Lomonosov Moscow State University

Moscow, Russian Federation
http://www.econ.msu.ru

Preprint series of the economic department 007/2023

«Использование нейронных сетей для прогнозирования стоимости акций на основе новостных данных»

Борисенко Георгий Александрович

Москва 2023

Аннотация

Данная работа посвящена прогнозированию стоимости акций крупных российских компаний, торгующихся на Московской бирже, на основе новостей. В качестве моделей для прогноза используются нейронные сети трансформеры. Более того, в анализе участвуют и классические методы машинного обучения для сравнения с нейросетевым подходом. В качестве новостных данных используются крупные российские новостные источники и Телеграмм-каналы. Также производится сравнение моделей, обученных на разных источниках.

В результате исследования получено, что классические методы машинного обучения справляются лучше с данной задачей в общем случае, но нейросети также показывают хорошее качество. Также в работе даются рекомендации по выбору источника новостей и выбора постановки задачи.

Ключевые слова: стоимость акций, новости, нейросетевой подход, Telegram

JEL-коды: C63, G14

Оглавление

Введение

Согласно теории эффективного рынка [Fama et al., 1969][1], капитализация публичных компаний зависит от событий, которые происходят вокруг нее. Соответственно, имея доступ к этой информации, 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 Телеграмма, где публикация новостей происходит максимально быстро. Таким образом, объектом исследования данной работы являются новости о публичных компаниях, а предметом — их взаимосвязь с движением цен акций компаний, к которым они относятся.

На данный момент лучше всего справляются с задачей обработки естественных языков современные архитектуры нейронных сетей, так как они способны агрегировать в себе огромное число выученных взаимосвязей о имеющейся информации, зачастую они способны даже превосходить результат человека. Таким образом, «ключ» к рынку может быть найден с помощью глубокого семантического анализа полученных новостей.

Цель данной работы состоит в построении моделей машинного обучения, способных предсказывать направление движения стоимости акций голубых фишек на Московской фондовой бирже, а также в выявлении слов и словосочетаний, которые оказывают влияние на эту динамику. В качестве моделей будут использоваться 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 Телеграмм каналов.

[1] Fama E. F. et al. The adjustment of stock prices to new information //International economic review. – 1969. – Т. 10. – №. 1. – С. 1-21.

Глава 1. Теоретические и практические основы прогнозирования цен акций на основе текстовых данных.

1.1 Обзор литературы

Одна из статей, цель которой — предсказание стоимости акции на основе новостного фона с помощью нейростетевого анализа, является [Li Y., Pan Y. 2021][2].

Авторы утверждают, что стоимость акции зависит от спроса и предложения на ценную бумагу. Они же, в свою очередь, зависят от множества различных факторов, которые можно определить из финансовых новостей, официальных публичных писем компаний, заявлений топ-менеджмента или финансовых отчетов.

В качестве новостных данных авторы используют крупнейшие новостные издания США. Однако они используют не весь текст интересующих статей, а только заголовки, так как по их заверениям использование полного текста может привести к сложностям при его анализе из-за возможного наличия шумовых слов, мешающих модели оценить его семантику. В качестве целевой метки используются данные по компаниям из индекса S&P 500, стоимость ценной бумаги берется по закрытию торгового дня.

Подготовка текстовых данных происходит с помощью Aware Dictionary and Sentiment Reasoner (VADER). Данная модель преобразовывает текстовые данные статьи в числовую характеристику, которая позволяет судить о степени позитивности или негативности новости. Данные по ценам акций авторы переформатировали в интервал от 0 до 1, чтобы избежать переобучения и увеличить точность.

Построенная модель машинного обучения представляет из себя комбинацию из нескольких рекуррентных нейронных сетей (RNN): нейросети долгой краткосрочной памяти (LSTM), управляемой рекуррентной нейронной сети (GRU) и полносвязная нейронная сеть. LSTM модель позволяет разделять новости про продолжительности их влияния, в то время как GRU модель имеет меньшее время вычислений. В зависимости от ситуации одна модель может быть лучше другой, поэтому необходима их комбинация.

После получения обработанных текстов, необходимо перейти к задаче классификации. С методами соответствующего анализа можно ознакомиться в статье [Vajrala A. 2019][3].

[2] 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

В статье предлагается оформить данные, полученные после подготовки текста в виде TF-IDF матрицы. Такая структура данных позволяет учесть то, как часто слово встречается в конкретном документе и как редко оно встречается в каком-либо другом. То есть учитываются не просто слова в тексте, а степень их принадлежности к какому-то классу или категории. Также плюс такого подхода заключается в том, что такую матрицу можно расширить на комбинации слов, то есть будет учитываться семантический смысл словосочетаний, однако при этом сильно возрастет размерность данных.

Автор приводит множество возможных вариантов моделей, которые могут быть использованы, но больше всего внимания уделяет нейронным сетям Двунаправленной долгой краткосрочной памяти (BLSTM). Данная структура нейронной сети в обработке текстов показывает себя хорошо, так как она способна “запоминать” события из прошлого, то есть как LSTM нейронная сеть, но при этом двунаправленная модификация позволяет учесть не только слова, которые шли до текущего слова, но и слова после. Такой подход позволяет учесть семантику настоящего текста, где смысл определенного слова не всегда зависит от смысла прошлых, но и будущих.

Трансформеры на данный момент — наиболее успешные архитектуры нейронных сетей для работы с текстовыми данными, так как они в отличие от RNN и CNN способны акцентировать внимание сразу на всем тексте новости. Поэтому просто необходимо опробовать трансформер в сравнении с CNN или RNN. Этим уже занимались в статье [Liu J. et al. 2019][4].

Авторы с помощью нейросети с архитектурой CapTE на основе трансформера делали предсказание направление движения акции до закрытия торгового дня по твитам с соответствующими названию компании тегами. То есть они учитывали не только объективную информацию, но и мнения любого пользователя Твиттера, который решил о ней высказаться.

В результате авторы получили, что трансформер превосходит сверточные и рекуррентные нейросети. Однако, как мне кажется, то, что они учитывали мнение всех инвесторов, могло добавить шума в модель, ведь рынок не всегда движется так, как этого ожидают активные пользователи Твиттера, поэтому стоит рассматривать только объективную или кажущейся на момент выхода новости объективной информацию.

[3] Vajrala A. Text Classification

[4] 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. – С. 66-73.

Также они рассматривали только предсказание на день вперед, но можно рассмотреть и другие временные интервалы.

1.2 Выводы из обзора литературы.

Проанализировав статьи по схожей с моей работой тематикой, можно сделать вывод, что прогнозирование стоимости акций на основе новостей с помощью нейросетей возможно.

Сверточные нейросети хорошо воспринимают локальный контекст слова, то есть они могут видеть заданное количество слов слева и справа от конкретного слова в некоторой последовательности, однако, они не могут выучить взаимосвязи слов, которые лежат за пределами этого окна. Рекуррентные нейросети воспринимают всю последовательность слов в предложении, однако они больше внимания уделяют именно концу последовательности, так как они работают последовательно и более ранние слова в предложении сетью просто забываются. Также один проход по данным занимает много времени, так как они не могут работать параллельно из-за своего дизайна. Более того, им надо много итераций для обучения, так как градиент при обратном распространении ошибки в рекуррентных сетях затухает.

Проблемы сверточных и рекуррентных нейросетей могут решить трансформеры, так как они одновременно анализируют весь текст новости одновременно, однако это делает модель очень тяжеловессной, то есть в ней присутствует огромное число параметров, и для обучения такой сети может понадобиться большое число вычислительных мощностей и данных. Однако веса трансформеров можно найти в Интернете. Эти веса будут заточены под общие языковые задачи, но модели можно дообучить так, чтобы сеть лучше воспринимала финансовые новости. Таким образом, в качестве нейросетевого подхода будет использована предобученная нейросеть трансформер.

Если говорить про выбор метрик для оценки результатов моделей, то можно сделать вывод, что для задачи классификации стоит использовать стандартные для этого метрики Accuracy, Precision. Recall, AUC-ROC, F1.

3 Данные

Весь код и ссылки на все данные и модели можно найти по ссылке на репозиторий на GitHub https://github.com/BorisenkoGeorgy/Disser_news_stocks/tree/dev.

3.1 Классические новости

В качестве источника традиционных новостных данных были выбраны крупные новостные издания России (РИА новсти, Комерстант, Лента Ру, Ведомости). Новости с них были получены с помощью самостоятельно реализованных на языке программирования Python парсинговых программ. Всего было получено 716740 уникальных текстов за интервал с 1 января 2010-ого года по 20 ноября 2022-ого года.

3.2 Новости из Телеграмма

В Телеграмме 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 покрыть информационное поле публичных компаний.

На данный момент на рынке ценных бумаг я нахожусь уже более 5 лет, последние 3 года я получаю новостную информацию о компаниях только из Телеграмм-каналов. Поэтому, в качестве базового набора каналов, я выбрал каналы, которыми пользуюсь сам. Далее с помощью самомстоятельно написанной парсинговой программы, реализованной на языке Python, я получил все новости из этих каналов.

Телеграмм-каналы часто пересылают новости друг от друга, и эта информация может быть получена также может быть получена при парсинге. Таким образом, из полученных мной данных можно также извлечь информацию о каналах, на которые ссылается конкретный выбранный канал. В результате я составил матрицу популярности каналов. То есть я определял популярность канала по тому, сколько каналов и как часто на этот канал ссылаются. Сделал отсечку, что в среднем должно быть 25 ссылок на некоторый канал, чтобы можно было его считать значимым. Некоторые каналы из полученных таким образом оказывались закрытыми каналами с торговыми сигналами, такие каналы я игнорировал, потому что они содержат в основном информацию о рекомендации по сделке, а не сами новости.

Далее, я итеративно совершал действия, описанные выше, пока к списку каналов не перестали добавляться новые каналы, таким образом, я считаю, мне удалось охватить максимально широкое информационное поле, ограничившись условно небольшим числом Телеграмм-каналов. Итого было получено 1043208 уникальных текстов за период, начиная от создания каждого канала до 15 января 2023.

3.3 Предобработка 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 \n и \t, затем очищены от стоп слов (различные предлоги, местоимения, частицы или другие части речи, которые часто встречаются в текстах и не несут значимой семантической нагрузки). Также при обработке текстов Телеграмм-каналов были удалены смайлики. После это все тексты были приведены к нижнему регистру и лемматизированы. Затем к полученным текстам был применен метод TF-IDF.

Нейросетевые модели используют уже предобученные эмбеддинги, которые могут работать с предложениями практически в сыром виде. Были удалены лишь специальные символы \n, \r и смайлики. Затем все слова в предложениях преобразуются к соответствующим номерам, соответствующим строкам обученного эмбеддинга.

3.4 Выбор ценных бумаг

Для анализа мной были выбраны ценные бумаги, входящие в Индекс Мосбиржи на ноябрь 2021, когда данная работа задумывалась. Из выборки я исключил такие компании (далее компании будут назваться их биржевыми тикерами) как OZON, VKCO, POGR и HHRU, так как на тот момент они недавно появлялись на бирже и из-за этого по ним собрано довольно мало информации, более того, на данный момент POGR находится в процессе банкротства и акции этой компании не торгуются на бирже.

Также из списка бумаг я исключил привилегированные акции SBERP и TATNP, так как в моей выборке присутствует SBER и TATN и их бумаги коррелируют практически с коэффициентом 1 за очень редкими исключениями.

Итоговый список тикеров ценных бумаг, для которых были построены модели машинного обучения, выглядит следующим образом: AFKS, AFLT, ALRS, CBOM, CHMF, DSKY, FEES, GAZP, GMKN, HYDR, IRAO, LKOH, LSRG, MAGN, MOEX, MTSS, NLMK, NVTK, PHOR, PIKK, PLZL, ROSN, RTKM, RUAL, SBER, SNGS, TATN, TCSG, TRNFP, VTBR, YNDX.

3.5 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 языке Python с сайта финансового портала Финам.ру[5]. По каждой компании были получены данные цены открытия, минимальной, максимальной, закрытия и объем торгов в денежном эквиваленте по интервалам 1, 5, 10, 15, 30 минут, 1 час и 1 день. Впоследствии были использованы только минутные данные, так как из них можно получить данные по всем остальным интервалам.

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.

3.6 Разметка данных с помощью регулярных выражений.

Некоторую часть информации, которая заложена в тексте можно получить из текстов путем поиска ключевых слов в этом тексте. 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.

[5] https://www.finam.ru/

С помощью регулярных выражений были отобраны новости, которые относятся к конкретным компаниям и секторам. Например, если в тексте новости встречается «Сбербанк», то эта новость относится к ПАО «Сбербанк». Регулярные выражения состояли не только из названий компаний, но и из их сокращенных названий, биржевых тикеров и названий дочерних предприятий. Для секторов были перечислены основные элементы секторов, например, элементы финансовых рынков для финансового сектора или различные металлы для сектора цветной металлургии.

4 Построение Бейзлайн-моделей

В качестве бейзлайн-моделей мной были выбраны модели Случайного леса и Градиентного бустинга над деревьями из-за своей относительной легкости работы с большими данными. Логистическая регрессия многой не применялась, так как очень плохо работает с данными большой размерности, также не применялся алгоритм SVM, так как его вычислительная сложность $O(n^2)$, что затрудняет его применение на большой выборке данных с использованием процессора, а ресурсы видеокарты лучше потратить на основную цель исследования — нейросети.

Для бейзлайна текстовые данные приводились к числовым с помощью метода TF-IDF. Метод применялся к прошедшим обработку текстам с удалением пунктуации, стоп-слов, лемматизации и приведению текстов к нижнему регистру. Максимальный размер векторного представления был ограничен 10000 сверху. После применения алгоритма оказалось, что этот максимум был достигнут. Также ограничением снизу выступала минимальная встречаемость токена не менее 5 раз. Под токеном подразумевается конкретное слово или пара слов (биграмма). Биграммы используются, чтобы учесть совместную встречаемость слов в текстах.

Подбор параметров для моделей классического машинного обучения проводился с помощью перебора по сетке параметров. Выбор оптимальных параметров осуществлялся с помощью кросс-валидации. Кросс-валидация использовалась обычная, а не для временных рядов, так как формально временного ряда в объясняющих переменных нет, а задача стоит в классификации новости по тексту. Сами тексты между собой имеют ограниченную временную структуру.

Перебиралась только глубина максимальная глубина деревьев, как один из самых важных параметров. Все параметры перебрать вычислительно долго. Обучение и кросс-валидация проводилось на всех доступных данных до 1 июня 2021 не

включительно. Под тестовую часть были выделены данные с 1 июня 2021 по 1 января 2022 не включительно. Для каждого временного интервала, каждой компании и каждого источника была построена отдельная модель классического машинного обучения. Каждая модель была сохранена и ее можно найти по ссылке в репозитории на GitHub

Чтобы показать наличие информации в текстах предсказания моделей сравнивались с двумя простыми бенчмарками. Предсказание случайным классом с вероятностями, полученными на обучающей части выборки и предсказание нейтральным классом.

4.1 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 хорошо. Полные результаты можно найти в репозитории GitHub.

Для алгоритма случайного леса были получены следующие результаты (Таблица 1).

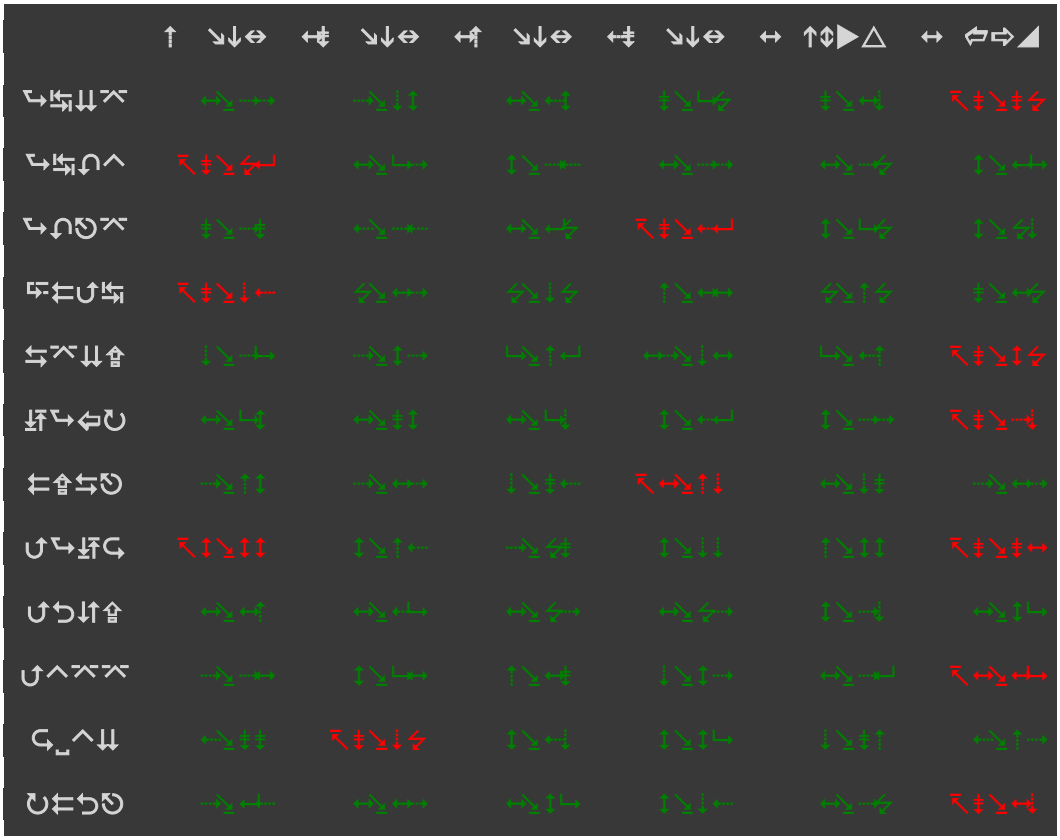


Таблица 1 Сокращенные результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Телеграмма на 3 класса алгоритмом случайного леса

Для алгоритма бустинга имеем минимальный эффект на Accuracy для качественно предсказанных компаний (Таблица 2):

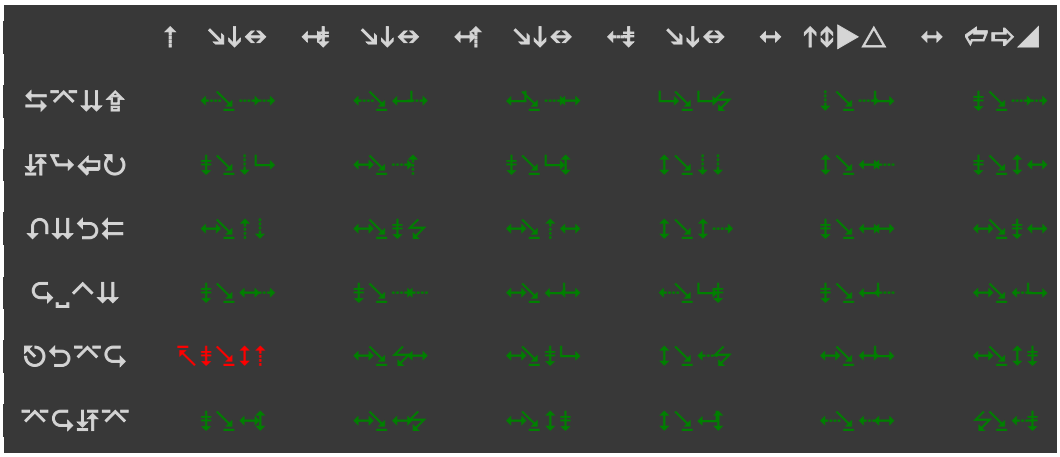


Таблица 2 Сокращенные результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Телеграмма на 3 класса алгоритмом бустинга над деревьями

4.2 Результаты для Традиционных источников новостей

В данном разделе будут рассмотрены полученные результаты для моделей случайного леса и градиентного бустинга над деревьями для классических новостных источников. Результаты в статье показаны неполные, а только те, где модели показали себя хорошо. С полными результатами можно ознакомиться в репозитории GitHub.

Рассмотрим результаты для случайного леса (Таблица 3):

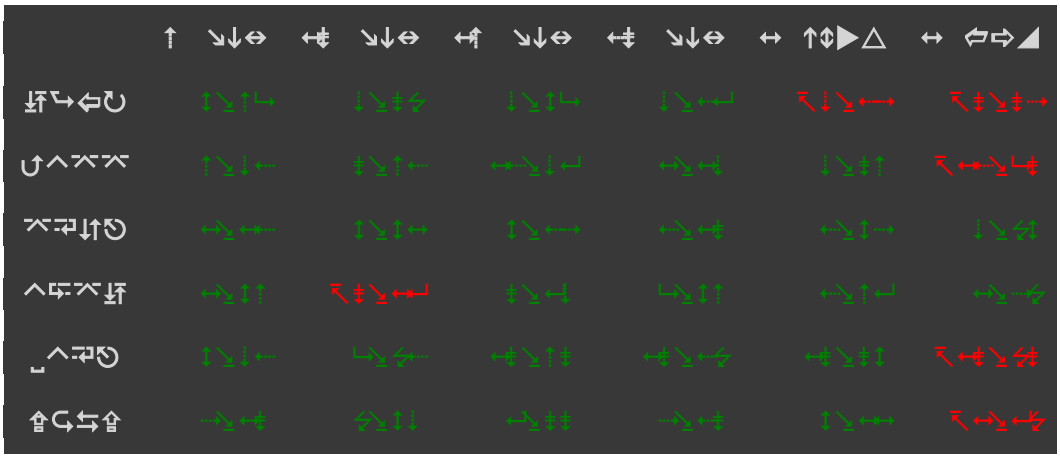


Таблица 3 Сокращенные результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Классических источников на 3 класса алгоритмом случайного леса

Теперь обратимся к результатам работы алгоритма градиентного бустинга над деревьями (Таблица 4).

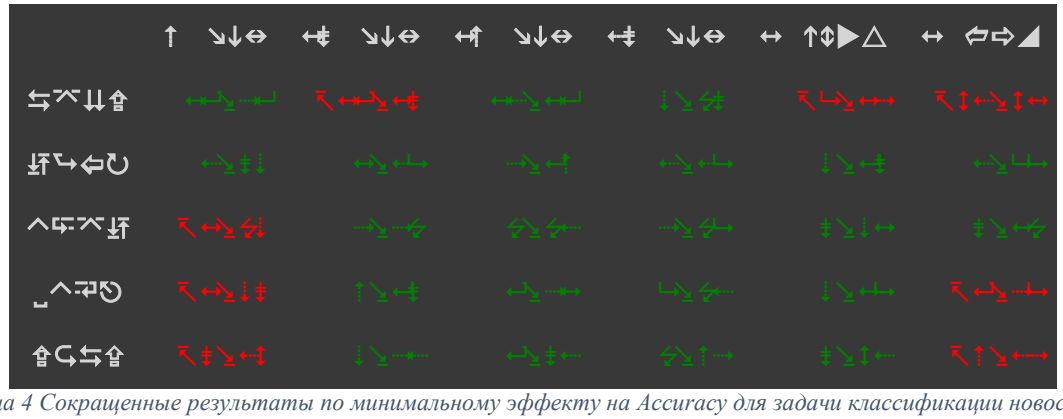


Таблица 4 Сокращенные результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Классических источников на 3 класса алгоритмом бустинга над деревьями

4.3 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. Таккже больше значимых эффектов наблюдается для новостей, полученных из Телеграмма, но все-таки для полноты картины необходимо свести все результаты в таблицу, чтобы сравнить между собой различные модели и источники (Таблица 5).

Задача	Источник	Модель	5 min	10 min	15 min	30 min	1 hour	1 day
3 класса	Телеграмм	Случайный лес	0.80%	1.65%	2.02%	1.78%	2.36%	0.45%
3 класса	Телеграмм	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%

Таблица 5 Сводная таблица по средним минимальным эффектам для Accuracy

5 Нейросетевой подход

В качестве нейросетей модели были опробованы несколько вариантов трансформеров с сайта Hugging Face[6]. Это сайт, на котором в отрытом доступе можно найти уже обученные веса для интересующих моделей. sbert_large_nlu_ru[7], ruRoberta-large[8] и distilrubert-base-cased-conversational[9]. Но в результате была выбрана только одна модель distilrubert-base-cased-conversational, так как из-за своих небольших размеров данная модель позволяла выбрать размер батча как 32. Остальные модели также возможно было обучить, но с меньшем размером батча, что влияло на сходимость, модели очень медленно обучались, как в смысле времени на одну эпоху, так и в скорости падения функции потерь даже на обучающей выборке.

Для работы получения ответов для задачи классификации необходимо было достроить классификатор поверх трансформера. Было опробовано два варианта:

1) Из выходов трансформера получать только выход, который отвечает за токен <CLS> и далее сверху добавить два линейных слоя с нелинейностями и Dropout между ними;

2) Обрабатывать все выходы трансформера, усредняя контекст и конкатенируя его с выходами токена. Это описано в статье, посвященной решению соревнования с помощью трансформера sbert[10]. Хоть для решения поставленной в работе задачи и была использована другая архитектура трансформера, данный метод оказался лучше.

Для обучения нейросети кросс-валидацию использовать затруднительно, так как это вычислительно трудно, 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. Итого данные были разбиты по интервалам от начала собранных данных до 1 января 2021 не включительно для обучающей выборки, от 1 января 2021 до 1 июня 2021 не включительно для валидационной выборки и от 1 июня 2021 до 1 января 2022 не включительно для тестовой выборки. Получаем, что тестовые выборки для нейросетевого подхода и классического ML совпадают.

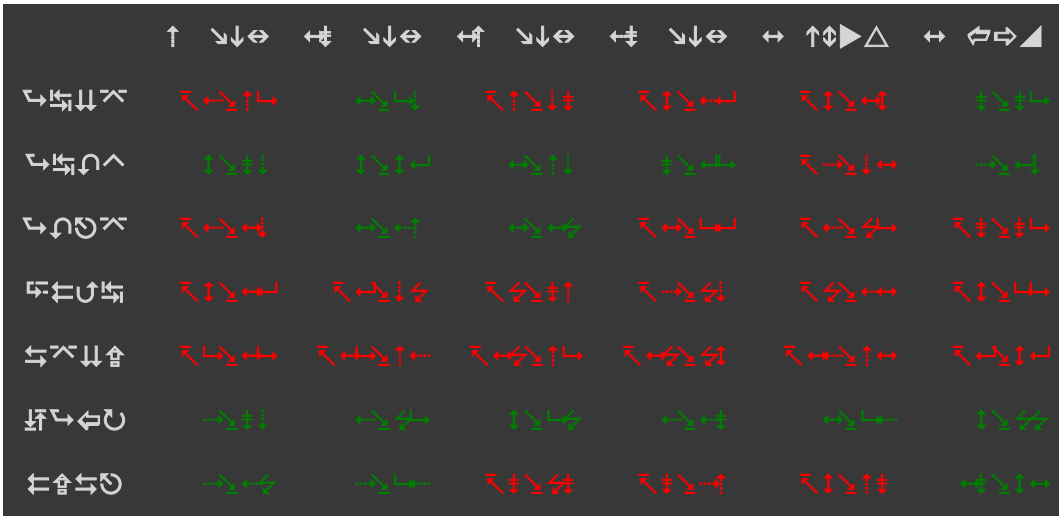
[6] https://huggingface.co/
[7] https://huggingface.co/ai-forever/sbert_large_nlu_ru
[8] https://huggingface.co/ai-forever/ruRoberta-large
[9] https://huggingface.co/DeepPavlov/distilrubert-base-cased-conversational
[10] https://habr.com/ru/company/sberdevices/blog/527576/

Формально модель может иметь множество выходов и предсказывать сразу несколько целевых переменных. Однако в таком подходе есть проблема. Нейросеть будет оптимизировать среднее значение (арифметическое или геометрическое, в зависимости от того, как задать итоговую функцию потерь) функций потерь для каждого выхода сети, а не каждый выход в отдельности. Из-за этого результаты сети будут получены вероятно хуже, так как на примере результатов классического ML видно, что модели не всегда справляются с задачей прогнозирования на всех интервалах хорошо. 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 хорошее качество. Не на всех временных 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.

Каждая модель обучалась на протяжении 8 эпох. После каждой эпохи на валидационной выборке вычислялись метрики Accuracy, Precision, Recall, F1-score, ROC-AUC и Каппа Коэна. На основе метрики Accuracy сохранялись оптимальные веса модели. После окончания обучения, оптимальные веса заново загружались в модель и выполнялось предсказание на тестовый набор данных.

5.1 Результаты для Телеграмма

Рассмотрим результаты нейросетвего подхода для классификации на 3 класса для Телеграмма (Таблица 6).



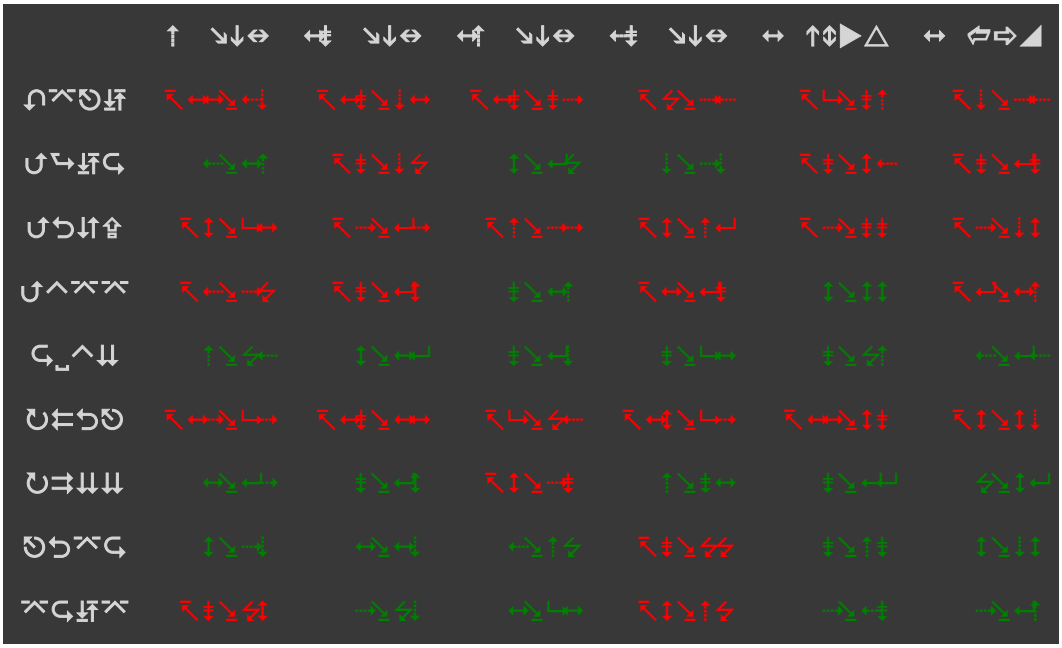


Таблица 6 Результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Телеграмма на 3 класса нейросетью

Теперь сравним разность минимальных эффектов для нейросетового подхода и случайного леса (Таблица 7).

Таблица 7 Сравнение по Accuracy Нейросетевого подхода и алгоритма случайного леса для задачи классификации на 3 класса по новостям из Телеграмма

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

5.2 Результаты для классических новостей

Рассмотрим результаты задачи классификации на 3 класса для классических новостных источников с помощью нейросети (Таблица 8).

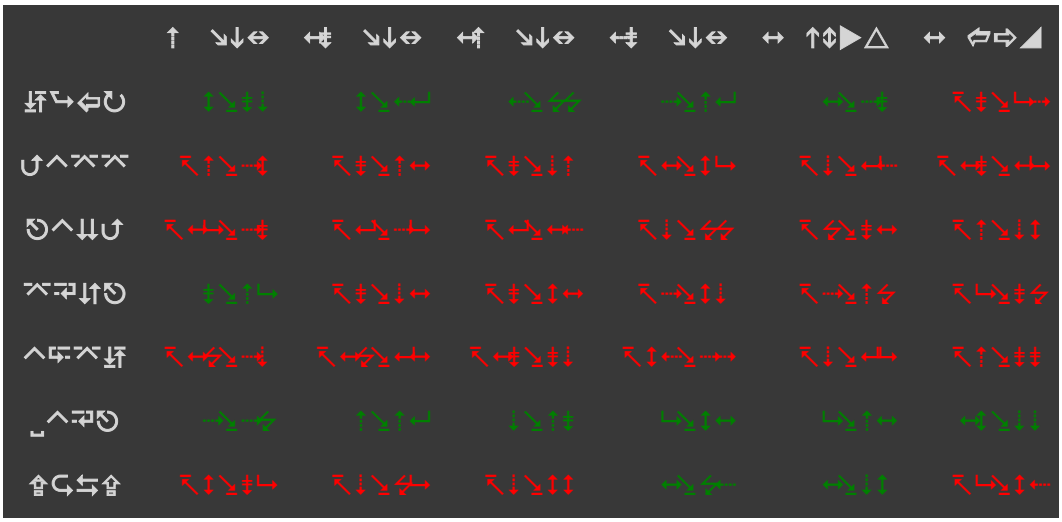


Таблица 1 Результаты по минимальному эффекту на Accuracy для задачи классификации новостей из Классических источников на 3 класса нейросетью

Теперь рассмотрим в сравнении с эффектами для случайного леса (Таблица 9).

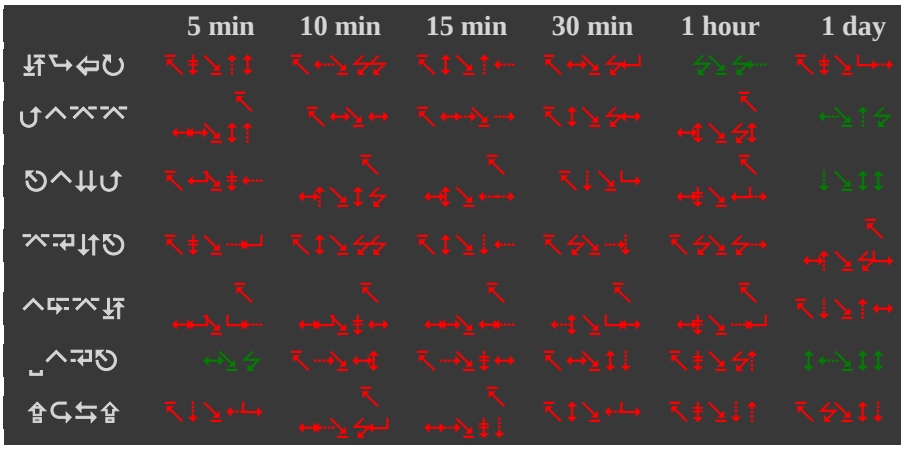


Таблица 2 Сравнение по Accuracy Нейросетевого подхода и алгоритма случайного леса для задачи классификации на 3 класса по новостям из Классических источников

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.

5.3 Итоговые результаты для нейросетевого подхода

Как видно из попарного сравнения таблиц метрик для нейросетевого подхода и случайного леса по результатам исследования получаем, что нейросети на российской фондовой 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. Именно поэтому нейросеть в данном случае смогла показать качество выше классического ML.

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

Превосходство случайного леса над нейросетью можно объяснить несколькими способами:

1. Кросс-валидация для случайного леса проводилась обычная, а не специальная для временных рядов, так как формально временной ряд не присутствует в независимых переменных, исследование заключается в влиянии 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 слишком вычислительно 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 данных их постов и комментариев социальных сетях, что не относится к области финансовых текстов. Дообучение нейросети частично эту проблему, но только частично. Для полного решения 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-так приведет к численному виду, который был

оптимален для задачи, на которой изначально обучалось это числовое представление, а не для задачи классификации финансовых текстов.

5.4 Интерпретация предсказаний нейросети

Несмотря на то, что качество для нейросети в среднем хуже, чем для случайного леса, у нейросетевого подхода есть преимущество — благодаря механизму самовнимания, заложенного в идею архитектуры трансформера, можно визуализировать, на что ориентируется модель при принятии решений о классификации. Компании и интервалы, для которых строились визуализации выбирались по F1-мере. Чем она больше, тем более разнообразные ответы модель давала в предсказании, что дает возможность выбирать наиболее интерпретируемые правильно классифицированные новости.

Рассмотрим примеры новостей Телеграмма из тестовой выборки на задаче 3-ех классов (так как для нее было получено наибольшее число отличных от случайных предсказаний), для которых нейросеть успешно определила метку класса и попробуем их проинтерпретировать.

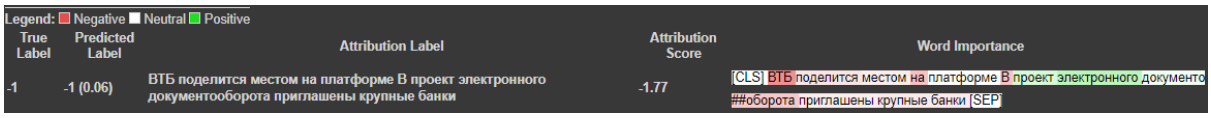
Рассмотрим новость о Сургутнефтегазе от 2021-12-30 из Телеграмм канала finascop: «Сургутнефтегаз ап растет вместе с курсом доллара».

Как видно из визуализации механизма внимания, нейросеть делает наибольший положительный акцент на связи нефти и газа с ростом доллара. Действительно, ПАО «Сургутнефтегаз» занимается добычей нейти 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. Более того, Сургутнефтегаз на момент выхода новости хранил большие суммы валюты в виде денежных средств на банковском счету[11]. В результате валютных

[11]https://www.tadviser.ru/index.php/%D0%9A%D0%BE%D0%BC%D0%BF%D0%B0%D0%BD%D0%B8%D1%8F:%D0%A1%D1%83%D1%80%D0%B3%D1%83%D1%82%D0%BD%D0%B5%D1%84%D1%82%D0%B5%D0%B3%D0%B0%D0%B7#:~:text=%D0%9A%D0%B0%D0%BA%20%D0%BF%D0%B5%D1%80%D0%B5%D0%B4%D0%B0%D0%B5%D1%82%20%C2%AB%D0%98%D0%BD%D1%82%D0%B5%D1%80%D1%84%D0%B0%D0%BA%D1%81%C2%BB%2C%20%D0%BB%D0%B8%D0%BA%D0%B2%D0%B8%D0%B4%D0%BD%D1%8B%D0%B5,%D0%B4%D0%BE%D1%81%D1%82%D0%B8%D0%B3%D0%BB%D0%B8%204%2C13%20%D1%82%D1%80%D0%BB%D0%BD%20%D1%80%D1%83%D0%B1%D0%BB%D0%B5%D0%B9.

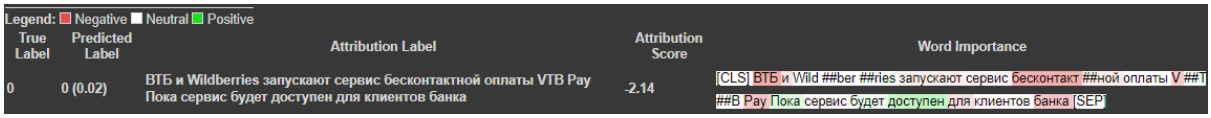
переоценок в годы резкого курса Сургутнефтегаз рекомендовал к выплате большие дивиденды, что приводило к росту стоимости его акций. В итоге имеем, что стоимость акций Сургутнефтегаза стремится к росту вместе с курсом доллара, что смогла определить нейросеть.

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



Банк ВТБ был одним из создателей в партнёрстве с «Ростелекомом» программы электронного документооборота для упрощения взаимодействий между гражданами, государством и бизнесом[12]. Однако на момент выхода новости в «Комерсанте» 2021-09-08 в развитие программы 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 акции и подчеркивает негативным весом, что ВТБ будет делится с крупными банками.

Также рассмотрим нейтральную новость из издания «Ведомости» от 2021-12-07. «ВТБ и Wildberries запускают сервис бесконтактной оплаты VTB PAY Пока сервис будет доступен для клиентов банка».



Здесь также модель выделяет некоторые положительные и отрицательные моменты новости, но относит ее все же к нейтральному классу. На тот момент Apple Pay и Sumsung Pay в России еще работали и недобности 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.

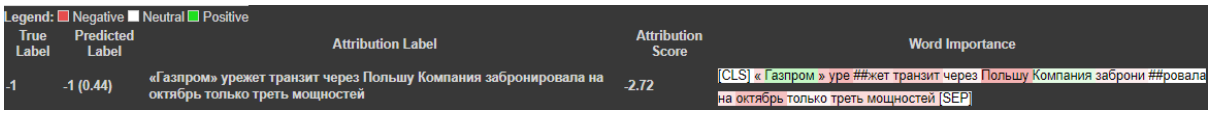
Теперь рассмотрим новость также о ПАО «ВТБ» из издания Лента.ру от 2021-07-28. «ВТБ создаст экосистему рынка имущественных торгов». Модель больше всего внимания акцентирует на том, что ВТБ займется созданием некоторого проекта по имуществу. Банку этот проект будет выгоден с точки зрения реализации заложенного

[12] https://regnum.ru/news/polit/3366298.html

имущества, которое попало в собственность банка. Тем самым банк будет быстрее избавляться от ненужных ему активов и сможет эффективнее направлять свободные денежные средства в операционную деятельность. Рынок положительно отреагировал на эту новость, и модель смогла предсказать это.

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) акции ВТБ упали примерно на 40%. А с момента IPO (апрель 2007) на 70%.

Еще стоит обратиться к самой популярно 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 Коммерсанте 2021-09-20.



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.

6 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, между собой.

Также удалось выявить, что новости из Телеграмма — более надежный источник с точки зрения качества предсказания моделей по сравнению с классическими новостными источниками в лице крупных изданий.

К сожалению, для нейросетевого подхода в общем случае не удалось получить качество лучше, чем для случайного леса. Однако в частном случае было получено, что нейросеть лучше прогнозирует движение акций для GAZP. Также удалось показать, что нейросеть делает свои предсказания обосновано путем визуализации матриц внимания и их содержательной интерпретации.

Также при написании работы было реализовано 5 парсинговых программ для получения новостей из использованных источников. Код для них хранится в свободном доступе и может быть использован исследователями в дальнейшем. Более того, все собранные данные также хранятся в свободном доступе.

В качестве идей для дальнейших исследований возможностей машинного обучения прогнозировать движение акций на российском фондовом рынке можно попробовать добавить в модели авторегрессинную компоненту. Также в модели можно добавить временные ряды биржевых товаров и валюты.

7 Список литературы

1. Aizawa A. An information-theoretic perspective of tf–idf measures //Information Processing & Management. – 2003. – Т. 39. – №. 1. – С. 45-65. [10]
2. Atsalakis G. S., Valavanis K. P. Surveying stock market forecasting techniques–Part II: Soft computing methods //Expert Systems with applications. – 2009. – Т. 36. – №. 3. – С. 5932-5941.
3. Biau G., Scornet E. A random forest guided tour //Test. – 2016. – Т. 25. – С. 197-227. [9]
4. De Fortuny E. J. et al. Evaluating and understanding text-based stock price prediction models //Information Processing & Management. – 2014. – Т. 50. – №. 2. – С. 426-441.
5. Fama E. F. et al. The adjustment of stock prices to new information //International economic review. – 1969. – Т. 10. – №. 1. – С. 1-21. [1]
6. Kannan S. et al. Preprocessing techniques for text mining //International Journal of Computer Science & Communication Networks. – 2014. – Т. 5. – №. 1. – С. 7-16. [3]
7. Kozhevnikov V. A., Pankratova E. S. RESEARCH OF TEXT PRE-PROCESSING METHODS FOR PREPARING DATA IN RUSSIAN FOR MACHINE LEARNING //Theoretical & Applied Science. – 2020. – №. 4. – С. 313-320 [4]
8. 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. – С. 1-11. [2]
9. 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. – С. 66-73. [6]
10. Luan Y., Lin S. Research on text classification based on CNN and LSTM //2019 IEEE international conference on artificial intelligence and computer applications (ICAICA). – IEEE, 2019. – С. 352-355.
11. Mittal A., Goel A. Stock prediction using twitter sentiment analysis //Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf). – 2012. – Т. 15. – С. 2352. [13]
12. Natekin A., Knoll A. Gradient boosting machines, a tutorial //Frontiers in neurorobotics. – 2013. – Т. 7. – С. 21. [9]

13. Peng Y., Jiang H. Leverage financial news to predict stock price movements using word embeddings and deep neural networks //arXiv preprint arXiv:1506.07220. – 2015
14. Roll R. Eugene F. Fama, Lawrence Fisher, Michael C. Jensen //Modern Developments in Investment Management: A Book of Readings. – 1978. – С. 177.
15. Rong X. word2vec parameter learning explained //arXiv preprint arXiv:1411.2738. – 2014.
16. Vajrala A. Text Classification. – 2019. [5]
17. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30. [11]
18. Volodin S. N., Kuranov G. M., Yakubov A. P. Impact of Political News: Evidence from Russia //Scientific Annals of Economics and Business. – 2017. – Т. 64. – №. 3. – С. 271-287. [7]
19. 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 (Volume 1: Long Papers). – 2018. – С. 1970-1979.