Г.А. БОРИСЕНКО
аспирант экономического факультета МГУ имени М.В. Ломоносова

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ АКЦИЙ НА ОСНОВЕ НОВОСТНЫХ ДАННЫХ[1]

Данная работа посвящена прогнозированию движения стоимости акций крупных российских компаний, представленных в индексе Московской биржи, на основе новостных данных. В качестве моделей для прогноза используются нейронные сети трансформеры, а также и классические методы машинного обучения. В качестве новостных данных используются крупные российские новостные источники и Telegram-каналы по экономике и финансам. Задача решается в двух постановках: классификация на 2 класса (цена акции окажется выше/ниже текущей) и классификация на 3 класса (цена акции окажется выше/примерно на том же уровне/ниже текущей). В результате исследования было выявлено, что классические методы машинного обучения справляются лучше с данной задачей в общем случае, но нейронные сети также показывают хорошее качество для крупных компаний.

Ключевые слова: *стоимость акций, новости, нейросети.*

УДК: 336.763.21, 004.032.26

EDN: KJVYKJ

DOI: 10.52180/2073-6487_2024_5_211_232

Введение

Согласно гипотезе эффективного рынка (подробнее см.: [4]), вся существенная информация немедленно и в полной мере отражается на рыночной стоимости ценной бумаги. Это значит, что никакие новости или события не могут обеспечить кому-либо из акторов преимущество в прогнозировании будущей цены актива, поскольку все значимые факторы уже учтены рынком.

[1] Выражаю искреннюю благодарность Гурову Илье Николаевичу, д.э.н. (CFA, экономический факультет МГУ имени М.В. Ломоносова), Демиденко Татьяне Ивановне, к.э.н. (РГЭУ (РИНХ)), Мирзояну Ашоту Гамлетовичу (экономический факультет МГУ имени М.В. Ломоносова) за ценные советы при написании данной работы на стадии подготовки препринта.

Но, даже при том что сильная форма эффективности рынка[2] недостижима, тем не менее существуют возможности для того, чтобы на основе анализа новостных данных предсказывать изменения будущей стоимости актива.

Капитализация публичных компаний зависит от событий, которые происходят как в сфере самого бизнеса, так и за ее пределами. Соответственно, имея доступ к этой информации, при недостижимости сильной формы эффективности рынка, можно делать предсказания о стоимости акций компании в будущем.

Наиболее доступным источником информации о компаниях являются публикации крупных новостных изданий, однако информация там публикуется с задержкой. Чтобы преодолеть этот недостаток, можно воспользоваться новостями из социальных сетей, в частности из Telegram-каналов, где публикация новостей происходит максимально быстро.

Объектом данного исследования являются новости о публичных компаниях, а предметом – взаимосвязь новостей о таких компаниях с движением цен их акций.

На данный момент лучше всего справляются с задачей обработки естественных языков современные архитектуры нейронных сетей, так как они способны агрегировать в себе огромное число взаимосвязей в имеющейся информации; зачастую они способны даже превосходить результат работы человека. Таким образом, с помощью глубокого семантического анализа полученных новостей можно попробовать уточнить предсказание о движении стоимости акций компаний, к которым новость относится. Более того, в случае статистически значимых результатов можно будет говорить о возможностях искусственного интеллекта использовать информацию из новостей для более успешной торговли на неэффективных, с точки зрения разных по силе форм эффективности, рынках.

Цель данной работы состоит в построении моделей машинного обучения, способных предсказывать направление движения стоимости акций крупных компаний на Московской бирже. В качестве моделей будут использоваться как классические методы машинного обучения, так и нейросетевые подходы, с целью сравнения качества их прогнозов. Более того, будет проведен сравнительный анализ предсказательной силы моделей для разных входных данных: традиционных новостей из крупных новостных ресурсов и новостей из Telegram-каналов[3].

---

[2] Сильная форма эффективности рынка – если стоимость рыночного актива полностью отражает всю информацию – прошлую, публичную и внутреннюю (инсайдерскую информацию, которая известна узкому кругу лиц в силу служебного положения, или иных обстоятельств). Ru/Wikipedia.org/wiki.

[3] Основные понятия и термины, используемые в статье, приведены в Приложении.

1. Обзор литературы

Прогнозирование движения акций интересовало людей с момента создания фондовых бирж. Однако строить краткосрочные прогнозы на основе новостных данных стало возможным относительно недавно, что объясняется большой размерностью данных, с одной стороны, а с другой – отсутствием до недавнего времени необходимых мощностей.

Одной из первых работ, где использовались новости для предсказания движения стоимости акций, стала статья Фама [4]. В ней автор пользовался моделью Rainbow для текстовой классификации и моделью Naive Bayes. Со временем вычислительные мощности росли, а алгоритмы машинного обучения развивались. При работе с текстами стали использовать продвинутые методы векторизации TF-IDF и Word2vec, а также более сложные методы машинного обучения, такие как SVM [3], «случайный лес» [13] и бустинг [10].

Нейросетевые подходы решения задачи также активно развивались. В исследованиях находили применение сверточные [14] и рекуррентные нейронные сети [6], а также наиболее популярный вид нейросетей для работы с текстовыми данными на сегодняшний день – трансформеры [7].

В источниках данных наблюдается меньшая вариативность. Как правило, авторы используют новости крупных изданий [6] или новости в Twitter [8]; иногда встречаются более нестандартные источники, например комментарии пользователей Twitter по поводу изменения цен на активы [10].

К сожалению, русскоязычных статей по прогнозированию стоимости акций, торгующихся на Московской бирже, с помощью нейросетей на основе новостей в открытом доступе при подготовке статьи не было найдено. На данный момент в опубликованных русскоязычных статьях используются только авторегрессионные модели (см., например: [1]). Если же говорить об анализе именно новостей, то существует научная работа о влиянии политических новостей на акции различных секторов [12]. Но ее главная цель – оценить степень воздействия новостей на акции секторов, цель же данного исследования — предсказывать направление движения акций.

2. Используемая информация и данные[4]

2.1. Классические новости

В качестве источника традиционных новостных данных были выбраны крупные новостные издания России (РИА «Новости», «Ком-

[4] Весь код и ссылки на все данные можно найти по ссылке на репозиторий на GitHub https://github.com/BorisenkoGeorgy/Disser_news_stocks (дата обращения: 24.10.2024).

мерсантъ», Лента.ру, «Ведомости»), данные из которых были получены с помощью самостоятельно реализованных автором на языке Python парсинговых программ. Всего было получено 716 740 уникальных текстов, начиная с 1 января 2010 г., по 20 ноября 2022 г.

2.2. Новости из Telegram

В Telegram существует множество различных каналов, основная идея которых — освещение актуальных событий о публичных компаниях. Их существенно больше, чем крупных новостных источников, поэтому необходима процедура отбора каналов, чтобы максимально охватить информационное поле публичных компаний. Стартовый набор каналов был выбран автором на основе экспертного мнения. Все эти каналы крупные и публикуют новости часто и в коротком формате. Далее, с помощью самостоятельно написанной автором парсинговой программы, были получены новости из этих каналов[5].

Telegram-каналы часто пересылают новости друг другу, и эта информация также может быть получена при парсинге. Таким образом, из полученных данных можно извлечь информацию о каналах, на которые ссылается выбранный канал. В результате была составлена матрица популярности каналов. Популярность канала определялась по тому, сколько каналов и как часто на него ссылаются. По полученным значениям данной матрицы можно выделить наиболее популярные каналы, которые не содержатся в исходной выборке, и получить все новости из них.

Данные действия совершались итеративно, пока к списку каналов не перестали добавляться новые каналы. Таким образом удалось охватить максимально широкое информационное поле, ограничившись условно небольшим числом Telegram-каналов. В результате было получено 1 043 208 уникальных текстов за период, начиная от создания каждого канала, до 15 января 2023 г.

2.3. Предварительная обработка текстов

Для работы с текстовыми данными их необходимо преобразовать к виду, понятному для модели, то есть к числовому.

Для классического ML все текстовые данные были очищены от пунктуации и различных служебных символов, таких как $\backslash n$ и $\backslash t$, затем очищены от стоп-слов (различные предлоги, местоимения, частицы). Также при обработке текстов Telegram каналов были удалены смайлики. После этого все тексты были приведены к нижнему регистру и лемматизированы с помощью пакета pymorphy2 для Python.

[5] Парсеры – программы, которые помогают собирать и систематизировать данные. Информацию можно брать как со своего веб-ресурса, так и с других сайтов.

Нейросетевые модели используют уже предварительно обученные эмбеддинги, которые могут работать с предложениями практически в сыром виде. Были удалены лишь специальные символы \n, \r и смайлики.

2.4. Выборка ценных бумаг для исследования

Для анализа были выбраны ценные бумаги, входящие в Индекс МосБиржи на ноябрь 2021 г. Из выборки были исключены такие компании (далее компании будут называться их биржевыми тикерами), как OZON, VKCO, POGR, HHRU, POGR, так как на тот момент они недавно появились на бирже, поэтому по ним было собрано мало информации.

Также из списка ценных бумаг были исключены привилегированные акции SBERP, TATNP, SNGSP, так как в моей выборке присутствуют обыкновенные акции соответствующих эмитентов. Итоговый список компаний, информация о которых анализировалась в данном исследовании, представлен в табл. 1.

2.5. Получение данных о стоимости ценных бумаг и создание целевой переменной

Данные о стоимости ценных бумаг были также получены с помощью самостоятельно реализованной парсинговой программы на Python с сайта финансового портала Финам.ру[6]. По каждой компании были получены данные цены открытия, минимальной, максимальной, закрытия и объем торгов в денежном эквиваленте по интервалам 5, 10, 15, 30 минут, 1 час и 1 день.

При классификации на 2 класса целевая переменная равнялась 0, если после выхода новости цена актива снизилась, и 1, если выросла.

При классификации на 3 класса задается предпосылка, что не каждая новость оказывает значимое влияние на движение акций компании. За интервал в n предыдущих минут от момента выхода новости вычислялась средневзвешенная цена и ее стандартное отклонение. Если в следующие n минут средневзвешенная цена отклонялась от средневзвешенной за последние n минут более, чем на полтора стандартных отклонения, то такая новость относилась к классу +1 или –1 в зависимости от направления отклонения (позитивный/негативный класс). Если же средневзвешенная цена остается в рамках полутора стандартных отклонений, то такой новости присваивается класс 0, то есть нейтральный.

[6] Официальный сайт АО «Инвестиционный холдинг Финам» https://www.finam.ru/ (дата обращения: 10.01.2023).

Таблица 1

Список компаний, участвовавших в исследовании

Тикер	Компания	Сектор
AFKS	ПАО АФК «Система»	Различные сектора
AFLT	ПАО «Аэрофлот»	Транспортный
ALRS	ПАО «Алроса»	Несырьевые полезные ископаемые
CBOM	ПАО «Московский кредитный банк»	Ритейл
CHMF	ПАО «Северсталь»	Черная металлургия
DSKY	ПАО «Детский мир»	Ритейл
FEES	ПАО «Россети»	Электроэнергетика
GAZP	ПАО «Газпром»	Нефтегазовый
GMKN	ПАО «Норильский никель»	Цветная металлургия
HYDR	ПАО «РусГидро»	Электроэнергетика
IRAO	ПАО «Интер РАО»	Электроэнергетика
LKOH	ПАО «Лукойл»	Нефтегазовый
LSRG	ПАО «ЛСР»	Девелопмент
MAGN	ПАО «ММК»	Черная металлургия
MOEX	ПАО «Московская биржа»	Финансовый
MTSS	ПАО «МТС»	Телекоммуникации
NLMK	ПАО «НЛМК»	Черная металлургия
NVTK	ПАО «Новатэк»	Нефтегазовый
PHOR	ПАО «Фосагро»	Химическая и нефтехимическая
PIKK	ПАО «ПИК»	Девелопмент
PLZL	ПАО «Полюс»	Золотодобыча
ROSN	ПАО «Роснефть»	Нефтегаз
RTKM	ПАО «Ростелеком»	Телекоммуникации
RUAL	ПАО «Русал»	Цветная металлургия
SBER	ПАО «Сбербанк»	Финансовый
SNGS	ПАО «Сургутнефтегаз»	Нефтегазовый
TATN	ПАО «Татнефть»	Нефтегазовый
TCSG	ПАО «ТКС Групп»	Финансовый
TRNFP	ПАО «Транснефть»	Нефтегазовый
VTBR	ПАО «ВТБ»	Финансовый
YNDX	ПАО «Yandex»	Телекоммуникации

Источник: составлено автором.

2.6. Разметка данных с помощью регулярных выражений

Некоторую часть информации, которая заложена в тексте, можно извлечь путем поиска ключевых слов. В языках программирования (в том числе и на Python) это реализовано через функционал Регулярных выражений (regex). То есть, если задать набор ключевых слов, можно проверить, какие из них входят в текст, и разметить все тексты соответствующим образом.

С помощью регулярных выражений были отобраны новости, которые относятся к конкретным компаниям. Например, если в тексте новости встречается «Сбербанк», то эта новость относится к ПАО «Сбербанк». Все регулярные выражения были составлены самостоятельно, их список можно найти на GitHub. Регулярные выражения состояли не только из названий компаний, но и из их сокращенных названий, биржевых тикеров и названий дочерних предприятий.

3. Построение Бейзлайн-моделей

В качестве бейзлайн-моделей были выбраны модели «Случайного леса» [2] и Градиентного бустинга над деревьями [9] из-за своей относительной легкости работы с большими данными. Также они показывали себя хорошо в прошлых исследованиях из Обзора литературы и превосходили по качеству модели Наивного Байеса и Логистической регрессии. SVM, хоть и показывала сравнимое качество в исследованиях, является вычислительно дорогой моделью. Так как было получено 1,7 млн текстов, то работа с SVM будет затруднительна.

Для бейзлайна текстовые данные приводились к числовым с помощью метода TF-IDF. Параметры моделей и TF-IDF можно найти в приложенной выше ссылке на Github. Подбор параметров для моделей классического машинного обучения проводился с помощью перебора по сетке на основе кросс-валидации.

Обучение и кросс-валидация проводились на всех доступных данных до 1 июня 2021 г. не включительно. Под тестовую часть были выделены данные с 1 июня 2021 г. по 1 января 2022 г. не включительно. Для каждого временного интервала, каждой компании и каждого источника была построена отдельная модель классического машинного обучения.

Чтобы показать наличие информации в текстах предсказания моделей сравнивались с простым бенчмарком — предсказание случайным классом с вероятностями, полученными на обучающей части выборки.

3.1. Результаты для Телеграмма

В данном разделе будут рассмотрены полученные результаты для моделей «случайного леса» и градиентного бустинга над деревьями для

новостей из Телеграмма. Результаты в статье демонстрируются только для компаний, которые моделям удалось предсказать качественно, то есть на хотя бы 5 из 6 временных интервалов модели показывают качество предсказания, статистически отличные от случайного. Полные результаты можно найти в репозитории GitHub.

Значения в таблицах представляют из себя следующее. Было получено среднее качество предсказания случайным и его стандартное отклонение для метрики Accuracy. В таблицах представлена разность Accuracy для модели и для суммы среднего и стандартного отклонения. Назовем это значение минимальным ожидаемым эффектом от применения модели.

Рассмотрим пример. Пусть имеем значение 2,5 в некоторой ячейке. Если случайное предсказание показывает качество в среднем 52% с стандартным отклонением в 1,5%, то наша модель предсказывает с точностью 52 + 1,5 + 2,5 = 56%.

Для алгоритма «случайный лес» были получены следующие результаты (см. табл. 2).

Таблица 2

Результаты по минимальному ожидаемому эффекту, полученные для задачи классификации новостей из Telegram на 3 класса алгоритмом «случайного леса»

Тикер\Интервал	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
AFKS	1,44	4,62	1,32	0,97	0,16	–0,07
AFLT	–0,78	1,94	2,43	1,44	1,47	2,81
ALRS	0,40	3,43	1,87	–0,38	2,97	2,76
CHMF	–0,63	7,14	7,67	5,11	7,57	0,17
DSKY	6,49	4,24	9,58	14,61	9,35	–0,27
GAZP	1,92	1,02	1,96	2,38	2,44	–0,46
HYDR	4,52	4,14	6,03	–1,56	1,60	4,14
MAGN	–2,22	2,53	4,70	2,66	5,22	–0,01
MOEX	1,15	1,39	1,74	1,74	2,46	1,29
MTSS	4,41	2,91	5,10	6,24	1,48	–1,19
NVTK	3,00	–0,67	2,36	2,29	6,05	3,54
PHOR	4,83	1,14	1,29	2,63	1,47	–0,16
PIKK	–1,88	4,85	5,43	2,83	2,69	0,07
ROSN	2,84	4,43	4,16	0,54	0,70	1,09
SNGS	6,96	6,36	4,10	1,36	9,93	8,47

Источник: составлено автором.

Для алгоритма бустинга для качественно предсказанных компаний были получены следующие результаты (см. табл. 3).

Таблица 3

Результаты по минимальному ожидаемому эффекту для задачи классификации новостей из Telegram на 3 класса алгоритмом бустинга над деревьями

Тикер\Интервал	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
DSKY	3,44	3,84	8,41	9,97	6,49	0,44
GAZP	0,69	1,45	0,92	2,66	2,13	0,21
LKOH	1,56	1,07	1,51	2,24	0,11	1,01
NVTK	0,14	0,43	1,81	3,90	0,83	1,39
ROSN	–0,25	1,71	1,09	2,37	1,19	1,20
SNGS	0,12	1,17	1,20	2,82	3,31	7,30

Источник: составлено автором.

3.2. Результаты для участвующих в исследовании классических источников новостей

В данном разделе будут рассмотрены полученные результаты для моделей «случайного леса» и «градиентного бустинга над деревьями» для классических новостных источников.

Рассмотрим результаты для «случайного леса» (см. табл. 4).

Таблица 4

Результаты по минимальному ожидаемому эффекту для задачи классификации новостей из классических источников на 3 класса алгоритмом «случайного леса»

Тикер\Интервал	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
GAZP	2,59	6,07	6,29	6,38	–6,34	–0,04
MTSS	5,63	0,53	13,68	1,16	6,05	–13,90
SBER	1,13	2,21	2,34	3,10	3,24	6,72
TCSG	1,25	–0,18	0,86	9,25	3,58	1,47
VTBR	2,63	9,73	10,50	10,37	10,02	–10,70
YNDX	4,10	7,26	8,00	4,30	2,11	–1,87

Источник: составлено автором.

Теперь обратимся к результатам работы алгоритма градиентного бустинга над деревьями (см. табл. 5).

Таблица 5

Результаты по минимальному ожидаемому эффекту для задачи классификации новостей из классических источников на 3 класса алгоритмом бустинга над деревьями

Тикер\Интервал	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
DSKY	18,48	–18,10	13,18	6,70	–9,14	–23,21
GAZP	3,06	1,39	4,85	3,39	6,80	3,99
TCSG	–1,76	4,47	7,73	4,79	0,61	0,17
VTBR	–1,60	5,80	8,41	9,73	6,19	–8,49
YNDX	–0,32	6,43	8,03	7,54	0,23	–5,34

Источник: составлено автором.

3.3. Итоги построения бейзлайн-моделей

Как видно из приведенных выше таблиц, прогнозы для задачи классификации на 3 класса случайный лес показывает больше значимых эффектов по сравнению с бустингом. Также больше значимых эффектов наблюдается для новостей, полученных из Telegram, но все-таки для полноты картины необходимо свести все результаты в таблицу, чтобы сравнить между собой различные модели и источники (см. табл. 6).

Таблица 6

Сводная таблица по средним минимальным эффектам, %

Задача	Источник	Модель	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
3 класса	Telegram	Случайный лес	0,80	1,65	2,02	1,78	2,36	0,45
3 класса	Telegram	Бустинг	–0,34	–0,07	0,38	0,81	0,50	–0,68
3 класса	Новости	Случайный лес	–4,42	–3,11	–1,91	–1,79	–1,57	–6,30
3 класса	Новости	Бустинг	–5,41	–5,32	–2,92	–2,15	–2,68	–5,76

Источник: составлено автором.

4. Нейросетевой подход

В качестве нейросетей модели были опробованы несколько вариантов трансформеров [11] с сайта Hugging Face[7]. Это сайт, на котором в отрытом доступе можно найти уже обученные веса для интересующих моделей sbert_large_nlu_ru[8], ruRoberta-large[9] и distilrubert-base-cased-conversational[10]. Но в результате была выбрана только одна модель distilrubert-base-cased-conversational, как оптимальный вариант по качеству/скорости обучения при ограниченных ресурсах.

Трансформеры были выбраны из-за своей огромной популярности для решения задач NLP. Популярность обусловлена высоким качеством решения задач с помощью архитектуры трансформера, а также за счет вычислительной эффективности, которая позволяет быстро проводить эксперименты. Более того, за счет популярности в открытом доступе существует множество уже обученных трансформеров под самые разные задачи. Остается выбрать наиболее подходящую модель и дообучить ее.

Для получения ответов для нашей задачи необходимо было достроить классификатор поверх трансформера (то есть на выходе из блока трансформера, который отвечает за обработку языка добавлялся полносвязный блок). Это производилось следующим образом. Все выходы трансформера усреднялись и конкатенировались с выходами трансформера для токена <CLS>[11].

Для обучения нейросети кросс-валидацию использовать затруднительно, так как это дорого (из-за большого количества параметров модели и большого объема данных), поэтому придется разбить данные на обучение, валидацию и тест. То есть формально обучение по сравнению с классическим ML будет происходить на немного разных выборках, но другого варианта из-за ограничений по ресурсам нет. Итоговые данные были разбиты по интервалам: от начала сбора данных до 1 января 2021 г. не включительно для обучающей выборки; от

---

[7] Официальный сайт Hugging Face. https://huggingface.co/ (дата обращения: 17.01.2023).

[8] Карточка модели Sbert Large на Hugging Face https://huggingface.co/ai-forever/sbert_large_nlu_ru (дата обращения: 17.01.2023).

[9] Карточка модели ruRoberta Large на Hugging Face https://huggingface.co/ai-forever/ruRoberta-large (дата обращения: 17.01.2023).

[10] Карточка модели Distilrubert-base-cased-conversational на Hugging Face https://huggingface.co/DeepPavlov/distilrubert-base-cased-conversational (дата обращения: 17.01.2023).

[11] Это описано в статье, посвященной решению задачи соревнования с помощью трансформера sbert – Обучение модели естественного языка с BERT и Tensorflow https://habr.com/ru/company/sberdevices/blog/527576/ (дата обращения: 17.01.2023).

1 января 2021 г. до 1 июня 2021 г. не включительно для валидационной выборки; и от 1 июня 2021 г. до 1 января 2022 г. не включительно для тестовой выборки. Получаем, что тестовые выборки для нейросетевого подхода и классического ML совпадают.

Нейросетевая модель может иметь множество выходов и предсказывать сразу несколько целевых переменных. Однако в таком подходе есть проблема: нейросеть будет оптимизировать среднее значение функций потерь для каждого выхода сети, а не каждый выход в отдельности. Вероятно, что из-за этого результаты сети будут хуже, так как на примере результатов классического ML видно, что модели не всегда хорошо справляются с задачей прогнозирования на всех интервалах. Для правильного сравнения нейросетевого подхода с классическим машинным обучением необходимо на каждую целевую переменную для каждой компании обучать свою модель, что займет очень много времени. Поэтому было принято решение сконцентрироваться на меньшем числе компаний, а именно на тех, для которых модели классического ML получили качество предсказания, статистически отличное от случайного. Не на всех временны*х* интервалах для этих компаний модели классического ML показывают себя лучше случайного предсказания, поэтому также будет возможность сравнить результаты нейросетевого подхода с классическим ML там, где второй не справляется.

4.1. Результаты для Telegram

Рассмотрим результаты нейросетевого подхода для классификации на 3 класса для Telegram (см. табл. 7).

Рассмотрим разность минимальных ожидаемых эффектов для нейросетового подхода и «случайного леса» (см. табл. 8).

Как видим, нейросеть справляется практически везде хуже по сравнению со «случайным лесом», за исключением прогнозирования акций Газпрома.

4.2. Результаты для классических новостей

Рассмотрим результаты задачи классификации на 3 класса для классических новостных источников с помощью нейросети (см. табл. 9).

Рассмотрим результаты в сравнении с эффектами для «случайного леса» (см. табл. 10).

В данном случае нейросеть очень сильно проигрывает случайному лесу. Из значимых эффектов явно выделяется VTBR на интервале в 1 день – там нейросеть показывает сильное превосходство в качестве.

4.3. Итоговые результаты для нейросетевого подхода

Как видно из попарного сравнения таблиц метрик для нейросетевого подхода и «случайного леса», по результатам исследования

Таблица 7

Результаты по минимальному ожидаемому эффекту для задачи классификации новостей из Telegram на 3 класса нейросетью

Тикер	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
AFKS	–3,59	1,96	–5,60	–2,38	–2,12	0,09
AFLT	2,06	2,28	1,56	0,89	–4,61	4,86
ALRS	–3,16	1,35	1,17	–1,98	–3,79	–0,09
CHMF	–2,18	–8,67	–7,05	–4,76	–7,31	–2,99
DSKY	–9,19	–19,53	–17,59	–17,72	–13,51	–8,28
GAZP	4,06	3,79	2,97	3,30	1,93	2,77
HYDR	4,37	4,93	–0,70	–0,45	–2,50	10,21
LSRG	–11,36	–10,61	–10,04	–7,43	–9,05	–6,43
MAGN	3,15	–0,67	2,87	6,46	–0,23	–0,80
MOEX	–2,91	–4,84	–5,44	–2,58	–4,00	–4,62
MTSS	–3,47	–0,82	0,15	–1,80	2,22	–8,15
NVTK	5,73	2,18	0,86	0,91	0,75	3,83
PHOR	–14,94	–10,11	–9,73	–12,94	–11,20	–2,26
PIKK	1,84	0,82	–2,40	5,01	0,88	7,28
ROSN	2,46	1,16	3,57	–0,77	0,50	2,62
SNGS	–0,72	4,76	1,91	–2,57	4,30	4,85

Источник: составлено автором.

Таблица 8

Сравнение по Accuracy нейросетевого подхода и алгоритма случайного леса для задачи классификации на 3 класса по новостям из Telegram

Тикер	**5 мин.**	**10 мин.**	**15 мин.**	**30 мин.**	**1 час**	**1 день**
AFKS	–5,06	–2,67	–6,87	–3,27	–2,23	0,22
AFLT	2,83	0,37	–0,97	–0,55	–6,17	1,98
ALRS	–3,66	–1,97	–0,67	–1,63	–6,72	–2,91
CHMF	–1,66	–15,71	–14,77	–9,91	–14,78	–3,04
DSKY	–15,51	–23,54	–27,28	–32,44	–22,77	–8,08
GAZP	2,12	2,79	1	0,93	–0,52	3,22
HYDR	–0,13	0,66	–6,73	1,13	–4,13	6,04
LSRG	–16,49	–13,83	–10,52	–11,06	–9,32	–4,95
MAGN	5,32	–3,04	–1,66	3,6	–5,47	–0,75
MOEX	–4,07	–6,31	–7,24	–4,32	–6,43	–5,88
MTSS	–7,81	–3,78	–4,91	–8,08	0,85	–7,17
NVTK	2,73	2,75	–1,27	–1,4	–5,3	0,33
PHOR	–19,87	–11,23	–11,16	–15,84	–12,65	–2,31
PIKK	3,7	–3,81	–7,81	2,19	–1,68	7,24
ROSN	–0,28	–3,33	–0,55	–1,35	–0,27	1,59
SNGS	–7,58	–1,9	–2,27	–3,77	–5,58	–3,45

Источник: составлено автором.

Таблица 9

Результаты по минимальному ожидаемому эффекту для задачи классификации новостей из классических источников на 3 класса нейросетью

Тикер	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
GAZP	2,06	2,38	3,77	4,58	1,40	–0,94
MTSS	–5,42	–0,51	–0,65	–1,29	–6,83	–10,19
RTKM	–19,40	–8,49	–8,13	–6,77	–7,01	–5,62
SBER	0,59	–0,61	–0,21	–4,26	–4,57	–9,07
TCSG	–17,46	–17,81	–10,06	–23,44	–6,89	–5,00
VTBR	4,47	5,58	6,50	9,21	9,51	12,66
YNDX	–2,09	–6,79	–6,22	1,73	1,62	–9,23

Источник: составлено автором.

Таблица 10

Сравнение по Accuracy нейросетевого подхода и алгоритма «случайного леса» для задачи классификации на 3 класса по новостям из классических источников

Тикер	5 мин.	10 мин.	15 мин.	30 мин.	1 час	1 день
GAZP	–0,52	–3,77	–2,53	–1,78	7,73	–0,94
MTSS	–11,25	–1,1	–14,4	–2,71	–12,72	3,57
RTKM	–8,03	–15,27	–12,34	–6,9	–10,84	6,22
SBER	–0,48	–2,77	–2,63	–7,46	–7,74	–15,79
TCSG	–18,93	–18,01	–11,13	–32,91	–10,48	–6,51
VTBR	1,7	–4,12	–4,01	–1,26	–0,75	23,22
YNDX	–6,39	–13,78	–14,06	–2,39	–0,65	–7,26

Источник: составлено автором.

получаем, что нейросети на Московской бирже показывают качество в среднем хуже «случайного леса» для всех постановок задач за исключением прогнозирования движения акций Газпрома. Это может быть связано с тем, что Газпром – лидер по упоминаемости в прессе, а для обучения нейросетей необходимо большое число данных. Возможно, именно поэтому нейросеть в данном случае смогла показать качество предсказания выше классического ML.

Еще для некоторых компаний на отдельных временны́х интервалах нейросеть показывала качество лучше «случайного леса», но там сложно однозначно подтвердить превосходство нейросетевого подхода, так как на соседних временны́х интервалах нейросеть уже проигрывала случайному лесу.

Превосходство «случайного леса» над нейросетью можно объяснить несколькими причинами:

1. Кросс-валидация для «случайного леса» проводилась обычная, а не специальная для временны́х рядов, так как формально временной ряд не присутствует в независимых переменных, а исследование заключается в оценке влияния текста новости на движение акций. При дальнейшем исследовании и добавлении авторегрессионности в модель так делать уже будет нельзя. Из-за этого модели классического ML были обучены на данных, которые максимально приближены во времени к данным для теста, то есть модель имела возможность обучаться на самых свежих данных относительно тестовых. Для нейросетей кросс-валидацию использовать дорого, поэтому полученные нейросети нельзя было обучить на самых свежих данных.

2. Векторные представления слов для классического ML были получены в результате алгоритма TF-IDF, то есть они были рассчитаны на основе имеющихся данных. Таким образом, эти векторные представления максимально точно (насколько это возможно ввиду простоты метода) описывали новостную область (область финансовых текстов). Векторные представления нейросети уже были обучены на комментариях из социальных сетей, что не относится к области финансовых текстов. Дообучение нейросети частично решает эту проблему, но все-таки для достижения наивысшего качества нейросеть необходимо учить с нуля.

3. При встрече в тестовой выборке с новым словом, которого не было в обучающей выборке, алгоритм TF-IDF его просто пропустит, а нейросетевой подход все-таки приведет к численному виду, который был оптимален для задачи, на которой изначально обучалось это числовое представление, а не для задачи классификации финансовых текстов.

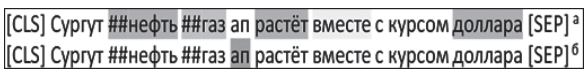
4.4. Интерпретация предсказаний нейросети

Несмотря на то, что качество для нейросети в среднем хуже, чем для «случайного леса», у нейросетевого подхода есть преимущество – благодаря механизму внимания, заложенного в идею архитектуры трансформера, можно визуализировать, на что ориентируется модель при принятии решений о классификации.

Рассмотрим примеры новостей Telegram из тестовой выборки на задаче трех классов (так как для нее было получено наибольшее

число отличных от случайных предсказаний), для которых нейросеть успешно определила метку класса, и попробуем их проинтерпретировать. (Знаки препинания специально удалены из новостей в тексте работы, так как используемая нейросеть не умеет с ними работать, а новости приходят ей именно в таком виде).

Рассмотрим новость о Сургутнефтегазе от 30.12.2021 из Telegram-канала finascop: «Сургутнефтегаз ап растет вместе с курсом доллара» на рис. 1.

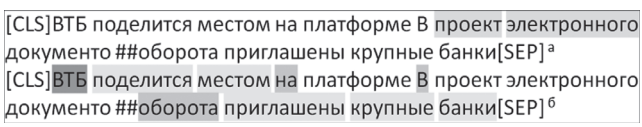


Источник: составлено автором.

Рис. 1. Визуализация предсказания для новости о Сургутнефтегазе.
а) Визуализация токенов, которые склоняют модель к позитивному ответу (цена растет); б) К негативному (цена падает)

Как видно из визуализации механизма внимания, нейросеть делает наибольший положительный акцент на связи нефти и газа с ростом доллара. Действительно, ПАО «Сургутнефтегаз» занимается добычей нефти и газа, которые в основном идут на экспорт. То есть доходы компании в рублях напрямую зависят от курса доллара: чем он выше, тем и выше доходы, а соответственно, и стоимость акций компании. И модель верно предсказывает, что акции компании на горизонте в один день покажут рост.

Теперь рассмотрим новости о ПАО «ВТБ». «ВТБ поделится местом на платформе В проект электронного документооборота приглашены крупные банки» (см. рис. 2).



Источник: составлено автором.

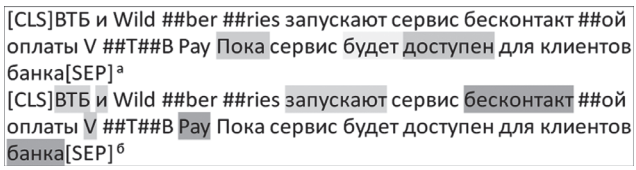
Рис. 2. Визуализация предсказания для новости о ВТБ.
а) Визуализация токенов, которые склоняют модель к позитивному ответу (цена растет); б) К негативному (цена падает)

Банк ВТБ в партнерстве с ПАО «Ростелеком» был одним из создателей программы электронного документооборота для упрощения взаимодействий между гражданами, государством и бизнесом[12]. Однако на момент выхода новости в «Комерсантъ» 08.09.2021 в развитие про-

[12] «Документооборот России под банковским контролем: зачем ВТБ наши данные?» https://regnum.ru/news/polit/3366298.html (дата обращения: 12.04.2023).

граммы были приглашены и другие банки. Соответственно, акции компании должны были отреагировать негативно, так как доля банка ВТБ в сегменте снизится. Модель верно угадывает направление движение акций и подчеркивает негативным весом, что ВТБ будет делиться доходами с крупными банками.

Также рассмотрим нейтральную новость из издания «Ведомости» от 07.12.2021. «ВТБ и Wildberries запускают сервис бесконтактной оплаты VTB PAY Пока сервис будет доступен для клиентов банка» (см. рис. 3).

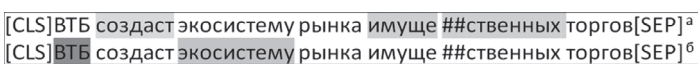


Источник: составлено автором.

Рис. 3. Визуализация предсказания для новости о ВТБ.
а) Визуализация токенов, которые склоняют модель к позитивному ответу (цена растет); б) К негативному (цена падает)

Здесь также модель выделяет некоторые положительные и отрицательные моменты новости, но относит ее все же к нейтральному классу. На тот момент Apple Pay и Samsung Pay в России еще работали, и надобности в подобном сервисе не было, поэтому, видимо, участники рынка никак на эту новость и не отреагировали. С одной стороны, она может принести дополнительную прибыль банку, но с другой — очень сложно бороться с конкурентами, которые уже широко распространены в сегменте бесконтактных платежей; поэтому новость неоднозначная, и модель верно это угадывает.

Теперь рассмотрим новость также о ПАО «ВТБ» из издания Лента.ру от 28.07.2021 «ВТБ создаст экосистему рынка имущественных торгов». Модель больше всего внимания акцентирует на том, что ВТБ займется созданием некоторого проекта по имуществу. Банку этот проект будет выгоден с точки зрения реализации заложенного имущества, которое попало в собственность банка. Тем самым банк будет быстрее избавляться от активов банкротов и сможет эффективнее управлять денежными средствами. Рынок положительно отреагировал на эту новость, и модель смогла предсказать это (см. рис. 4).

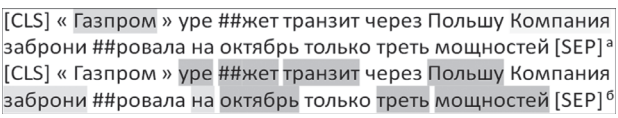


Источник: составлено автором.

Рис. 4. Визуализация предсказания для новости о Газпроме.
а) Визуализация токенов, которые склоняют модель к позитивному ответу (цена растет); б) К негативному (цена падает)

Также стоит обратить внимание, что в приведенных примерах о ПАО «ВТБ» токен ВТБ всегда выделяется негативно. Получается, что модель смогла распознать, что, с точки зрения инвестиции, банк «ВТБ» не самый лучший актив. С момента начала датасета по его конец (период 2010–2021 гг.) акции ВТБ упали примерно на 40%. А с момента IPO (апрель 2007 г.) на 70%.

Еще стоит обратиться к самой популярной и наиболее часто упоминаемой компании в новостях, к ПАО «Газпром». Рассмотрим новость: «Газпром урежет транзит через Польшу, Компания забронировала на октябрь только треть мощностей», опубликованную в «Коммерсантъ» 20. 09. 2021 (см. рис. 5).



Источник: составлено автором.

Рис. 5. Визуализация предсказания для новости о ВТБ.
а) Визуализация токенов, которые склоняют модель к позитивному ответу (цена растет); б) К негативному (цена падает)

Здесь явно виден акцент модели на «урежет транзит» и «треть мощностей». Газпром зарабатывает в основном на поставках газа за рубеж. И снижение поставок, естественно, приведет к снижению выручки и прибыли, что негативно сказывается на стоимости компании, и модель это правильно предсказывает.

Таким образом, можно сказать, что нейросеть действительно способна понимать суть полученных новостей, несмотря на то что в общем случае предсказательное качество нейросетей получилось хуже классического машинного обучения. Рассмотренные новости далеко не единственные в выборке правильно классифицированных. Данные примеры лишь показывают возможности нейросети.

Заключение

В данной работе были построены модели классического машинного обучения («случайный лес» и бустинг над деревьями) и глубокого обучения (нейросети) для прогнозирования движения цен акций публичных компаний из индекса Московской биржи и произведено сравнение моделей, обученных на разных источниках данных.

Также удалось выявить, что новости из Telegram – более надежный источник новостей, с точки зрения качества предсказания полученных моделей по сравнению с классическими новостями.

К сожалению, для нейросетевого подхода в общем случае не удалось получить качество лучше, чем для «случайного леса». Однако

в частном случае было подтверждено, что нейросеть лучше прогнозирует движение акций для GAZP. Удалось показать, что нейросеть делает свои предсказания обоснованно – путем визуализации матриц внимания и их содержательной интерпретации.

Если говорить о способности моделей машинного обучения предсказывать движение акций на неэффективных рынках по текстовым данным, то на статистически значимом уровне получено, что это возможно. То есть данная работа не просто доказывает, что российский рынок акций неэффективен в сильном смысле, а также показывает, что методы ИИ способны выявлять эту неэффективность.

Также при написании работы было реализовано пять парсинговых программ для получения новостей из выбранных источников. Код для них хранится в свободном доступе и может быть использован исследователями в дальнейшем. Более того, все собранные данные также возможно получить по запросу.

В качестве идей для дальнейших исследований возможностей машинного обучения прогнозировать движение акций на российском фондовом рынке можно попробовать добавить в модели авторегрессионную компоненту, а также добавить временные ряды биржевых товаров и валюты, усовершенствовать архитектуру нейросети для учета новостей в каком-то определенном окне в прошлом, причем можно взять большую сеть, ответственную за работу с текстом.

ПРИЛОЖЕНИЕ

Основные понятия и термины, используемые в статье

NLP (Natural Language Processing) – обработка естественного языка. Так называется область в машинном обучении, которая посвящена работе с текстами.

Бейзлайн-модель – модель, которая является условно простым решением задачи. Именно с ее результатами сравниваются дальнейшие эксперименты по улучшению решения.

Векторизация – метод представления данных.

Дообучение – процесс обучения нейросети, однако обучается она не с нуля. Все параметры модели уже были обучены под другую близкую задачу. При дообучении же происходит настройка под интересующую исследователя задачу. Это позволяет сэкономить ресурсы и время на обучение.

Классический ML – набор методов машинного обучения, которые используются исключительно для решения типовых (как правило, табличных) задач, то есть это все, кроме нейросетей.

Кросс-валидация – способ оценки работоспособности модели, который предполагает поэтапное обучение модели и сравнение с буду-

щим. Например, у нас есть данные с 2010 по 2021 г. по цене акций Газпрома. Сначала мы обучаем модель на данных с 2010 по 2012 г., а предсказываем для 2013 г. Потом обучаем новую модель на данных с 2010 по 2013 г., а предсказываем для 2014 г. И так далее. Это позволяет проверить адекватность предсказаний модели с течением времени.

Лемматизация – приведение слова к начальной форме – к мужскому роду единственному числу для существительных и прилагательных и к инфинитиву для глаголов.

Машинное обучение (ML от англ. Machine Learning) – класс методов искусственного интеллекта для решения задач анализа данных.

Парсинг – процесс получения данных с веб-сайтов с помощью автоматизированных программ.

Предобученный – применяется к моделям или эмбеддингам, которые уже были обучены на какую-то задачу. То есть все параметры модели и эмбеддингов уже имеют некоторую предсказательную силу, а не инициализированы случайно.

Рекуррентная нейросеть – нейросеть, в основе которой лежит рекуррентный блок, суть которого заключается в том, что он обрабатывает данные последовательно, накапливая в себе информацию.

Сверточная нейросеть – нейросеть, в основе которой лежит блок свертки, суть которого заключается в том, что он обрабатывает сразу несколько соседних элементов последовательности, тем самым работает с локальной информацией в последовательности.

Трансформер – нейросеть, в основе которой лежит комбинация блоков «Трансформеров», суть которых заключается в работе сразу со всей последовательностью целиком (а не последовательно или в некотором окне); тем самым это дает возможность учитывать сложные взаимосвязи в данных.

Эмбеддинг – числовое представление любых данных. В работе говорится о числовых представлениях текстов.

ЛИТЕРАТУРА

1. *Кузнецов Р.С., Тумарова Т.Г.* Прогнозирование котировок акций ПАО Газпром с использованием нейронных сетей LSTM // Вестник Института экономики Российской академии наук. 2023. № 3. С. 84–98. DOI: 10.52180/2073-6487_2023_3_84_98.
2. *Biau G., Scornet E.* A random forest guided tour // Test. 2016. Vol. 25. Pp. 197–227. DOI: 10.1007/s11749-016-0481-7.
3. *De Fortuny E.J. et al.* Evaluating and understanding text-based stock price prediction models // Information Processing & Management. 2014. Vol. 50. No. 2. Pp. 426–441. DOI: 10.1016/j.ipm.2013.12.002.
4. *Fama E.F.* Efficient capital markets // Journal of finance. 1970. Vol. 25. No. 2. Pp. 383–417. DOI: 10.2307/2325486.

5. *Gidofalvi G., Elkan C*. Using news articles to predict stock price movements // Department of computer science and engineering, university of California. San Diego. 2001. Vol. 17. DOI: 10.1109/IJCNN.2018.8489208.
6. *Li Y., Pan Y*. A novel ensemble deep learning model for stock prediction based on stock prices and news // International Journal of Data Science and Analytics. 2022. Pp. 1–11. DOI: 10.1007/s41060-021-00279-9.
7. *Liu J. et al*. Transformer-based capsule network for stock movement prediction // Proceedings of the First Workshop on Financial Technology and Natural Language Processing. 2019. Pp. 66–73. DOI: 10.1016/j.eswa.2022.117239.
8. *Mittal A., Goel A.* Stock prediction using twitter sentiment analysis // Standford University, CS229. 2012. Vol. 15. P. 2352.
9. *Natekin A., Knoll A*. Gradient boosting machines, a tutorial // Frontiers in neurorobotics. 2013. Vol. 7. P. 21. DOI: 10.3389/fnbot.2013.00021.
10. *Sekioka S., Hatano R., Nishiyama H*. Market prediction using machine learning based on social media specific features // Artificial Life and Robotics. 2023. Vol. 28. No. 2. Pp. 410–417. DOI: 10.1007/s10015-023-00857-z.
11. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30. DOI: 10.48550/arXiv.17 06.03762.
12. *Volodin S.N., Kuranov G.M., Yakubov A.P.* Impact of Political News: Evidence from Russia // Scientific Annals of Economics and Business. 2017. Vol. 64. No. 3. Pp. 271–287. DOI: 10.1515/saeb-2017-0018.
13. *Xu Y., Cohen S.B.* Stock movement prediction from tweets and historical prices // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). 2018. Pp. 1970–1979. DOI: 10.18653/v1/P18-1183.
14. *Zhang J., Ye L., Lai Y*. Stock price prediction using CNN-BiLSTM-Attention model // Mathematics. 2023. Vol. 11. No. 9. P. 1985. DOI: 10.3390/math11091985.

REFERENCES

1. *Kuznetsov R.S., Tumarova T.G.* Forecasting stock prices of PJSC Gazprom using LSTM neural networks // Bulletin of the Institute of Economics of the Russian Academy of Sciences. 2023. No. 3. Pp. 84–98. DOI: 10.52180/2073-6487_2023_3_84_98. (In Russ.).
2. *Biau G., Scornet E*. A random forest guided tour // Test. 2016. Vol. 25. Pp. 197–227. DOI: 10.1007/s11749-016-0481-7.
3. *De Fortuny E.J. et al*. Evaluating and understanding text-based stock price prediction models // Information Processing & Management. 2014. Vol. 50. No. 2. Pp. 426–441. DOI: 10.1016/j.ipm.2013.12.002.
4. *Fama E.F*. Efficient capital markets // Journal of finance. 1970. Vol. 25. No. 2. Pp. 383–417. DOI: 10.2307/2325486.
5. *Gidofalvi G., Elkan C*. Using news articles to predict stock price movements // Department of computer science and engineering, university of California. San Diego. 2001. Vol. 17. DOI: 10.1109/IJCNN.2018.8489208.
6. *Li Y., Pan Y*. A novel ensemble deep learning model for stock prediction based on stock prices and news // International Journal of Data Science and Analytics. 2022. Pp. 1–11. DOI: 10.1007/s41060-021-00279-9.
7. *Liu J. et al*. Transformer-based capsule network for stock movement prediction // Proceedings of the First Workshop on Financial Technology and Natural Language Processing. 2019. Pp. 66–73. DOI: 10.1016/j.eswa.2022.117239.

8. *Mittal A., Goel A.* Stock prediction using twitter sentiment analysis // Standford University, CS229. 2012. Vol. 15. P. 2352.
9. *Natekin A., Knoll A.* Gradient boosting machines, a tutorial // Frontiers in neurorobotics. 2013. Vol. 7. P. 21. DOI: 10.3389/fnbot.2013.00021.
10. *Sekioka S., Hatano R., Nishiyama H.* Market prediction using machine learning based on social media specific features // Artificial Life and Robotics. 2023. Vol. 28. No. 2. Pp. 410–417. DOI: 10.1007/s10015-023-00857-z.
11. *Vaswani A. et al.* Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30. DOI: 10.48550/arXiv.1706.03762.
12. *Volodin S.N., Kuranov G.M., Yakubov A.P.* Impact of Political News: Evidence from Russia // Scientific Annals of Economics and Business. 2017. Vol. 64. No. 3. Pp. 271–287. DOI: 10.1515/saeb-2017-0018.
13. *Xu Y., Cohen S. B.* Stock movement prediction from tweets and historical prices // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). 2018. Pp. 1970–1979. DOI: 10.18653/v1/P18-1183.
14. *Zhang J., Ye L., Lai Y.* Stock price prediction using CNN-BiLSTM-Attention model // Mathematics. 2023. Vol. 11. No. 9. P. 1985. DOI: 10.3390/math11091985.

Дата поступления рукописи: *12.09.2024 г.*

СВЕДЕНИЯ ОБ АВТОРЕ

Борисенко Георгий Александрович – аспирант экономического факультета МГУ имени М.В. Ломоносова, Москва, Россия
ORCID: 0009-0000-8430-7744
borisenko.georgiy@bk.ru

ABOUT THE AUTHOR

Georgiy A. Borisenko – Postgraduate student, Faculty of Economics, Lomonosov Moscow State University, Moscow, Russia
ORCID: 0009-0000-8430-7744
borisenko.georgiy@bk.ru

NEURAL NETWORKS TO FORECAST STOCK PRICES BASED ON NEWS DATA

This work is devoted to forecasting the movement of the stock prices of large Russian companies represented in the Moscow Exchange Index, based on news data. Transformer neural networks as well as classical machine learning are used as forecast models. Large Russian news sources and Telegram channels on economics and finance concerns are used as news data. The problem is solved in two settings: classification into 2 classes (the share price will be higher/lower than the current one) and classification into 3 classes (the share price will be higher/approximately at the same level/lower than the current one). As a result of the study, it was found that classical machine learning methods cope better with this task in the general case, but neural networks also show good quality for large companies.

Keywords: *stock price, news, neural networks.*
JEL: C63, G14.