ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2024, том 520, с. 70–81

ИНФОРМАТИКА

УДК 004.8

ИНДЕКС ЭТИЧНОСТИ РОССИЙСКИХ БАНКОВ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2024 г. М. А. Сторчевой[1], П. А. Паршаков[2,3], С. Н. Паклина[2], А. В. Бузмаков[2], В. В. Кракович[1,2,*]

Представлено академиком РАН А. И. Аветисяном

Получено 20.03.2024 г.
После доработки 23.07.2024 г.
Принято к публикации 30.10.2024 г.

Измерение этичности компании является важным элементом в механизме регулирования поведения участников рынка, поскольку позволяет потребителям и регулирующим органам принимать более эффективные решения, что оказывает дисциплинирующее воздействие на компании. Мы протестировали различные способы машинного анализа отзывов потребителей российских банков и разработали Индекс этичности, который позволяет на основе отзывов потребителей рассчитывать количественную оценку этичности трех сотен российских банков за разные периоды времени с 2005 по 2022 г. Мы использовали метод “мешка слов” на основе Moral Foundations Dictionary (MFD) и обучение модели BERT на основе размеченной экспертами выборки 3 тыс. и 10 тыс. предложений. Полученный индекс был валидизирован на основе количества арбитражных дел с 2005 по 2022 г. (более этичные компании вовлечены в меньшее количество арбитражных дел в качестве ответчика), при этом только модель BERT прошла валидизацию, а модель на основе MFD – нет. Индекс этичности будет полезен как альтернативная метрика по отношению к популярным рейтингам ESG как для теоретических исследований о поведении компаний, так и для практических задач управления репутацией компании и формирования политики регулирования поведения участников рынка.

Ключевые слова: индекс, этичность, искусственный интеллект, NLP, BERT.

DOI: 10.31857/S2686954324060111, **EDN:** KKLCMK

1. ВВЕДЕНИЕ

В последние годы в мире наблюдается рост интереса к концепции устойчивого развития и ответственного инвестирования. Все больше инвесторов обращают внимание не только на финансовые показатели компаний, но и на их деятельность в области охраны окружающей среды, социальной ответственности и корпоративного управления (ESG), для оценки которой разрабатываются рейтинги или рэнкинги ESG [1, 2].

В России система рейтингов ESG также развивается [3, 4], однако их показания часто расходятся в виду различной методологии рейтинговых агентств. Кроме того, большинство существующих рейтингов основаны на данных корпоративной отчетности компаний, которая неизбежно носит приукрашивающий или даже декларативный характер. В этим условиях актуальным является развитие альтернативных способов измерения социальной ответственности компании, построенных на других источниках данных – в частности, на новостях, сообщениях в социальных сетях, отзывах пользователей и т. п. [5].

В данной работе предпринята попытка разработать индекс этичности российских компаний на основе открытых данных с использованием методов обработки естественного языка (NLP). Под этичностью компании мы понимаем следование нормам этичного бизнеса, принятым в обществе: честность при заключении и выполнении соглашений, отказ от злоупотребления рыночной властью, справедливое вознаграждение, уважительная коммуникация и т. п. Этичность компании по отношению к заинтересованным сторонам должна являться составной частью любого рейтинга ESG, поскольку дополняет оценки добросовестности поведения компании, полученных другими путями (напр., через отчеты компании). Предложенная нами модель может использоваться как автономно для оцен-

[1] *Санкт-Петербургская школа экономики и менеджмента, НИУ ВШЭ в Санкт-Петербурге, Санкт-Петербург, Россия*

[2] *Международная лаборатория экономики нематериальных активов, НИУ ВШЭ в Перми, Пермь, Россия*

[3] *Московская школа управления СКОЛКОВО, Москва, Россия*

[*] *E-mail: mstorchevoy@hse.ru*

ки этичности любой компании или построения рейтинга этичности компаний в отрасли, а также как компонент в построении более сложных систем оценки социальной ответственности, например, рейтингов ESG.

Для сбора текстов из открытых источников можно использовать технологии веб-скрейпинга (web-scraping) – скрипты, скачивающие HTML-страницы с нужных сайтов и автоматически преобразующие их в более удобный для последующего анализа формат данных. В дальнейшем для обработки этой текстовой базы данных запускаются другие скрипты, которые сначала обучают модель NLP на эталонной выборке, а затем используют обученную модель расчета индекса этичности по всей совокупности компаний и отзывов. Далее в нашей статье мы сделаем обзор существующих исследований по данной тематике, а затем расскажем о нашем исследовании и его результатах.

2. ОБЗОР ЛИТЕРАТУРЫ

2.1. Измерение этичности через опросы

Попытки измерения этичности компании начались еще в 1970-х, когда начались исследования в области корпоративной социальной ответственности (КСО), однако ученые долгое время испытывали существенные ограничения в доступных данных для измерения этичности. В первых исследованиях они просто опрашивали некоторых стейкхолдеров, что они думают о репутации компании. Похожие исследования проводились также в литературе по маркетингу, где ученые пытались измерить мнение потребителей о социальной ответственности компании и его влияние на покупательское поведение. Например, в работе [6] попытались оценить, что думают потребители о социальной ответственности различных компаний. В работе [7] измерили, как информация о неэтичной трудовой практике компании влияет на лояльность ее клиентов. В работе [8] измеряется мнение потребителей о компании по шести измерениям: корпоративные пожертвования, участие в жизни общества, позиция по проблемам женщин, позиция по вопросам этнических меньшинств, позиция по вопросам геев и позиция по вопросам меньшинств с ограниченными возможностями.

Были попытки создать отдельный индекс этического поведения компании. Например, в работе [9] был предложен индекс CPE (*consumer perceived ethicality*) как субъективное восприятие общей этичности компании. Автор на основе 20-ти углубленных интервью разработал классификацию этических проблем для каждого стейкхолдера (напр., для потребителей этические проблемы могут касаться ценообразования, маркировки, рекламы, таргетинга, качества продукции, обслуживания, свободы выбора). Позже индекс CPE использовался во множестве других исследованиях в области маркетинга [10].

В 1980–90-х годах стали появляться проекты по измерению *корпоративной репутации* компаний путем опроса экспертов и представителей бизнеса. В 1983 г. деловой журнал Fortune начал составлять ежегодный рейтинг *America's Most Admired Companies*. Вскоре за ним последовали другие ведущие деловые издания: в 1991 г. журнал Management Today запустил рейтинг *Britain's Most Admired Companies*, в 1992 г. журнал *Asian Business* запустил рейтинг *Asia's Most Admired Companies*, в 1994 г. газета Financial Times начала составлять рейтинг *Europe's Most Respected Companies*. Во всех случаях опросы строились по ряду общих критериев: качество управления, качество продукции, инновационность, коммуникация, социальная и экологическая ответственность и т. д. Ограничением данной методологии было то, что рейтинги были ориентированы только на крупнейшие компании, а оценка давалась самими бизнесменами, а не заинтересованными сторонами (напр., потребителями), которые могут обладать гораздо более полной информацией о компании.

В 2000 г. консалтинговая фирма Reputation Institute представила свой *индекс корпоративной репутации*, который строился на основе опросов многих заинтересованных сторон, а не только представителей отрасли. Среди вопросов была в том числе оценка социальной и экологической ответственности компании, ее отношение к потребителям и т. д. [11]. Позже этот индекс трансформировался в показатель RepTrack®, при этом почти половина его показателей была связана с этичностью поведения. Однако проблема этого инструмента также заключалась в очень ограниченном охвате компаний – в индексе участвовали только крупнейшие компании.

2.2. Измерение этичности через NLP

В 2010-х годах начинаются попытки измерения этичности компании с помощью NLP (natural language processing). Поскольку объем текстов в интернете растет в геометрической прогрессии, такие инструменты позволяют проводить крупномасштабные исследования, потенциально охватывающие целые рынки и отрасли. Этот подход требует решения двух задач: 1) определить тему текста (например, относящуюся к какому-то компоненту ESG), 2) определить тональность текста (т. е. положительную или отрицательную оценку поведения компании). Для анализа используются несколько типов текстов: 1) документы компании (нефинан-

совый отчет, сайт, внутренние документы), 2) посты и обзоры в социальных сетях, 3) новости.

Одна из первых попыток применить машинное обучение и анализ настроений к онлайн-обзорам потребителей была предпринята в работе [12]. К концу 2020-х гг. уже вышли десятки статей, посвященных анализу текстов потребителей. Например, в работе [13] использовали модель BERT для анализа сообщений пользователей в X (ранее Twitter), а также для определения их топиков по классификации MSCI, а также тональности. В работе [14] проанализировали 10.4 миллиона анонимных отзывов сотрудников с помощью модели с векторным представлением слов, чтобы получить представление о внутренних ESG-практиках компании.

В данной литературе почти нет работ, посвященных измерению этичности банков. В статье [15] осуществляется систематический обзор литературы по 68 статьям с 2017 по 2020 год, посвященной анализу отзывов потребителей на основе NLP. Рассматриваемые статьи изучают самые разные сектора (отели, авиакомпании, рестораны, аэропорты, туризм, искусство и музеи), но среди них нет банковской отрасли. Кроме того, примерно половина данных статей посвящена анализу тональности, а другая половина – выявлению фейковых отзывов или формированию прогнозных рекомендаций, но никто не рассчитывает оценку этичности поведения компаний. В обзоре литературы [16] делается акцент также и на тональности, но нет оценки этичности компании и не анализируется банковская отрасль.

Существует только две статьи, которые частично соответствуют предмету нашего исследования. В работе [17] использовали около 20 тыс. постов в социальных сетях о двух финских банках и нейронную сеть для автоматической классификации постов на основе размеченной вручную выборки. Авторы использовали шесть топиков репутации: качество, надежность, ответственность, успешность, приятность и инновационность и добились распознавания этих топиков с вероятностью 60–70%. В нашем исследовании проводится аналогичный анализ, но на основе нашей собственной классификации топиков, и далее мы строим индекс этического поведения банка, чего не делают авторы данного исследования. Во второй статье [18] применяется текстовый анализ и анализ настроений 30 тыс. отзывов клиентов о 29 ведущих банках, опубликованных на сайте bankbazaar.com с 2014 по 2021 год. Авторы определили основные атрибуты банковских услуг (с помощью набора слов) в отзывах и научились прогнозировать оценку по пятизвездочной шкале, которую выставляют банкам клиенты, размещая свои отзывы на сайте.

Как видно из этого обзора, существуют значительные возможности для применения продвинутых методов NLP к разнообразным неструктурированным текстовым данным с целью получения новых знаний о социальной ответственности и этичности компаний. В нашем исследовании мы разработаем модель ИИ для оценки этичности компаний в банковской отрасли.

3. РАЗРАБОТКА ИНДЕКСА И ВЫБОРКА ДАННЫХ

3.1. Источники информации

В качестве источников данных для анализа были выбраны сайты, посвященные размещению отзывов о деятельности банков, а также сайты, содержащие новости о банках и их страницы в социальных сетях – Банки.ру (https://www.banki.ru), IRecommend.ru (https://irecommend.ru), Отзовик (https://otzovik.com) – за период с марта 2005 по март 2021 года. В выборку были включены 330 банков, которые осуществляли деятельность в РФ на момент сбора данных. Также в рамках проекта была собрана база данных постов российских банков в социальной сети Вконтакте. Основная работа по обучению и валидации моделей была проведена на основе отзывов с портала *Банки.ру* – платформы, предлагающей финансовые онлайн-сервисы и являющейся крупным медиаресурсом в области финансов[1].

3.2. Разметка данных

Первая часть эмпирической части работы заключалась в разметке предложений из отзывов о банках на сайте Banki.ru. Целью этого этапа было формирование базы данных для обучения модели. Разметка проводилась в ручном режиме с выборочной ручной и сплошной автоматизированной проверкой результатов разметки. Разметчики оценивали отдельные предложения из отзывов без возможности прочитать отзыв целиком. Разметчики на основе подготовленной инструкции оценивали этический сентимент (тон) предложений, который принимал следующие значения: “+” в случае, если предложение позитивно оценивает этическую категорию, “–” в случае отрицательной оцен-

[1] Благодаря данной платформе пользователи имеют возможность найти и сравнить между собой различные предложения по финансовым продуктам и отправить заявку на их получение напрямую в банк, микрофинансовую, страховую или инвестиционную компанию. Кроме того, клиенты финансовых организаций могут оставить свои отзывы о деятельности компаний. По данным портала, более 1 млн реальных отзывов было оставлено пользователями об уровне обслуживания и качестве услуг банков и других финансовых организаций.

ки и “?” в случае нейтральной или неопределенной оценки.

Разметка проводилась в два этапа. На первом этапе было размечено 3060 предложений, каждое из которых было независимо оценено тремя разметчиками для обеспечения согласованности и устойчивости оценок. Всего разметку проводили пять разметчиков. Таким образом, каждый разметчик оценил около 1835 предложений. На первом этапе все три разметчика указали одинаковый тон для 71% предложений. Затем было проведено обсуждение спорных случаев, в результате которого еще 4% (122) предложений с категорий “?” были перекодированы в отрицательную или положительную категорию.

На втором этапе проводилась разметка 7000 предложений из верхних 5% по прогнозу вероятности этического сентимента (равное количество случаев положительной и отрицательной оценки этичности), то есть по 4300 предложения на каждого разметчика. Предварительный прогноз сентимента на втором этапе считался на основании модели BERT (подробнее о процедуре обучения в следующих разделах), обученной по первым 3060 размеченным предложениям. На данном этапе согласованность оценок сентимента повысилась. Теперь в 81% предложений три разметчика выбрали одинаковый тон, при этом из них только в 10% этот тон был неопределенным (вместо 30% ранее).

3.3. Обучение BERT

Для оценки сентимента предложений использовалась нейросетевая модель BERT. По причине того, что процесс разметки состоял из двух этапов (3000 предложений и 7000 предложений), по итогам каждого этапа модель обучалась заново. Сначала модель обучалась на 3000 размеченных предложений. Предсказание этой модели в дальнейшем помогло уменьшить долю нейтральных предложений среди следующих 7000 предложений, предложенных для ручной разметки. По окончании второго этапа разметки была обучена новая модель на всех 10000 размеченных предложений. При этом процесс обучения модели BERT совпадал как при обучении первой модели, так и при обучении второй модели. Рассмотрим этот процесс подробнее.

В силу того, что обучение модели BERT с нуля требует значительных вычислительных ресурсов, это обучение проводилось в режиме точной подстройки (fine tuning). В качестве исходной модели была взята модель blanchefort/rubert-base-cased-sentiment-rurewiews[2], обученная на отзывах на русском языке по широкой группе товаров. Данная модель представляет из себя дообученную модель от DeepPavlov, одной из лучших моделей общего назначения для русского языка. Модель blanchefort/rubert-base-cased-sentiment-rurewiews была выбрана как модель, обученная на выборке данных наиболее близких к используемым в рамках данного исследования. Данная модель обучена для русского языка, а также обучена на отзывах, которые представляют собой специальный неформальный жанр короткого текстового сообщения с высказыванием личного мнения автора по поводу некоторого товара. Однако в отличие от этой модели в данном исследовании анализируются отзывы о банковских услугах. Кроме того, цель нашей модели – предсказание этического сентимента, в отличие от исходной модели, которая предсказывает обычный сентимент. По этой причине требовалось дообучить модель.

Дообучение модели проводилось в несколько этапов с постепенным “размораживанием” весов модели. Под “замороженными” весами модели подразумеваются веса, которые не меняются в процессе оптимизации модели. Рассмотрим процесс оптимизации подробнее. Модель BERT состоит из 12-ти соединенных последовательно слоев архитектуры трансформер. Этот подход для каждого входного слова, а также специальных токенов, соответствующих началу (‘[CLS]’) и концу (‘[SEP]’) предложения, рассчитывает вектор эмбеддинга размерности 768. Затем эмбеддинг для токена ‘[CLS]’ пропускается через полносвязный слой, преобразуя эти 768 значений в 3 значения для этического сентимента, каждый из которых соответствует своему классу. Полученные значения с помощью функции SoftMax преобразуются в вероятности, показывающие, что с точки зрения модели то или иное предложение принадлежит каждому из 3 классов. В дальнейшем эти вероятности используются внутри функции потерь известной как кросс-энтропия:

$$Loss\,(Y,\widehat{p}) = \frac{1}{N}\sum_{i=0}^{N} Y_i \cdot \log \widehat{p}_i. \qquad (1)$$

По результатам расчета этой функции потерь можно рассчитать градиент по всем незамороженным весам модели. Этот градиент применяется в оптимизаторе для обновления всех незамороженных весов. При обучении использовался алгоритм стохастического градиентного спуска с размером батча 16 предложений, то есть на этапе одного шага оптимизации выбирались случайные 16 предложений неиспользованные ранее, на которых и рассчитывалась функция потерь с последующим расчетом градиента. Размер батча выбирался максимально допустимым с учетом ограничения видеопамяти на расчетной машине.

[2]https://huggingface.co/blanchefort/rubert-base-cased-sentiment-rurewiews

Было замечено, что достаточно трех эпох для сходимости модели для одного множества замороженных весов (здесь эпоха – это множество шагов оптимизации, за которое все предложения обучающей выборки обрабатываются ровно один раз). Соответственно, в процессе обучения для каждого набора замороженных весов использовалось ровно три эпохи. Для достижения максимального качества итоговой модели применялся подход с постепенным размораживанием весов модели начиная с последнего слоя. Так, на первом этапе обучения модели были разморожены веса только последнего полносвязного слоя. Через три эпохи были дополнительно разморожены веса 12-го слоя архитектуры трансформер, и снова обучение длилось три эпохи. Далее следовал 11-й слой, 10-й слой и т. д. до 5-го слоя. Первые 4 слоя архитектуры трансформер не размораживались ввиду объема видеопамяти на расчетной машине.

Процесс обучения всегда проводился на обучающей выборке данных, содержащей 80% от исходной выборки. Все остальные наблюдения использовались в‘качестве тестовой выборки данных.

3.4. Результаты обучения модели: 3 000 предложений

Модель, обученная на 3 000 предложений, дала точность около 60%. Эта точность была показана как на тестовой, так и на обучающей выборке данных, что говорит об отсутствии эффекта переобучении. При этом, как было показано ранее, распределение классов в соответствующей выборке данных было неравномерным. Ниже показана матрица сопряженности, т. е. как соотносились предсказанные и фактические классы на всей выборке данных[3]. Как видно из таблицы 1, в выборке данных преобладают отрицательные и нейтральные отзывы. Также модель предпочитает нейтральные отзывы по разметке помечать как отрицательные.

Такую модель на практике было бы затруднительно использовать, поэтому она использовалась при составлении второго задания для разметчиков. С помощью этой модели для разметки отбирались только те предложения, которые были предсказаны моделью как положительные или отрицательные. Это позволило отбалансировать выборку на втором этапе разметки, в результате чего получалось меньше нейтральных отзывов.

3.5. Результаты обучения модели: 10 000 предложений

Модель, обученная на 10 000 предложений с более качественно отбалансированными классами, показала как на обучающей, так и на тестовой выборке качество около 85%. В этом случае также не наблюдалось переобучения. Ниже показана матрица сопряженности для всех 10 000 уникальных предложений[4]. Данная модель отличает предложения с отрицательным и положительным этическим сентиментом. Основные ошибки модели связаны с нейтральным классом. Суммируя, 85% точности является достаточно высоким показателем, чтобы после агрегации предсказаний по всем отзывам получать данные, отражающие этическое восприятие банков клиентами.

3.6. Агрегирование оценок

Следующим этапом при построении индекса этичности стало агрегирование оценок отдельных предложений и текстов в единый числовой показатель. В данном случае присутствует четыре уровня агрегирования:

1) с уровня отдельных предложений в отзыве на уровень всего отзыва. Например, в отзыве может быть 10 предложений, из которых два содержат положительную оценку банка, три предложения – нейтральны, и еще пять предложений содержат отрицательную оценку;

2) с уровня разных отзывов на одном ресурсе до уровня всего ресурса. Например, на ресурсе Banki.ru для банка X было найдено 2 положительных отзыва, 3 нейтральных и 5 отрицательных;

3) с разных ресурсов одного типа в индекс этичности по данному типу ресурса. Например, индекс этичности для банка X на основе Banki.ru равен 4.2, на основе iRecommend.ru – 3.8, на основе сайта Otzovik.ru – 3.5;

4) агрегирование индексов с ресурсов разного типа в один индекс. Например, индекс этичности по сайтам отзывов равен 4.2, индекс этичности по СМИ равен 4.5, а индекс этичности по социальным сетям равен 4.0.

Существующие исследования [19–22] не предлагают единого подхода к агрегированию данных о сентименте ни для общих случаев, ни для специфики текстовых отзывов клиентов. Выбор подхода имеет вероятность сказаться на точности итоговых данных. Иллюстрацией затруднений с агрегированием сентимента отдельных предложений может быть следующий пример гипотетического от-

[3] Важно отметить, что речь идет о 3060 уникальных предложений, каждое из которых было размечено ровно три раза. Таким образом общее количество размеченных предложений – 9180. Тестирование и предсказание проводилось именно на этом множестве предложений с дубликатами. Это позволило учесть субъективность работы разметчиков, т. к. модель при обучении на противоречивых предложениях училась также понижать свою степень уверенности в ситуациях сложных для классификации человеком. Отдельно отметим тот факт, что при разбиении на обучающее и тестовое множество все дубликаты каждого предложения всегда попадали только в одно из этих множеств.

[4] Речь идет о 10229 уникальных предложений, каждое из которых было размечено по три раза.

Таблица 1. Тональность отзывов в обучающей и тестовой выборках (3 тыс. предложений)

	Класс в обучающей выборке		
Предсказанный класс	Позитивный	Нейтральный	Негативный
Позитивный	463	172	97
Нейтральный	259	830	657
Отрицательный	278	2248	4176

Таблица 2. Тональность отзывов в обучающей и тестовой выборках (10 тыс. предложений)

	Класс в обучающей выборке		
Предсказанный класс	Позитивный	Нейтральный	Негативный
Позитивный	7258	917	357
Нейтральный	454	3950	795
Отрицательный	261	1626	15075

зыва: “Банк совершенно отвратительный! [негативное предложение по смыслу и по разметке] Мне потребовалось совсем мало времени чтобы понять, насколько может быть приятно оттуда уйти. [негативное предложение по смыслу, но положительное по разметке, т. к. “приятно” будет вероятно иметь больший вес при оценке сентимента]”.

Как правило, перед исследователями стоит два выбора при планировании процедуры агрегирования данных сентимента:

- уровень агрегирования (предложение, отзыв, сущность – банк или другая организация / событие и т. п.) [19, 20];
- механизм агрегирования: невзвешенные оценки (simple averaging, majority voting) или взвешенные оценки на уровне леммы, слова, предложения, отзыва и сущности. В ряде случаев [21, 22] предлагаются и автоматические подходы к вычислению весов, в частности через методы сетевого анализа.

Для реализации целей данного исследования было проведено агрегирование пятью разными способами, для того чтобы получить возможность сравнить результаты и сделать выбор. Таблица 3 содержит краткое обобщение предложенных механизмов агрегирования сентиментов в отзывах.

Ниже приведена матрица корреляции индексов этичности банков, рассчитанных по пяти методам, описанным выше.

3.7. Индекс этичности на основе MFD

Для сравнения индекса, рассчитанного на основе модели BERT, был также рассчитан индекс на основе распространенной теории моральных оснований (moral foundations theory). На данной теории основан Словарь моральных основ (англ. Moral Foundations Dictionary, MFD) [23], в котором выделяются пять базовых моральных принципов (moralfoundations.org):

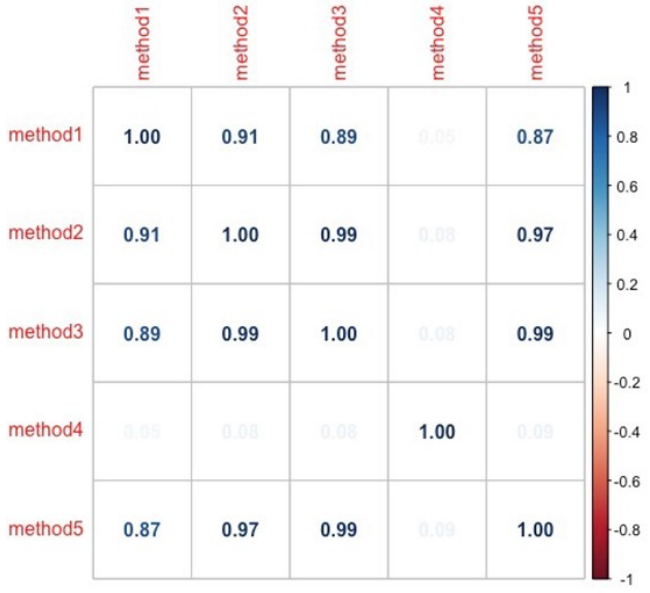

Рис 1. Матрица корреляции индексов этичности по пяти методам агрегирования.

1) забота (англ. care) – связан с процессом формирования систем привязанности и способностью чувствовать боль других (противоположность – вред, англ. harm);

2) справедливость (англ. fairness) – связан с эволюционным процессом взаимного альтруизма (противоположность – жульничество);

3) лояльность – связан с человеческой тенденцией создавать коалиции и переходить между ними (противоположность – предательство);

4) авторитет – связан с формированием иерархических социальных взаимодействий (противоположность – свержение, англ. subversion);

Таблица 3. Различные механизмы агрегирования оценок

Способ	Оценка предложения	Оценка отзыва	Оценка банка
1	Простое округление	Простое округление	Net Promoter Score
2	Простое округление	Взвешивание по количеству предложений	
3	Точная вероятность	Взвешивание по количеству предложений	
4	Точная вероятность	Взвешивание по количеству предложений и по количеству отзывов	
5	Точная вероятность с учетом длины предложения	Взвешивание по количеству предложений	

5) чистота – связан с психологией отвращения и осквернения (противоположность – порочность).

В Словаре моральных основ выделяется 11 категорий: перечисленные выше моральные принципы, их противоположности, а также общая категория моральности (англ. morality). Словарь составлен в форме масок слов. Например, “sympath*”, относящийся к категории “забота”, означает, что в словарь включаются слова, начинающиеся на “sympath”: “sympathy”, “sympathetic”, “sympathize” и др. Мы перевели словарь на русский язык, стараясь сохранить оригинальную универсальность, чтобы все слова, соответствующие маске из английского словаря, а также их синонимы и различные падежные формы в русском языке, попадали под ту же категорию. Например, “sympath*”, упомянутый ранее, мы перевели как “сочувств*”; этой маске соответствуют слова: “сочувствие”, “сочувственно”, “сочувственник”, “сочувственный”, “сочувствовать”, “сочувствующая”, “сочувствующий” и их падежные формы. Все переводы были проверены и утверждены экспертом по этике.

В основе расчета индекса этичности банков, отзывов и предложений на основе MFD словаря лежит идея о том, что индекс этичности предложения может быть определен на основании этичности входящих в него слов, индекс этичности отзыва на основании индекса этичности составляющих его предложений, а индекс этичности банка через индекс этичности отзывов.

Для определения этичности слов каждое слово из словаря MFD было охарактеризовано либо как этичное, либо неэтичное, в соответствие с англоязычной классификацией слов на 6 групп: harm, fairness, ingroup, authority, purity, morality. Каждое слово из словаря было переведено на русский язык, причем допускалось несколько русскоязычных переводов одному англоязычному слову. В результате имеющимся 322 англоязычным словам были сопоставлены 353 русскоязычных, примеры переводов представлены в таблице 4.

Далее для каждого предложения рассчитывалось количество упоминаний переведенных слов.

Таблица 4. Пример перевода слов для словаря MFD

original	foundation	translation
warring	harm	воюющ* воинств* агресс*
fight*	harm	борьб* ссор* конфликт* вражд*
violen*	harm	жесток*
hurt*	harm	ранит* страдан*
kill	harm	уби*

Среди всех 2 132 706 предложений в 35% случаев (756 935 предложений) упоминалось хотя бы одно слово из словаря. В результате для каждого предложения было посчитано количество этичных и неэтичных слов, что позволило применить различные способы агрегации данной информации в индекс этичности отзыва и банка в целом.

Для агрегирования тональности (этичности) предложений в индекс банка были применены два некоррелированных способа агрегации, описанные выше: взвешенная по количеству предложений оценка majority voting и взвешенная по количеству предложений и по количеству отзывов оценка. Сравнительный анализ результатов показал, что в первом случае наибольшим индексом этичности обладают банки, в немногочисленных отзывах которых чаще всего присутствуют только этичные слова, в то время как второй индекс придает большее значение не только самому уровню этичности, но и то, в каком количестве отзывов относительно других банков встречаются индикаторы этичности, что неизбежно выводит на первые места банки с большим количеством отзывов.

3.8. Валидация индекса этичности для банков

С целью валидации индекса этической репутации, рассчитанного различными способами, был осуществлен сбор данных по арбитражным делам, в которых банки выступали ответчиками. Данный показатель был выбран, так как арбитражные дела в отношении банков возникают в том числе в случаях неэтичного поведения банков по отношению

к своим клиентам и партнерам. Таким образом, количество арбитражных дел выступает в качестве прокси-показателя неэтичного поведения банков. Данные были собраны за период с 2005 по 2022 годы. Из 330-ти анализируемых банков, 145 банков выступали ответчиками не менее чем по 10-ти арбитражным делам, 57 банков выступали ответчиками хотя бы по 100 арбитражным делам. Для дальнейшего анализа были исключены банки, которые не выступали в качестве ответчиков по арбитражным делам за рассматриваемый период.

Для того чтобы оценить качество предлагаемого индекса, была рассчитана корреляционная связь между различными вариантами расчета индекса и количеством арбитражных дел как прокси-показателя этичности компаний. Результаты корреляционного и графического анализа представлены в таблицах и на рисунках ниже.

Всего было рассмотрено четыре различных варианта расчета индекса. Во-первых, тональность текстов, лежащих в основе индекса, предсказывалась двумя способами: по модели BERT и словарю MFD. Во-вторых, после определения тональности анализируемых текстов использовалось два способа расчета индексов.

Способ № 1 предполагает расчет индекса по формуле:

$$index = \text{доля позитивных отзывов} - \text{доля негативных отзывов}. \quad (2)$$

Способ № 2 предполагает коррекцию $index$ по формуле:

$$index_{safe} = \big(2(index - mean(index) > 0) - 1\big) \times \big(\max[|index - \text{mean}(index)| - sd(index), 0]\big). \quad (3)$$

Этот индекс представляет из себя консервативную версию первого индекса, учитывающего качество расчета индекса. Так, например, если первый индекс был рассчитан всего по паре отзывов, то мы не можем доверять такой оценке. Тогда, вычитая стандартное отклонение, мы получим более безопасную оценку.

В таблице 5 представлены коэффициенты корреляции и уровни статистической значимости между индексами, рассчитанными разными способами, и количеством арбитражных дел. По результатам корреляционного анализа индекс, рассчитанный двумя способами на основе тональности текстов, определенной по модели BERT, оказался обратно и значимо связан с количеством арбитражных дел. В этом случае такой способ расчета индекса подтверждает, что он отражает уровень этичности деятельности банка. В случае индекса, рассчитанного по словарю MFD и способу № 1, значимой взаимосвязи не было обнаружено. Однако в случае расчета по способу № 2 корреляция значимая и положительная, что свидетельствует о том, что данный способ отражает этичность банка, оцененную через прокси-показатель количества арбитражных дел.

Кроме того, была протестирована гипотеза об отложенном эффекте неэтичного поведения банков. Предполагается, что претензии и недовольство клиентов и партнеров могут отражаться через количество арбитражных дел не сразу, а спустя некоторое время, так как процесс подготовки судебного иска требует временных затрат. В последнем столбце таблицы 1 представлены корреляционные коэффициенты между индексом этичности, рассчитанным различными способами, и количеством арбитражных дел в следующем квартале. В целом, гипотеза подтвердилась, в абсолютном выражении коэффициенты корреляции выше, чем в случае поквартального сравнения значения индекса и количества арбитражных дел. Другими словами, менее этичное поведение банка, измеренное через предложенный индекс в квартале t, ассоциируется с большим количеством арбитражных дел в квартале $t + 1$. Данный результат важен также в контексте дальнейшего анализа влияния этичности банка, измеренного с помощью предлагаемого индекса, на финансовые результаты банка, поскольку выявленный отложенный эффект неэтичного поведения может также наблюдаться в отношении финансовых результатов.

В дополнении к вышеописанным результатам взаимосвязь между индексом этичности, рассчитанным различными способами, и арбитражными делами была проанализирована в разрезе различных источников текстовой информации о деятельности банков. В качестве источников текстовой информации использовались три типа ресурсов — сайты, агрегирующие отзывы клиентов, новости и страницы банков в социальных сетях. Представляется важным определить, влияет ли источник и тип текстовой информации для расчета индекса на его взаимосвязь с количеством арбитражных дел. Коэффициенты корреляции между индексом и количеством арбитражных дел в разрезе способов расчета и источников информации представлены в таблице 6. Поскольку предшествующий анализ показал, что наилучшие результаты показывает индекс, учитывающий тональность, определенной по модели BERT, данный корреляционный анализ представлен только для этой модели.

Как видно из таблицы 6, для всех источников информации сохранилась отрицательная и статистически значимая взаимосвязь между индексом на основе модели BERT и количеством арбитражных дел вне зависимости от способа расчета индекса. Если сравнивать различные источники текстовой

Таблица 5. Коэффициенты корреляции между количеством арбитражных дел и индексом этичности, рассчитанным различными способами

Модель оценки тональности	Способ расчета индекса	Количество арбитражных дел	Количество арбитражных дел в следующем квартале
Модель BERT	Способ № 1	–0.1282***	–0.1389***
	Способ № 2	–0.1553***	–0.1698***
Словарь MFD	Способ № 1	0.0257	0.0316*
	Способ № 2	0.1034***	0.1064***

* – уровень значимости 10%, ** – уровень значимости 5%, *** – уровень значимости 1%

Таблица 6. Коэффициенты корреляции между количеством арбитражных дел и индексом этичности, рассчитанным различными способами, в разрезе источников информации

Источник текстовой информации	Способ расчета индекса	Количество арбитражных дел	Количество арбитражных дел в следующем квартале
Отзывы	Способ № 1	–0.0385***	–0.0437***
	Способ № 2	–0.1177***	–0.1234***
Новости	Способ № 1	–0.0395*	–0.0332
	Способ № 2	–0.1223***	–0.1089***
Социальные сети	Способ № 1	–0.1308***	–0.1224***
	Способ № 2	–0.1901***	–0.1842***

* – уровень значимости 10%, ** – уровень значимости 5%, *** – уровень значимости 1%

информации, более высокие корреляционные коэффициенты в абсолютном выражении наблюдаются в случае социальных сетей. Это означает, что предлагаемый индекс, рассчитанный для текстов из социальных сетей, в большей степени связан с количеством арбитражных дел. Кроме того, важно отметить, что вне зависимости от источника информации индекс этичности, рассчитанный по способу № 2, в большей степени связан с количеством арбитражных дел, используемым в качестве прокси-показателя этичности банка, чем индекс, рассчитанный способом № 1.

Важно отметить, что и в разрезе различных источников информации может наблюдаться отложенный эффект неэтичного поведения банков. Однако взаимосвязь между индексом и количеством арбитражных дел в следующем квартале сильнее по сравнению с корреляцией между индексом и количеством дел в текущем квартале только в случае отзывов. Это может быть связано с тем, что новости, как правило, отражают значимые события, которые приводят к незамедлительным последствиям. В случае с социальными сетями можно предположить, что информация там обновляется быстрее остальных источников и поэтому в меньше степени может быть связана с показателями в следующем квартале.

В процессе конструирования индекса было важно учесть тот факт, что для менее известных банков наблюдается меньшее количество текстовой информации в виде отзывов, новостей и сообщений в социальных сетях. Для того чтобы оценить, как меняется взаимосвязь между индексом, рассчитанным различными способами, и количеством арбитражных дел в таком случае была рассчитана корреляционная взаимосвязь в зависимости от фильтра на количество единиц текстовой информации (отзывов, новостей, сообщений в социальных сетях). Данный анализ представлен на рис. 5 и 6. Каждая точка на рисунке отображает значение корреляционного коэффициента в зависимости от значения фильтра на количество текстовой информации. Таким образом, представляется возможным оценить, как меняется сила взаимосвязи между показателями, если мы ставим условие на минимальное количество текстовой информации для расчета корреляции.

Рис. 2 и 3 отражают коэффициенты корреляции между количеством арбитражных дел и индекса этичности по словарю BERT, рассчитанного способом № 1 и № 2 соответственно, в зависимости от *фильтра на количество единиц текстовой информации*. Как видно из графиков, в обоих случаях при ужесточении требования на количество текстовой информации сила взаимосвязи между показателями усиливается. Все представленные коэффициенты корреляции значимы на 1% уровне. Это означает, что *чем больше текстовой информации доступно*

для анализа, тем более точным становится предлагаемый индекс этичности деятельности банков, если принять количество арбитражных дел за прокси-показатель. Например, если мы ограничим выборку банками, для которых доступно не менее 4 тысяч единиц текстовой информации, корреляция между количеством арбитражных дел и индексом этичности достигнет в среднем –0.5 (около 100 таких наблюдений в выборке), что является весьма сильной взаимосвязью.

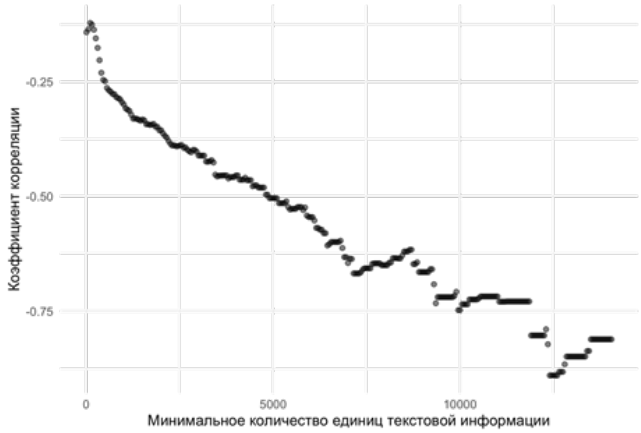

Рис 2. Коэффициент корреляции между количеством арбитражных дел и индекса этичности по словарю BERT, рассчитанного способом № 1, в зависимости от фильтра на количество единиц текстовой информации.

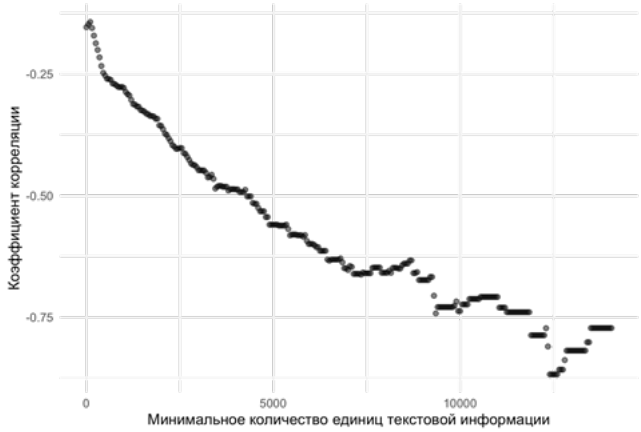

Рис 3. Коэффициент корреляции между количеством арбитражных дел и индекса этичности по словарю BERT, рассчитанного способом № 2, в зависимости от фильтра на количество единиц текстовой информации.

4. ЗАКЛЮЧЕНИЕ

Основным результатом данной работы является алгоритм расчета Индекса этичности российских банков на основе обученной нейросетевой модели BERT. Алгоритм позволяет проанализировать массив отзывов, которые оставили потребители о своем взаимодействии с банком за указанный период (месяц, квартал, год) и рассчитать индекс этичности в промежутке [–1; 1], характеризующий степень этичности поведения банка по отношению к потребителям. Как видно из расчетов, данные индексы у разных банков различны, изменяются во времени (что отражает изменение поведения банка по отношению к потребителям) и коррелируют с количеством арбитражных исков к данном банку в соответствующих периодах. Визуализированные индексы для различных банков с 2005 до 2023 гг. можно посмотреть на сайте index-ai.ethics.hse.ru. Данный эмпирический результат интересен в свете аналогичных зарубежных исследований измерения этичности банков [24, 25].

Разработанный алгоритм измерения этичности будет полезен для дальнейших разработок индексов и рейтингов добросовестности поведения на базе искусственного интеллекта. Авторы предполагают использовать его в качестве компонента при разработке интегрального показателя (индекса, рейтинга) этичности компании, который можно будет применять к любым отраслям и который будет включать в себя не только отзывы потребителей, но также и другие источники информации (веб-сайты, новости, отчеты, результаты проверок регуляторов и т. д.).

БЛАГОДАРНОСТИ

Авторы благодарят Романа Соломатина за помощь в осуществлении расчетов и Алексея Масютина за помощь при подготовке статьи к публикации.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследования выполнено в рамках программы фундаментальных исследований НИУ ВШЭ.

СПИСОК ЛИТЕРАТУРЫ

1. *Гришанкова С. Д.* Рейтинги ESG. ESG-трансформация как вектор устойчивого развития: В трех томах. Том 2. Под общ. ред. К. Е. Турбиной и И. Ю. Юргенса. М.: Издательство “Аспект Пресс”, 2022.
2. *La Torre M., Cardi M., Leo S., & Schettini Gherardini J.* ESG Ratings, Scores, and Opinions: The State of the Art in Literature. Contemporary Issues in Sustainable Finance, 2023. С. 61–102.
3. *Игнатова О. В.* ESG-рейтинги российского бизнеса. РИСК: Ресурсы, Информация, Снабжение, Конкуренция. 2022. № 1.
4. *Петров В. О., Стариков И. В., Фурщик М. А.* Особенности отечественных ESG-рейтингов // Журнал Бюджет. 2022. № 4.
5. *Казаков А., Денисова С., Барсола И., Калугина Е., Молчанова И., Егоров И., Костерина А.*

et al. ESGify: автоматизированная классификация экологических, социальных и управленческих рисков // Доклады Российской академии наук. 2023. Т. 514. № 2.

6. *Brown T. J., & Dacin P. A.* The company and the product: Corporate associations and consumer product responses. Journal of marketing, 61(1), 1997.
7. *Folkes V. S., & Kamins M. A.* Effects of information about firms' ethical and unethical actions on consumers' attitudes. Journal of consumer psychology, 8(3), 1999.
8. *Sen S., & Bhattacharya C. B.* Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. Journal of marketing Research, 38(2), 2001.
9. *Brunk K.H.* Exploring origins of ethical company/brand perceptions—A consumer perspective of corporate ethics. Journal of business research, 63(3), 2010.
10. *Khan I., & Fatma M.* Understanding the Influence of CPE on Brand Image and Brand Commitment: The Mediating Role of Brand Identification. Sustainability, 15(3), 2023.
11. *Fombrun C. J., Gardberg N. A., & Sever J. M.* The Reputation Quotient SM: A multi-stakeholder measure of corporate reputation. Journal of brand management, 7, 2000.
12. *Yang C. C., Tang X., Wong Y. C., & Wei C. P.* Understanding online consumer review opinions with sentiment analysis using machine learning. Pacific Asia Journal of the Association for Information Systems, 2(3), 2010.
13. *Sokolov A., Mostovoy J., Ding J., & Seco L.* ESG Index from Tweets and News Articles. Proceedings of the 2020 Workshop on NLP Business Applications. 2020.
14. *Briscoe-Tran H.* Do employees have useful information about firms' ESG practices? Fisher College of Business Working Paper, 2023.
15. *Jain P. K., Pamula R., & Srivastava G.* A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer science review, 41(1), 2021.
16. *Wankhade M., Rao A. C. S., & Kulkarni C.* A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 2022.
17. *Rantanen A., Salminen J., Ginter F., & Jansen B. J.* Classifying online corporate reputation with machine learning: a study in the banking domain. Internet Research, 30(1), 2020.
18. *Agrawal S. R., & Mittal D.* Optimizing customer engagement content strategy in retail and E-tail: Available on online product review videos. Journal of Retailing and Consumer Services, 67, 2022.
19. *de Kok S., Punt L., van den Puttelaar R., Ranta K., Schouten K., & Frasincar F.* Review-Aggregated Aspect-Based Sentiment Analysis with Ontology Features. Progress in Artificial Intelligence, 7(4), 2018. https://doi.org/10.1007/s13748-018-0163-7
20. *Sanei A., Cheng J., Adams B.* The Impacts of Sentiments and Tones in Community-Generated Issue Discussions. IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering, CHASE, 2021. https://doi.org/10.1109/CHASE52884.2021.00009
21. *Mirtalaie M. A., Hussain O. K.* Sentiment Aggregation of Targeted Features by Capturing Their Dependencies: Making Sense from Customer Reviews. International Journal of Information Management, 53, 2020. https://doi.org/10.1016/j.ijinfomgt.2020.102097.2020
22. *Basiri M. E., Kabiri A., Abdar M., Mashwani W. K., Yen N. Y., Hung J. C.* The Effect of Aggregation Methods on Sentiment Classification in Persian Reviews. Enterprise Information Systems, 14(9–10), 2020. https://doi.org/10.1080/17517575.2019.1669829
23. *Graham J., Haidt J., & Nosek B. A.* Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology, 96(5), 2009.
24. *Halamka R., & Teplý P.* The effect of ethics on banks' financial performance. Prague Economic Papers, 26(3), 2017.
25. *Alotaibi K. O., Mubarak I. A. S., & Alhammadi S.* Perceptions of Concerned Parties about Governance and Business Ethics in Kuwaiti Banks, June, 2020.

AI-BASED ETHICS INDEX OF RUSSIAN BANKS

M. A. Storchevoy[a], P. A. Parshakov[b,c], S. N. Paklina[b], A. V. Buzmakov[b], V. V. Krakovich[a]

[a] *St. Petersburg School of Economics and Management, National Research University Higher School of Economics in St. Petersburg, St. Petersburg, Russia*

[b] *International Laboratory of Intangible Asset Economics, National Research University Higher School of Economics in Perm, Perm, Russia*

[c] *Moscow School of Management SKOLKOVO, Moscow, Russia*

Presented by the Academician of the RAS A. I. Avetisyan

Measuring a company's ethics is an important element in the mechanism of regulating the behavior of market participants, as it allows consumers and regulators to make better decisions, which has a disciplining effect on companies. We tested various methods of machine analysis of consumer feedback from Russian banks and developed an Ethics Index that allows us to calculate a quantitative assessment of the ethics of three hundred Russian banks based on consumer feedback for different time periods from 2005 to 2022. We used a bag-of-words method based on the Moral Foundations Dictionary (MFD) and BERT model training based on a 3,000- and 10,000-sentence sample marked up by experts. The resulting index was validated based on the number of arbitration cases from 2005 to 2022 (more ethical companies are involved in fewer arbitration cases as a defendant), with only the BERT model validated and the MFD-based model not. The ethicality index will be useful as an alternative metric to the popular ESG ratings both for theoretical research on company behavior and for practical tasks of managing company reputation and forming policies to regulate the behavior of market participants.

Keywords: index, ethics, artificial intelligence, NLP, BERT.