы модели на тестовом наборе оценивается уже моделью, полученной в конце эпохи.

Итоговое качество работы модели после 35 эпох (5 эпох с одним слоем, 5 эпох с двумя слоями и 25 эпох с 3 слоями) составляет около 85 % по метрике точности и около 70 % по метрике правдоподобия.

Рис. 4. Динамика точности модели в зависимости от эпохи обучения (3-й этап, 3 слоя)

5. Модель инъекции отчетов прокурорских проверок. Опишем модель инъекции предложений отчетов прокурорских проверок в эмбеддинговое пространство сайтов компаний. Предположим, что для каждого предложения с сайта компании можно получить векторное описание посредством модели, оцененной в п. 4. Целью является построение модели, которая каждому предложению из текста прокурорской проверки сопоставляет такой вектор, что расстояние между ними и вектором какого-либо предложения с сайта компании, посчитанным ранее описанной моделью, будет минимальным. Это позволит рассчитать вероятность наличия связи между предложением с сайта компании и предложением прокурорской проверки. Общую структуру процедуры можно описать следующим образом:

1) embedding_1 = Model_1(sentence_site);
2) embedding_2 = Model_2(sentence_procuror);
3) $d = \|$embedding_1-embedding_2$\|$;
4) $p = f(d)$.

В ней Model_1 — это модель, описанная в п. 4. Model_2 — новая модель, также являющаяся подстройкой модели BERT. Расстояние d вычисляется как расстояние Эвклида между эмбеддингами, полученными моделями Model_1 и Model_2. Вероятность p — это вероятность того, что предложение с сайта компании (sentence_site) и предложение из прокурорской проверки (sentence_procuror) относятся к одной компании. Эта вероятность вычисляется на основании расстояния d по формуле, приведенной ранее (см. с. 83).

Для обеспечения объективности оценивания качества работы модели разбиение производилось по ИНН компании. Таким образом, в обучающую выборку вошло 80 % компаний, для которых проводилась хотя бы одна проверка, а в тестовую — 20 %. В качестве предложений с сайта брались все доступные предложения с сайтов металлургических компаний, для которых осуществлялась хотя бы одна проверка.

Для каждой подвыборки были сформированы наблюдения для обучения и тестирования модели. Каждое наблюдение представляет собой пару предложений. Первое предложение в паре — произвольное предложение с сайта компании, второе — произ-

вольное предложение из соответствующей выборки текстов прокурорских проверок. Каждая пара размечается как положительный пример, если предложения с сайта и из текста прокурорских проверок относятся к одному ИНН (т. е. встречаются на сайте и в тексте прокурорских проверок, относящихся к одной компании). Если предложения относятся к разным компаниям, то пример считается отрицательным. Именно эту метку и требуется уметь предсказывать при обучении модели Model_2. Обученная таким образом модель может предсказывать вероятность того, что два предложения относятся к одной компании. Обучение модели производилось аналогично тому, как описано в п. 4.

Обучение проводилось методом подстройки, начиная с модели DeepPavlov/rubert-base-cased-sentence. Для обучения модели обновлялись веса последних трех слоев исходной модели BERT. При этом на первых 5 эпохах обновлялись веса только последнего слоя, на следующих 5 эпохах — только веса последних двух слоев и, наконец, на последующих эпохах — веса последних трех слоев. Модель после 30-й эпохи третьего шага обучения (или 40-й эпохи в общем) показывает точность около 64.7 %. В целом данная точность, с одной стороны, является невысокой, но с другой — все еще позволяет отделять релевантные пары предложений с сайта и из прокурорских проверок с точностью существенно выше шанса.

6. Интерпретация результатов. Для интерпретируемости результатов по полученным предсказанным проверкам были построены паутинные диаграммы. С помощью модели, описанной в п. 5, каждому предложению из текста сайта сопоставляется такой вектор из предложений прокурорских проверок, что расстояние между ними и вектором какого-либо предложения с сайта компании, посчитанным ранее описанной моделью, будет минимальным. Это позволит рассчитать вероятность наличия связи между предложением с сайта компании и предложением прокурорской проверки.

Далее эти предложения «предсказанных» прокурорских проверок подаются на вход модели LDA, описанной в п. 3.1 с целью их категоризации. Таким образом получены представленность каждой из тем (категорий риска) в «предсказанных» предложениях прокурорских проверок, и для каждой компании можно выделить наиболее вероятный риск, согласно предсказаниям модели. Например, для одной из компаний машиностроительной отрасли по результатам модели наиболее вероятен риск проверки соблюдения Трудового кодекса и безопасности труда (рис. 5).

Рис. 5. Пример «предсказанных» рисков на основе текстов сайтов

Разработанная модель позволяет выявить, какие предложения с сайта компании связаны с повышенным риском проверки. Представим элементы текстов, собранных с сайта анализируемой компании, которые были классифицированы моделью как те, которые связаны с определенными типами ESG-рисков (выделены соответствующие рис. 5 типы рисков): «анодируем **магниевые сплавы** для всех отраслей промышленности; данная линия имеет современное качественное исполнение ***с применением долговечных материалов***, проста и надежна в эксплуатации; отличительной особенностью системы являются ее бесшумность и отличные эксплуатационные свойства; наиболее чистые тона получаются при окрашивании **оксидных пленок на алюминии и его сплавах с магнием или марганцем**».

7. Заключение. В рамках предложенного подхода была оценена вероятность того, что предложения с сайта компании и из прокурорской проверки относятся к одной компании. Полученная в рамках второго подхода модель позволяет определять релевантные пары предложений с точностью 64.7 % и создавать «предсказанные» тексты проверок для тех компаний, у которых проверок не было. Таким образом, на основе информации с сайта компании можно предсказывать содержание наиболее вероятных проверок, т. е. наиболее проблемных аспектов деятельности компании с точки зрения ESG-риск-факторов.

Путем разработки подхода для автоматизированного выявления ESG-рисков на основе собранной текстовой информации о компании и данных КНМ можно значительно сократить время и усилия, затрачиваемые на анализ ESG-риск-факторов. Анализ таких данных может помочь выявить потенциальные проблемы и нарушения, связанные с окружающей средой, социальной ответственностью и корпоративным управлением, которые могут повлиять на деловую репутацию компании и инвестиционные решения. Данный подход может быть полезным для инвестиционных фондов, банков, страховых компаний, надзорных ведомств, промышленных компаний (в рамках самодиагностики) и других организаций, которые заинтересованы в оценке ESG-рисков и интеграции ESG-факторов в свои бизнес-модели.

В дальнейшем рекомендуется продолжить исследования в этом направлении, протестировав различные модели BERT для оценки того, какая из них обеспечивает наилучшие результаты в задаче предсказания и анализа ESG-рисков. Предложенный способ можно применять для анализа не только рисков, но и факторов инвестиционной привлекательности, количественный подход к анализу которой обсуждался, например, в работе [29], где применялись регрессионный и кластерный анализы для оценки условий инвестиционной активности регионов.

Литература

1. *Gao W., Liu Z.* Green credit and corporate ESG performance: Evidence from China // Finance Research Letters. 2023. Vol. 55. Art. N. 103940.

2. *Devlin J., Chang M.-W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint. arXiv: 1810.04508, 2019. https://arxiv.org/abs/1810.04805v2

3. *Singh A. K., Zhang Y., Anu.* Understanding the evolution of environment, social and governance research: Novel implications from bibliometric and network analysis // Evaluation Review. 2022. Vol. 47. N 2. P. 350–386.

4. *Pavani K.* A study on risk assessment and financial management on ESG // International Journal of Research Publication and Reviews. 2024. Vol. 5. N 5. P. 3624–3632.

5. *De Giuli M. E., Grechi D., Tanda A.* What do we know about ESG and risk? A systematic and bibliometric review // Corporate Social Responsibility and Environmental Management. 2023. Vol. 31. N 2. P. 1096–1108.

6. *Tiwari R., Sharma N., Sharma N. K.* Categorizing and understanding the evolution of literature on ESG investments: A bibliometric analysis // A Journal of Business Perspective. 2023. https://doi.org/10.1177/09722629.231197574

7. *Kansal P., Malhotra K., Neelam.* Recent trends on Environmental, Social and Governance Research: A bibliometric analysis // Metamorphosis: A Journal of Management Research. 2024. Vol. 23. N 1. P. 7–22.

8. *Ziolo M., Bak I., Spoz A.* Incorporating ESG risk in companies' business models: State of research and energy sector case studies // Energies. 2023. Vol. 16. N 4. Art. N 1809.

9. *Augustin B., Julsain H., Sager M.* Integrating ESG risk analysis into a macro investment strategy // CIBC Asset Management Team Report — CIBC, 2021. URL: https://www.cibc.com/en/asset-management/insights/responsible-investing/integrating-esg-risk-analysis.html (дата обращения: 15 ноября 2024 г.).

10. *Gallucci C., Santulli R., Lagasio V.* The conceptualization of Environmental, Social and Governance risks in portfolio studies: A systematic literature review // Socio-economic Planning Sciences. 2022. Vol. 84. Art. N 101382.

11. *Sokolov A., Mostovoy J., Ding J., Seco L.* Building machine learning systems for automated ESG scoring // The Journal of Impact and ESG Investing. 2021. Vol. 1. N 3. P. 39–50.

12. *Sokolov A., Mostovoy J., Ding J., Seco L.* Building machine learning systems for automated ESG scoring // The Journal of Impact and ESG Investing. 2021. Vol. 1. Iss. 3. P. 39–50. https://doi.org/10.3905/jesg.2021.1.010

13. *Luccioni A., Baylor E., Duchene N.* Analyzing sustainability reports using natural language processing // arXiv preprint. arXiv: 2011.08073, 2020. https://arxiv.org/abs/2011.08073v2

14. *Yim T. Y., Zhang Y., Tan W., Lam T.-W., Yiu S. M.* Meticulously analyzing ESG disclosure: A data-driven approach // 2023 International Conference on Big Data (IEEE 2023). 2023. P. 2884–2889.

15. *Yang W., Rong X.* Duration dynamics: Fin-turbo's rapid route to ESG impact insight // Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing (FinNLP). Torino, Italia: Association for Computational Linguistics, 2024. P. 188–196. URL: https://aclanthology.org/2024.finnlp-1.18/ (дата обращения: 15 ноября 2024 г.).

16. *Ruberg N., Pereira R. B., Stein M. L.* GreenAI — An NLP approach to ESG financing // Anais do II Brazilian workshop on artificial intelligence in finance (BWAIF 2023). Sociedade Brasileira de Computacao. 2023. P. 37–48.

17. *Schimanski T., Reding A., Reding N., Bingler J., Kraus M., Leippold M.* Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication // Finance Research Letters. 2024. Vol. 61. Art. N 104979. https://doi.org/10.1016/j.frl.2024.104979

18. *Hernandez W., Tylinski K., Moore A., Roche N., Vadgama N., Treiblmaier H., Shangguan J., Tasca P., Xu J.* Evolution of ESG-focused DLT research: An NLP analysis of the literature // arXiv preprint. arXiv: 2308.12420, 2023. https://arxiv.org/abs/2308.12420v3

19. *Mehra S., Louka R., Zhang Y.* ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices // Computer Science & Information Technology. 2022. P. 183–190. https://doi.org/10.5121/csit.2022.120616

20. *Lee H., Lee S. H., Park H., Kim J. H., Jung H. S.* ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models // Heliyon. 2024. Vol. 10. Iss. 4. Art. N e26404. https://doi.org/10.1016/j.heliyon.2024.e26404

21. *Pontes E. L., Benjannet M., Ming L. K.* Leveraging BERT language models for multi-lingual ESG issue identification // Proceedings of 5th Workshop on Financial Technology and Natural Language Processing and the 2nd Multimodal AI For Financial Forecasting (FinNLP). Macao: Association for Computational Linguistics, 2023. P. 121–126. URL: https://aclanthology.org/2023.finnlp-1.13/ (дата обращения: 15 ноября 2024 г.).

22. *Kannan N., Seki Y.* Textual evidence extraction for ESG scores // Proceedings of 5th Workshop on Financial Technology and Natural Language Processing and the 2nd Multimodal AI For Financial Forecasting (FinNLP). Macao: Association for Computational Linguistics, 2023. P. 45–54. URL: https://aclanthology.org/2023.finnlp-1.4/ (дата обращения: 15 ноября 2024 г.).

23. *Goel T., Chauhan V., Sangwan S., Verma I., Dasgupta T., Dey L.* TCS WITM 2022@FinSim4-ESG: Augmenting BERT with linguistic and semantic features for ESG data classification // Proceedings of 4th Workshop on Financial Technology and Natural Language Processing (FinNLP). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. P. 235–242. URL: https://aclanthology.org/2022.finnlp-1.32/ (дата обращения: 15 ноября 2024 г.).

24. *Banerjee N., Sarkar A., Chakraborty S., Ghosh S., Naskar S.* Fine-tuning language models for predicting the impact of events associated to financial news articles // Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from

Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing (FinNLP). Torino, Italia: Association for Computational Linguistics, 2024. P. 244–247. URL: https://aclanthology.org/2024.finnlp-1.25/ (дата обращения: 15 ноября 2024 г.).

25. *Pasch S., Ehnes D.* NLP for responsible finance: Fine-tuning transformer-based models for ESG // 2022 International Conference on Big Data (IEEE 2022). 2022. Vol. 33. P. 3532–3536.

26. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.* Attention is all you need // Advances in Neural Information Processing Systems. 2017. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 15 ноября 2024 г.)

27. *Reimers N., Gurevych I.* Sentence embeddings using Siamese BERT-Networks // Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019. P. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

28. *Kuratov Y., Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint. arXiv: 1905.07213, 2023. https://arxiv.org/abs/1905.07213v1

29. *Ци Д., Буре В. М.* Исследование инвестиционной привлекательности на основе кластерного анализа // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 2. С. 199–211. https://doi.org/10.21638/11701/spbu10.2023.206

Статья поступила в редакцию 7 марта 2024 г.
Статья принята к печати 12 декабря 2024 г.

Контактная информация:

Бузмаков Алексей Владимирович — канд. физ.-мат. наук; https://orcid.org/0000-0002-9317-8785, avbuzmakov@hse.ru

Кирпищиков Дмитрий Андреевич — https://orcid.org/0000-0003-3440-5842, dakirpishchikov@hse.ru

Найденова Юлия Николаевна — канд. экон. наук; https://orcid.org/0000-0002-5838-1331, yunaydenova@hse.ru

Паклина София Николаевна — канд. экон. наук; https://orcid.org/0000-0001-9666-989X, snpaklina@hse.ru

Паршаков Петр Андреевич — канд. экон. наук; https://orcid.org/0000-0002-1805-2680, pparshakov@hse.ru

Соломатин Роман Игоревич — https://orcid.org/0009-0004-0559-9910, risolomatin@gmail.com

Сотириади Назар Сергеевич — sotiriadi.n.s@sberbank.ru

The modification of the SBERT language model for identifying ESG risks based on textual data from companies and supervisory activities

A. V. Buzmakov[1], *D. A. Kirpishchikov*[1], *Yu. N. Naidenova*[1], *S. N. Paklina*[1], *P. A. Parshakov*[1], *R. I. Solomatin*[1], *N. S. Sotiriadi*[2]

[1] HSE University, 37, bul. Gagarina, Perm', 614000, Russian Federation
[2] PJSC Sberbank, 19, ul. Vavilova, Moscow, 117312, Russian Federation

For citation: Buzmakov A. V., Kirpishchikov D. A., Naidenova Yu. N., Paklina S. N., Parshakov P. A., Solomatin R. I., Sotiriadi N. S. The modification of the SBERT language model for identifying ESG risks based on textual data from companies and supervisory activities. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2025, vol. 21, iss. 1, pp. 75–91. https://doi.org/10.21638/spbu10.2025.106 (In Russian)

An approach has been developed to identify risks associated with companies' environmental impact, social responsibility, and governance quality (Environmental, Social, and Governan-

ce — ESG risks) based on textual information about the company. To achieve this, a modification of the SBERT language model is proposed with a clearly defined distance function for the embedding space. The model is trained on data from supervisory activities and texts of corporate websites. An example of interpretation of the model's result is provided.

Keywords: ESG, natural language processing model, model training, topic modeling, website.

References

1. Gao W., Liu Z. Green credit and corporate ESG performance: Evidence from China. *Finance Research Letters*, 2023, vol. 55. art. no. 103940.

2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint. arXiv: 1810.04508*, 2019. https://arxiv.org/abs/1810.04805v2

3. Singh A. K., Zhang Y., Anu. Understanding the evolution of environment, social and governance research: Novel implications from bibliometric and network analysis. *Evaluation Review*, 2022, vol. 47, no. 2, pp. 350–386.

4. Pavani K. A study on risk assessment and financial management on ESG. *International Journal of Research Publication and Reviews*, 2024, vol. 5, no. 5, pp. 3624–3632.

5. De Giuli M. E., Grechi D., Tanda A. What do we know about ESG and risk? A systematic and bibliometric review. *Corporate Social Responsibility and Environmental Management*, 2023, vol. 31, no. 2, pp. 1096–1108.

6. Tiwari R., Sharma N., Sharma N. K. Categorizing and understanding the evolution of literature on ESG investments: A bibliometric analysis. *A Journal of Business Perspective*, 2023. https://doi.org/10.1177/09722629.231197574

7. Kansal P., Malhotra K., Neelam. Recent trends on Environmental, Social and Governance Research: A bibliometric analysis. *Metamorphosis: A Journal of Management Research*, 2024, vol. 23, no. 1, pp. 7–22.

8. Ziolo M., Bak I., Spoz A. Incorporating ESG risk in companies' business models: State of research and energy sector case studies. *Energies*, 2023, vol. 16, no. 4, art. no. 1809.

9. Augustin B., Julsain H., Sager M. Integrating ESG risk analysis into a macro investment strategy. *CIBC Asset Management Team Report — CIBC*, 2021. Available at: https://www.cibc.com/en/asset-management/insights/responsible-investing/integrating-esg-risk-analysis.html (accessed: November 15, 2024).

10. Gallucci C., Santulli R., Lagasio V. The conceptualization of Environmental, Social and Governance risks in portfolio studies: A systematic literature review. *Socio-economic Planning Sciences*, 2022, vol. 84, art. no. 101382.

11. Sokolov A., Mostovoy J., Ding J., Seco L. Building machine learning systems for automated ESG scoring. *The Journal of Impact and ESG Investing*, 2021, vol. 1, no. 3, pp. 39–50.

12. Sokolov A., Mostovoy J., Ding J., Seco L. Building machine learning systems for automated ESG scoring. *The Journal of Impact and ESG Investing*, 2021, vol. 1, iss. 3, pp. 39–50. https://doi.org/10.3905/jesg.2021.1.010

13. Luccioni A., Baylor E., Duchene N. Analyzing sustainability reports using natural language processing. *arXiv preprint. arXiv: 2011.08073*, 2020. https://arxiv.org/abs/2011.08073v2

14. Yim T. Y., Zhang Y., Tan W., Lam T.-W., Yiu S. M. Meticulously analyzing ESG disclosure: A data-driven approach. *2023 International Conference on Big Data (IEEE 2023)*, 2023, pp. 2884–2889.

15. Yang W., Rong X. Duration dynamics: Fin-turbo's rapid route to ESG impact insight. *Proceedings of Joint Workshop of the 7^{th} Financial Technology and Natural Language Processing, the 5^{th} Knowledge Discovery from Unstructured Data in Financial Services, and the 4^{th} Workshop on Economics and Natural Language Processing (FinNLP)*. Torino, Italia, Association for Computational Linguistics Publ., 2024, pp. 188–196. Available at: https://aclanthology.org/2024.finnlp-1.18/ (accessed: November 15, 2024).

16. Ruberg N., Pereira R. B., Stein M. L. GreenAI — An NLP approach to ESG financing. *Anais do II Brazilian Workshop on Artificial Intelligence in Finance (BWAIF 2023)*. Sociedade Brasileira de Computacao, 2023, pp. 37–48.

17. Schimanski T., Reding A., Reding N., Bingler J., Kraus M., Leippold M. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 2024, vol. 61, art. no. 104979. https://doi.org/10.1016/j.frl.2024.104979

18. Hernandez W., Tylinski K., Moore A., Roche N., Vadgama N., Treiblmaier H., Shangguan J., Tasca P., Xu J. Evolution of ESG-focused DLT research: An NLP analysis of the literature. *arXiv preprint. arXiv: 2308.12420*, 2023. https://arxiv.org/abs/2308.12420v3

19. Mehra S., Louka R., Zhang Y. ESGBERT: Language model to help with classification tasks

related to companies' environmental, social, and governance practices. *Computer Science & Information Technology*, 2022, pp. 183–190. https://doi.org/10.5121/csit.2022.120616

20. Lee H., Lee S. H., Park H., Kim J. H., Jung H. S. ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon*, 2024, vol. 10, iss. 4, art. no. e26404. https://doi.org/10.1016/j.heliyon.2024.e26404

21. Pontes E. L., Benjannet M., Ming L. K. Leveraging BERT language models for multi-lingual ESG issue identification. *Proceedings of 5^{th} Workshop on Financial Technology and Natural Language Processing and the 2^{nd} Multimodal AI For Financial Forecasting (FinNLP)*. Macao, Association for Computational Linguistics Publ., 2023, pp. 121–126. Available at: https://aclanthology.org/2023.finnlp-1.13/ (accessed: November 15, 2024).

22. Kannan N., Seki Y. Textual evidence extraction for ESG scores. *Proceedings of 5^{th} Workshop on Financial Technology and Natural Language Processing and the 2^{nd} Multimodal AI For Financial Forecasting (FinNLP)*. Macao, Association for Computational Linguistics Publ., 2023, pp. 45–54. Available at: https://aclanthology.org/2023.finnlp-1.4/ (accessed: November 15, 2024).

23. Goel T., Chauhan V., Sangwan S., Verma I., Dasgupta T., Dey L. TCS WITM 2022@FinSim4-ESG: Augmenting BERT with linguistic and semantic features for ESG data classification. *Proceedings of 4^{th} Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Abu Dhabi, United Arab Emirates, Association for Computational Linguistics Publ., 2022, pp. 235–242. Available at: https://aclanthology.org/2022.finnlp-1.32/ (accessed: November 15, 2024).

24. Banerjee N., Sarkar A., Chakraborty S., Ghosh S., Naskar S. Fine-tuning language models for predicting the impact of events associated to financial news articles. *Proceedings of Joint Workshop of the 7^{th} Financial Technology and Natural Language Processing, the 5^{th} Knowledge Discovery from Unstructured Data in Financial Services, and the 4^{th} Workshop on Economics and Natural Language Processing (FinNLP)*. Torino, Italia, Association for Computational Linguistics Publ., 2024, pp. 244–247. Available at: https://aclanthology.org/2024.finnlp-1.25/ (accessed: November 15, 2024).

25. Pasch S., Ehnes D. NLP for responsible finance: Fine-tuning transformer-based models for ESG. *2022 International Conference on Big Data (IEEE 2022)*, 2022, vol. 33, pp. 3532–3536.

26. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, vol. 30. Available at:
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed: November 15, 2024).

27. Reimers N., Gurevych I. Sentence embeddings using Siamese BERT-Networks. *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9^{th} International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, Association for Computational Linguistics, 2019, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

28. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint. arXiv: 1905.07213*, 2023. http://arxiv.org/abs/1905.07213

29. Qi D., Bure V. M. Issledovanie investitsionnoi privlekatel'nosti na osnove klasternogo analiza [Research of investment attractiveness based on cluster analysis]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2023, vol. 19, iss. 2, pp. 199–211. https://doi.org/10.21638/11701/spbu10.2023.206 (In Russian)

Received: March 7, 2024.
Accepted: December 12, 2024.

A u t h o r s' i n f o r m a t i o n:

Aleksey V. Buzmakov — PhD in Physics and Mathematics; https://orcid.org/0000-0002-9317-8785, avbuzmakov@hse.ru

Dmitriy A. Kirpishchikov — https://orcid.org/0000-0003-3440-5842, dakirpishchikov@hse.ru

Yuliya N. Naidenova — PhD in Economics; https://orcid.org/0000-0002-5838-1331, yunaydenova@hse.ru

Sofiya N. Paklina — PhD in Economics; https://orcid.org/0000-0001-9666-989X, snpaklina@hse.ru

Petr A. Parshakov — PhD in Economics; https://orcid.org/0000-0002-1805-2680, pparshakov@hse.ru

Roman I. Solomatin — https://orcid.org/0009-0004-0559-9910, risolomatin@gmail.com

Nazar S. Sotiriadi — sotiriadi.n.s@sberbank.ru