ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 385–394

УДК 517.54

ПОЛУЧАЕМ ЛИ МЫ ПОЛЬЗУ ОТ КАТЕГОРИЗАЦИИ ПОТОКА НОВОСТЕЙ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ ЦЕН АКЦИЙ?

© 2023 г. Т. Д. Куликова[1], Е. Ю. Ковтун[2,*], С. А. Буденный[3]

Представлено академиком РАН А. А. Шананиным
Поступило 04.09.2023 г.
После доработки 08.09.2023 г.
Принято к публикации 18.10.2023 г.

Машинное обучение широко применяется в задаче прогнозирования цен на акции публичных компаний. Для построения более точной прогностической модели необходимо учитывать не только исторические данные о ценах на акции, но и относящиеся к ним знания из внешнего мира. Такую ценную информацию может дать эмоциональная окраска финансовых новостей, связанных с компанией. Однако финансовые новости могут быть разделены на различные тематические группы, например, *Macro* (новости, относящиеся к теме "макроэкономика"), *Market* (новости о ситуации на различных рынках) или *Product news* (новости о продукте, который производит та или иная компания). В задачах исследования рынка обычно не принимается во внимание такая категоризация. В данной работе мы стремимся восполнить этот пробел и изучить эффект от учета разделения новостей на группы по тематическому признаку в задаче прогнозирования цен на акции публичных компаний. Сначала мы получаем индикаторы настроения финансовых новостей, затем классифицируем их поток на 20 заранее определенных тем с помощью предварительно обученной модели. Кроме того, мы проводим эксперименты с несколькими хорошо зарекомендовавшими себя моделями для прогнозирования временных рядов, включая темпоральную конволюционную сеть (Temporal Convolutional Network), D-линейную модель (D-Linear), трансформатор (Transformer) и темпоральный синтезирующий трансформатор (Temporal Fusion Transformer). Результаты нашего исследования показывают, что использование информации об эмоциональной окраске новостей из отдельных тематических групп способствует улучшению метрик работы моделей глубокого обучения по сравнению с подходом, когда рассматриваются все новости без какого-либо разделения.

Ключевые слова: Финансовые новости, Фондовый рынок, BERT, Тематическая классификация, Анализ эмоциональной окрашенности текста, Прогнозирование временных рядов, Глубокое обучение, Внешние данные

DOI: 10.31857/S2686954323601926, **EDN:** CWFYFQ

1. ВВЕДЕНИЕ

Связь между фондовым рынком компании и относящимися к ней новостями и событиями всегда была одной из самых актуальных тем для обсуждения. Представленная в нашей работе область исследования применяется в самых разных социальных и экономических сферах. Например, состояние рынков крупных фармацевтических компаний зависит от последних анонсов клинических исследований, поскольку за ними пристально следит общественность [1]. Более того, рост популярности интернет-трейдинга в последние годы усилил потребность в обращении к информационному фону. Многие трейдеры стали использовать информацию о настроениях новостей, которую можно получить с помощью компьютерных алгоритмов, способных быстро определять, являются ли новостная статья или сообщение в Twitter позитивно или негативно окрашенными. Главный вопрос заключается в том, какие именно новости следует учитывать для лучшего понимания будущих рыночных тенденций компаний. В данном исследовании мы пытаемся понять, имеет ли смысл разделять новости на тематические группы при передаче их в качестве внешних данных в модель глубокого обучения для прогнозирования цен на акции. Мы оцениваем, дают ли некоторые тематические группы возможность добиться большей эффективности модели или

[1]*Национальный исследовательский университет Высшая школа экономики; Факультет компьютерных наук, Москва, Россия*
[2]*Лаборатория искусственного интеллекта Sber AI Lab, Москва, Россия*
[3]*Институт искусственного интеллекта AIRI, Москва, Россия*
*****E-mail: eykovtun@sberbank.ru*

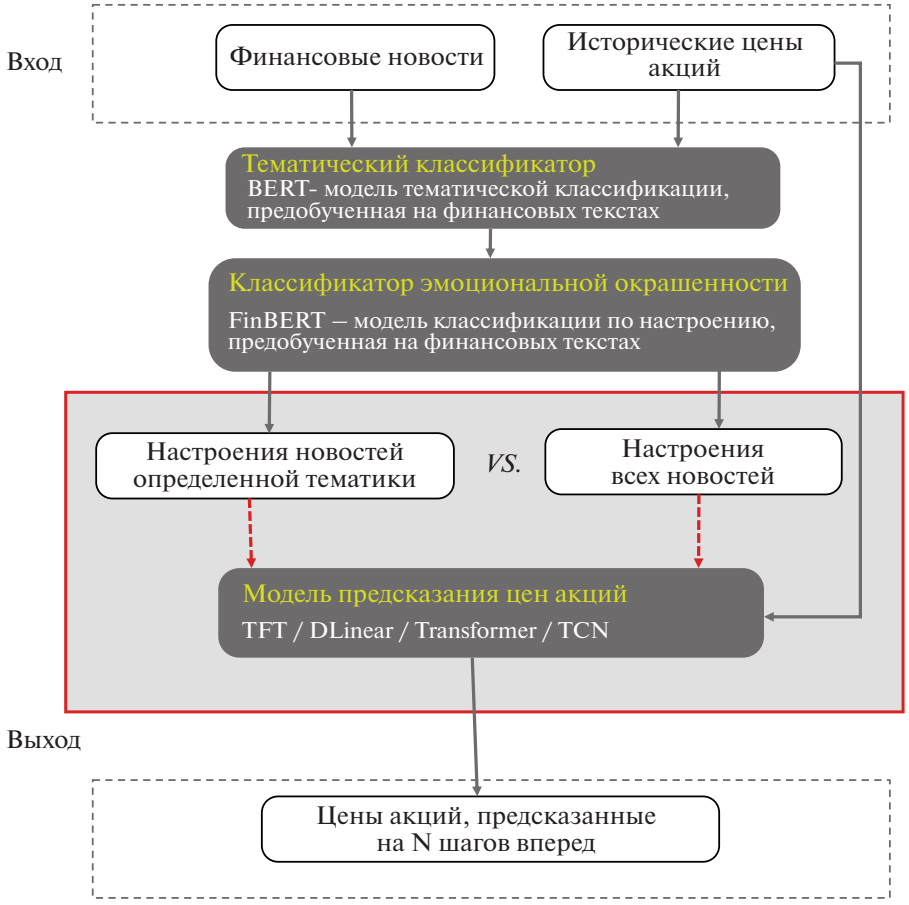

Рис. 1. Схема сравнения использования настроений всех новостей и настроений новостей определенной тематики в задаче прогнозирования цен на акции.

все же лучше передавать модели всю информацию без какого-либо разделения. Для проведения экспериментов мы выбрали пять крупных технологических публичных компаний: Apple, Amazon, Google, Netflix и Tesla. Схема нашего исследования представлена на рис. 1.

2. ОБЗОР ЛИТЕРАТУРЫ

Основным направлением исследований, описанных в нашей статье, является повышение качества прогнозирования цен на акции крупных компаний. Данное направление прогнозирования временных рядов рассматривается в большом количестве работ. В частности, в работе [2] сравнивается качество прогнозов цен на акции, сделанных различными алгоритмами. Для эксперимента выбраны четыре алгоритма ML (Machine learning) и DL (Deep learning). Среди них искусственные нейронные сети, модель Support Vector Regression, случайный лес и модель Long Short-Term Memory. Полученные результаты свидетельствуют о том, что лучшие метрики показывает модель LSTM, а значит, можно предположить, что глубокое обучение работает лучше, чем классическое машинное обучение. Основная идея другого исследования [3] заключается в сравнении 12 алгоритмов машинного обучения для классификации с использованием различных типов внешних данных в качестве дополнительной информации, а именно: новостей из средств массовой информации, финансовых новостей и двух упомянутых типов новостей одновременно. Результаты показывают, что на независимом тестовом наборе данных в качестве модели с информацией о новостях из средств массовой информации как дополнительным признаком лучше всего показал себя алгоритм Random Forest. При использовании информации только из финансовых новостей Random Forest также оказался лучшим. Необходимо отметить, что модель, в которой использовался весь новостной поток, показала наихудшие результаты по сравнению с другими моделями.

Категоризация новостных потоков является естественной проблемой, возникающей при работе с потоками новостей. В работе [4] представлен способ кластеризации новостей для моделирования их

тематики. Основная идея заключается в использовании алгоритма иерархической агломерационной кластеризации. Построенное иерархическое дерево кластеров позволяет выбрать оптимальное количество кластеров на определенной высоте дерева, что удобно, когда мы заранее не знаем количество кластеров. Многие исследователи уделяют особое внимание влиянию настроения новостей, характерных для конкретной области, на финансовые активы. В исследовании [5] авторы изучают зависимость цен на акции от эмоциональной окраски новостей в конкретном домене. Они рассматривают два значимых рынка – рынок акций и нефти – и оценивают два важных финансовых актива: промышленный индекс Доу-Джонса (DJIA) и нефть марки West Texas Intermediate (WTI). Новости для проведения данного исследования собирались с веб-сайтов, блогов и онлайн баз данных. Вектор временного ряда настроений использовался в формуле модели векторной авторегрессии в качестве признака.

Кроме того, исследователи постоянно изучают возможности повышения эффективности прогнозирования цен на акции с использованием информации о настроении различного рода фоновой информации. В частности, в работе [6] авторы собирают информацию о сообщениях на форуме EastMoney, чтобы составить словарь настроений и рассчитать индексы настроений инвесторов. В статье [7] ученые рассматривают проблему прогнозирования трендов на фондовом рынке с учетом технических индикаторов и настроений текстов социальных сетей. В исследовании [8] также изучается использование настроений инвесторов из социальных сетей, для чего собирается информация из StockTwits, социальной медиаплатформы для инвесторов. Для анализа настроений в данном исследовании используется FinBERT – предварительно обученная языковая модель, предназначенная для анализа настроений финансовых текстов. В качестве задачи рассматривается прогнозирование будущего движения биржевого индекса SPDR S&P 500 Index Exchange Traded Funds. Их результаты показывают, что использование модели FinBERT для анализа настроений дает наилучшие результаты. Авторы статьи [9] строят модель прогнозирования цен на акции на основе сети долговременной памяти, основанной на внимании, используя исторические данные о ценах, технические индикаторы и информацию о настроениях в социальных сетях. Кроме того, они предлагают модель классификации настроений BERT с точной настройкой на определение эмоциональной окраски текста и лексикон настроений для оценки вероятностей принадлежности текстов из социальных сетей к классам позитивно или негативно окрашенных. Согласно полученным результатам, модель BERT с точной настройкой лучше справляется с задачей классификации текстов по настроению, а настроения, рассчитанные с помощью модели BERT, обеспечивают более высокую точность прогнозирования, чем рассчитанные с помощью лексикона настроений.

3. МЕТОДОЛОГИЯ

Цель нашей работы – определить, следует ли разделять поток новостей о крупных публичных компаниях на категории, и какая стратегия обработки новостного потока способствует более высокому качеству прогнозов цен на акции компаний. Результаты исследования помогут улучшить наше представление о зависимости тенденций рынка от различных видов и представлений внешней информации. Кроме того, полученные знания могут быть использованы для определения того, какую информацию следует учитывать при анализе и прогнозировании будущих изменений на рынках публичных компаний.

Схема нашего исследования, представленная на рис. 1, состоит из следующих этапов:

• **Извлечение данных.** Для нашего исследования мы используем набор финансовых новостей, относящихся к 5 крупным публичным компаниям: Apple, Amazon, Google, Netflix и Tesla, а точнее, заголовки новостей и исторические данные о ценах на акции указанных компаний.

• **Тематическая разметка данных.** Мы используем предварительно обученную модель для определения тематической направленности каждой новости в соответствии с ее текущим заголовком. В результате мы получаем несколько тематических групп.

• **Разметка данных по эмоциональной окрашенности.** Для определения настроений новостных заголовков мы используем модель FinBERT, которая была обучена на финансовых текстах и настроена на предсказание меток настроений. Если по заголовку новости понятно, что она позитивная, ей присваивается метка “positive”, если заголовок имеет негативный окрас – метка “negative”, в противном случае новость считается нейтральной и получает метку “neutral”.

• **Передача внешней информации в модель.** Для каждого дня мы рассчитываем отношение количества положительных новостей к общему количеству положительных и отрицательных новостей, вышедших в этот день, и называем эти значения “настроениями”. Временной ряд таких значений определим как временной ряд настроений новостей.

• **Модель прогнозирования цен на акции.** Исторические цены акций с временными рядами настроений, построенными с использованием различных подходов, затем передаются в 4 различные модели: Темпоральная конволюционная сеть (Temporal Convolutional Network), D-Linear,

трансформатор (Transformer) и трансформатор темпорального слияния (Temporal Fusion Transformer). Затем мы получаем прогнозы на N шагов вперед.

В этом подразделе мы рассмотрим значимые части алгоритма разбиения финансовых новостей на тематические группы и определения настроений в новостях.

Для разделения обширной коллекции новостей на тематические группы мы использовали алгоритм классификации. Первым делом мы нашли набор данных с размеченными финансовыми текстами. Это twitter-financial-news-topic [10] – англоязычный набор данных, включающий твиты, связанные с финансовыми темами и новостями. Документы в этом наборе разделены на 20 тем:

- Обновления в аналитике
- ФРС | Центральный банк
- Новости о компании | продукте,
- Казначейские облигации | Корпоративный долг
- Дивиденды
- Доходы
- Энергетика | Нефть
- Финансовые показатели
- Валюты
- Общие новости | Мнения
- Золото | Металлы | Материалы
- IPO
- Слияния и поглощения | Инвестиции
- Макроэкономика
- Рынки
- Политика
- Кадровые изменения
- Комментарии о фондовом рынке
- Динамика изменения цен акций

Затем мы выбрали модель [11], обученную на описанном наборе размеченных данных. Она представляет собой модель на основе BERT (Bidirectional Encoder Representations from Transformers). Первоначально эта модель была обучена на корпоративных отчетах, расшифровках звонков о доходах, отчетах аналитиков, т.е. на текстах финансовой коммуникации. Затем она была обучена на 10 тыс. предложений из аналитических отчетов, которые были предварительно аннотированы метками позитивного, негативного и нейтрального настроений. Завершающим этапом стало обучение модели на ранее описанном наборе данных. Точность модели составила 0.911, что позволило сделать вывод о пригодности данной модели для задач тематической классификации. Наконец, мы применили описанную модель к нашим данным.

На предыдущем этапе мы получили 20 тематических групп финансовых новостей. Далее будет описано, как полученные тематические группы были аннотированы метками настроений. Для маркировки финансовых новостей мы использовали модель finbert-tone [12]. Мы применили эту модель к каждой группе новостей и в результате получили 20 тематических групп финансовых новостей с присвоенными каждой новости метками настроения (положительное, отрицательное, нейтральное).

На данном этапе исследования изучалось изменение качества прогнозирования после добавления внешних данных в модели глубокого обучения. В качестве моделей, которые мы использовали в наших экспериментах, были выбраны темпоральная конволюционная сеть (TCN) [13], D-линейная [14], трансформер [15] и темпоральный трансформер слияния (TFT) [16]. Среди них есть две модели с архитектурой на основе трансформеров. Одна из них – модель трансформатора темпорального слияния. По данным работы [16], модель TFT превосходит большинство существующих моделей Deep Learning для прогнозирования временных рядов. Она имеет архитектуру, основанную на внимании, которая обеспечивает высокопроизводительное прогнозирование на много шагов вперед, изучая временные связи. Основной особенностью, по которой модель TFT выделяется среди других решений, является ее способность эффективно строить представление признаков для каждого типа входных данных и обеспечивать качественное прогнозирование для широкого круга задач. Однако авторы исследования [14] пытаются оспорить успешность такой архитектуры и доказывают, что модель D-Linear превосходит существующие модели на основе трансформаторов. Модель TCN, которую мы также рассматриваем в нашем исследовании, согласно статье [13] и нашему собственному опыту, показывает самую высокую скорость обучения, конкурируя с другими моделями. Преимуществом всех описанных моделей является возможность построения широкогоризонтных прогнозов, что позволяет принимать решения, думая на несколько шагов вперед. Для реализации архитектуры указанных моделей мы использовали библиотеку Darts [17]. В модель прогнозирования передается 15 значений, соответствующих 15 предыдущим дням. Этот параметр называется *длиной входного чанка*. Чтобы проанализировать достаточное количество событий, в качестве такой длины мы взяли около двух недель. Горизонт прогнозирования для всех моделей был установлен равным 3. Остальные параметры моделей принимали значения по умолчанию. Все модели обучались 30 эпох с функцией потерь MSE (средняя квадратичная ошибка).

Общий алгоритм подготовки данных перед их передачей в модель:

• Создание временного ряда из набора данных об исторических ценах. Периодичность временного ряда – рабочие дни

.• Заполнение недостающих значений во временных рядах цен на акции.

• Скалирование значений временного ряда цен на акции в диапазон [0, 1].

• Создание временного ряда из набора данных о настроениях новостей, указав в качестве частоты временного ряда все дни.

• Заполнение недостающих значений временных рядов с настроениями.

• Скалирование временных рядов настроений

• Преобразование набора данных о настроениях во временной ряд с периодичностью, соответствующей рабочим дням.

• Усечение временных рядов цен акций и настроений таким образом, чтобы они охватывали одни и те же периоды.

Как уже было упомянуто в разделе 3.2.1, после этапа тематической разметки мы получили 20 тематических групп. Для наших экспериментов мы выбрали 5 из 20 групп, исходя из общего количества новостей, относящихся к конкретной теме, и степени важности данной темы для нашей задачи.

Мы рассматриваем три различных метода построения временных рядов с использованием внешних данных:

• **Использование только одной тематической группы новостей.** Здесь мы фильтруем поток новостей компании по определенной тематической метке (например, *Stock Commentary*). Временной ряд представляет собой ежедневное соотношение между количеством положительных новостей по выбранной теме и общим количеством положительных и отрицательных новостей по данной теме.

• **Построение тематических векторов.** Алгоритм формирования временного ряда здесь следующий: за каждый день мы берем новости из 5 выбранных тематических групп. Для каждой группы считается отношение между количеством положительных новостей из этой тематической группы и общим количеством положительных и отрицательных новостей, относящихся к данной группе. Таким образом, каждый день описывается вектором из 5 элементов, которые являются подсчитанными соотношениями.

• **Использование всех новостей.** В этом методе мы берем все новости о компании, не разделяя их на группы, и рассчитываем ежедневную долю положительных новостей, как в предыдущих методах.

Все временные ряды настроений и цен акций обрезаются таким образом, чтобы охватить одни и те же временные периоды для одной компании.

В библиотеке Darts есть важный параметр, называемый ковариатами. Ковариаты – это внешние данные, передаваемые в модель для улучшения прогноза. Ковариаты могут относиться к будущему (future covariates), прошлому (past covariates) или являться константами (static covariates). Ковариаты, относящиеся к прошлому, известны только в прошлом, относящиеся к будущему известны только в будущем, а статические ковариаты постоянны во времени. Поскольку ежедневные доли положительных новостей о компании в реальной жизни известны только в прошлом, мы передаем их в модель в качестве параметра *past covariates*.

Для расчета прогнозов на несколько временных периодов вперед мы применили метод исторических прогнозов (historical forecasts). На первом шаге данный алгоритм берет окно от начала временного ряда до временной точки, которая передается в качестве параметра *start*. В качестве параметра *start* мы передаем начало тестовой части временного ряда. Горизонту прогноза задаем значение в 3 шага. Параметр stride равен горизонту прогноза, поэтому метод предсказывает 3 точки, затем конец обучающего набора сдвигается вперед на три временных шага, и метод предсказывает следующие 3 точки. Этот процесс повторяется до конца ряда.

4. РЕЗУЛЬТАТЫ

Данные для нашего финального набора были взяты из архива финансовых новостей [18]. Этот набор данных был собран с сайтов investing.com, bloomber.com, seekingalpha.com, 247wallst.com, zacks.com и cnbc.com. Данные были предварительно обработаны путем удаления изображений, графики, рекламных блоков и знаков препинания. В набор вошли финансовые новости о более 800 публичных компаниях. Общее количество новостей составляет 221 513. После сбора данных мы удалили ненужные поля, чтобы осталась информация только о тикерах компаний, заголовках новостей и датах публикации новостей. Поле Ticker мы использовали для фильтрации заголовков новостей, относящихся только к Apple, Amazon, Google, Netflix и Tesla, для дальнейшей разметки данных.

По результатам тематической классификации можно сделать вывод, что наиболее многочисленными группами являются *Новости компании | продукта*, *Рынки*, *Комментарии о фондовом рынке*, *Изменения цен акций* и *Финансовые показатели*. Чтобы убедиться в разумности разделения новостного потока, мы извлекли из набора данных

Таблица 1. Количество меток сентиментов в каждой тематической группе. Последний столбец содержит процент эмоционально окрашенных новостей относительно размера тематичесой группы

Тема	Количество позитивных новостей	Количество негативных новостей	Количество нейтральных новостей	Позитивные + + Негативные, %
Обновления аналитики	1872	774	4377	38%
ФРС/Центральный банк	384	688	2077	34%
Валюты	742	1133	1672	53%
Дивиденды	409	55	744	38%
Доходы	4909	2108	9578	42%
Энергетика/Нефть	1605	2158	2671	58%
Финансовые показатели	11 363	5662	832	95%
Общие новости	1061	2506	8901	29%
Золото/Металлы/Материалы	691	615	1036	56%
Слияния и поглощения/Инвестиции	1076	318	7330	16%
IPO	103	87	643	23%
Правовое регулирование	225	1740	5391	27%
Макроэкономика	2582	3565	5063	55%
Рынки	7984	7472	9540	62%
Кадровые изменения	109	153	2428	10%
Политика	303	641	3675	20%
Новости о компании/продукте	8893	4275	31 233	30%
Комментарии о фондовом рынке	9650	1634	13 080	46%
Изменения цен акций	8862	5493	5560	72%
Казначейские облигации/корпоративный долг	335	521	1658	34%

финансовых новостей определенные группы, прочитали новости и попытались проанализировать, связаны ли они с конкретной темой и есть ли у них что-то общее. Мы заметили, что группы были определены разумно. Например, заголовок *Акции с самым высоким рейтингом роста для покупки 15 июня* был отнесен к группе *Комментарии о фондовом рынке*. Новости об экономических изменениях в различных странах, например, *Экономика Китая замедляется*, были определены к категории *Макроэкономика*. Группа *Общие новости* включает в себя новости, которые напрямую не связаны с экономикой, но могут косвенно влиять на экономику, повседневную жизнь и сознание людей. Например, *Неконтролируемый рост лесных пожаров в Калифорнии вынуждает закрывать школы*. Такие заголовки, как *Сообщается, что Apple закупает оборудование для производства OLED*, отражающие обновления, связанные с производством популярной продукции, относятся к категории *Новости продукта*. В группе *Рынки* представлены новости, описывающие общее состояние рынков разных стран, например, *Фондовый рынок Бельгии вырос на момент закрытия торгов*.

Для оценки качества работы модели, учитывающей различные тематические группы, мы подсчитывали процентное соотношение суммы положительных и отрицательных меток в общем количестве меток, так как нейтральные новости не представляют особого интереса. Результаты можно увидеть в табл. 1.

Мы начинаем с построения базовых предсказаний модели. Базовые прогнозы цен акций повторяют значения, полученные на предыдущих этапах. Таким образом, мы рассчитываем метрику MAPE для реальных значений и значений со сдвигом на 3 дня назад и считаем это базовым результатом. В результате мы заключили, что работа моделей TCN и Transformer могла бы быть более удовлетворительной, так как метрики, полученные по их прогнозам, не превзошли базовые результаты. В итоге, мы решили рассмотреть только метрики для моделей DLinear и TFT, поскольку их результаты в большинстве случаев превосходили базовые. Рассчитанные метрики для всех

Таблица 2. Тестовые метрики для Apple. Рассматриваемый период 23.11.2012–30.09.2019. Для TFT победителем является подход с подачей одной новостной тематики. Для DLinear предсказания без учета сентиментов новостей являются наилучшими, но лучший запуск демонстрирует, что подача одной новостной группы может дать более высокое качество

Модель	Оценка	Тип внешних данных	MAPE, %	MAE	R2
DLinear	Среднее	Общие новости	2.798	0.022	0.934
		Все настроения	2.837	0.022	0.931
		Без настроений	**2.577**	**0.020**	**0.943**
	Лучший запуск	**Новости о продукте**	**2.327**	**0.018**	**0.949**
TFT	Среднее	**Рынки**	**2.478**	**0.019**	**0.944**
		Все настроения	2.678	0.021	0.937
		Без настроений	4.253	0.033	0.835
	Лучший запуск	**Комментарии о фондовом рынке**	**2.333**	**0.018**	**0.947**
Baseline			2.967		

Таблица 3. Тестовые метрики для Amazon. Рассматриваемый период 12.10.2012–31.01.2020. Для TFT и DLinear подача одной новостной группы является наилучшей опцией

Модель	Оценка	Тип внешних данных	MAPE, %	MAE	R2
Dlinear	Среднее	**Рынки**	**2.256**	**0.019**	**0.791**
		Все настроения	2.295	0.020	0.783
		Без настроений	2.747	0.024	0.733
	Лучший запуск	**Рынки**	**1.786**	**0.015**	**0.869**
TFT	Среднее	**Рынки**	**2.719**	**0.023**	**0.705**
		Все настроения	3.117	0.027	0.630
		Без настроений	4.336	0.038	0.381
	Лучший запуск	**Комментарии о фондовом рынке**	**1.730**	**0.015**	**0.870**
Baseline			2.019		

Таблица 4. Тестовые метрики для Google. Рассматриваемый период 14.08.2012–12.09.2019. Для TFT и DLinear подход с подачей одной новостной группы является наилучшим. Однако в некоторых случаях TFT без сентиментов новостей показывает более высокие результаты

Модель	Оценка	Тип внешних данных	MAPE, %	MAE	R2
Dlinear	Среднее	**Макроэкономика**	**2.477**	**0.021**	**0.746**
		Все настроения	2.503	0.022	0.740
		Без настроений	2.591	0.022	0.720
	Лучший запуск	**Макроэкономика**	**2.337**	**0.020**	**0.768**
TFT	Среднее	**Общие новости**	**3.004**	**0.026**	**0.663**
		Все настроения	3.050	0.026	0.653
		Без настроений	3.238	0.028	0.612
	Лучший запуск	**Без настроений**	**2.286**	**0.020**	**0.788**
Baseline			2.838		

Таблица 5. Тестовые метрики для Netflix. Рассматриваемый период 24.04.2013–18.12.2019. Для TFT победителем является подход с подачей одной новостной группы, а для Dlinear лучшие результаты достигаются без учета сентиментов новостей

Модель	Оценка	Тип внешних данных	MAPE, %	MAE	R2
Dlinear	Среднее	Общие новости	2.946	0.023	0.886
		Все настроения	3.065	0.023	0.879
		Без настроений	**2.882**	**0.022**	**0.891**
	Лучший запуск	**Без настроений**	**2.619**	**0.020**	**0.908**
TFT	Среднее	**Новости о продукте**	**3.066**	**0.023**	**0.884**
		Все настроения	3.684	0.028	0.835
		Без настроений	3.509	0.027	0.850
	Лучший запуск	**Комментарии о фондовом рынке**	**2.512**	**0.019**	**0.919**
Baseline			3.152		

Таблица 6. Тестовые метрики для Tesla. Рассматриваемый период 04.04.2014–09.12.2019. Для TFT и DLinear подход с одной новостной группой является наилучшим. Лучший запуск для TFT был при учете вектора настроений

Модель	Eval	Тип внешних данных	MAPE, %	MAE	R2
Dlinear	Среднее	**Макроэкономика**	**8.039**	**0.036**	**0.926**
		Все настроения	8.088	0.037	0.921
		Без настроений	8.130	0.037	0.920
	Лучший запуск	**Макроэкономика**	**7.423**	**0.034**	**0.930**
TFT	Среднее	**Общие новости**	**8.597**	**0.039**	**0.917**
		Все настроения	9.595	0.042	0.903
		Без настроений	11.774	0.049	0.880
	Лучший запуск	**Тематические векторы**	**7.421**	**0.035**	**0.928**
Baseline			9.565		

пяти компаний с различными стратегиями учета новостных настроений приведены в табл. 2, 3, 4, 5 и 6. Визуализация различий в прогнозировании, обусловленных разными способами учета новостных настроений, представлена на рис. 2.

Согласно табл. 2–6, для компании Apple оптимальным является подход с использованием эмоционального наполнения одной тематической группы. То же самое справедливо для Amazon, Google и Tesla. Для Netflix, как показывают мет-

Рис. 2. Реальные цены на акции Amazon и предсказанные цены, полученные с помощью TFT модели с разными способами учета настроения новостей. Лучший запуск.

рики, модель DLinear лучше использовать без настроений, а TFT по-прежнему лучше работает с учетом одной тематической группы. С точки зрения моделей, подход с учетом одной тематики также является лучшим в большинстве экспериментов. В целом полученные результаты подтверждают, что в большинстве случаев показывать нашей модели глубокого обучения только одну тематическую группу новостей лучше, чем рассматривать весь поток новостей или обучать модель без учета новостей. Кроме того, мы определили, какие тематические группы с большей вероятностью улучшают прогнозы. Мы заметили, что эффективнее всего работают группы *Макроэкономика* и *Рынки*, на втором месте – *Комментарии о фондовом рынке*, затем – *Общие новости* и *Новости о компании и продукте*. Здесь можно провести параллели с табл. 1 и сделать вывод, что чем больше процент позитивных и негативных новостей в данной тематической группе, тем больше вероятность того, что эта группа поможет повысить эффективность работы модели. Чем выше эмоциональная насыщенность, тем сильнее влияние на модель.

5. ЗАКЛЮЧЕНИЕ

В данном исследовании мы показали, что в задаче прогнозирования цен на акции полезно проводить тематическую классификацию потока новостей. В большинстве случаев оптимальной стратегией являются разбиение новостного потока на тематические группы и передача в модель настроений из одной конкретной тематической группы, а не учет настроений из всего потока новостей. Более того, мы обнаружили, что улучшение прогноза за счет настроений из определенной тематической группы новостей связано с эмоциональной насыщенностью этой группы. В целом наше исследование способствует получению более глубокого представления о рынке, его процессах и поведении. Данное исследование может оказаться полезным для специалистов, занимающихся анализом рынков, специалистов по изучению данных, а также людей, заинтересованных в более глубоком понимании рыночных тенденций.

СПИСОК ЛИТЕРАТУРЫ

1. *Budennyy S., Kazakov A., Kovtun E., Zhukov L.* New drugs and stock market: a machine learning framework for predicting pharma market reaction to clinical trial announcements. Scientific Reports. 2023. V. 13. № 1. P. 12817.
2. *Nikou M., Mansourfar G., Bagherzadeh J.* Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting, Finance and Management. 2019. V. 26. № 4. P. 164–174.
3. *Khan W., Ghazanfar M.A., Azam M.A., Karami A., Alyoubi K.H., Alfakeeh A.S.* Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing. 2020. P. 1–24.
4. *Patel V., Patel A. C* Clustering news articles for topic detection. 2018.
5. *Kelly S., Ahmad K.* Estimating the impact of domain-specific news sentiment on financial assets. Knowledge-Based Systems. 2018. V. 150. P. 116–126.
6. *Mu G., Gao N., Wang Y., Dai L.* A stock price prediction model based on investor sentiment and optimized deep learning. IEEE Access PP, 2023. https://doi.org/10.1109/ACCESS.2023.3278790
7. *Wang Z., Hu Z., Li F., Ho S.-B.L.* Learning-based stock market trending analysis by incorporating social media sentiment analysis. 2021. https://api. semanticscholar.org/CorpusID:235526511
8. *Liu J.-X., Leu J.-S., Holst S.* Stock price movement prediction based on stock- twits investor sentiment using finbert and ensemble svm. PeerJ Computer Science. 2023. V. 9. P. 1403. https://doi.org/10.7717/peerj-cs.1403
9. *Ji Z., W P., Ling C., Zhu P.* Exploring the impact of investorвЂ™s sentiment tendency in varying input window length for stock price prediction. Multimedia Tools and Applications. 2023. V. 82. P. 1–35. https://doi.org/10.1007/s11042-023-14587-8
10. zeroshot/twitter-financial-news-topic. https://huggingface.co/datasets/zeroshot/ twitter-financial-news-topic. Accessed: 2023-05-17
11. finbert-tone-finetuned-finance-topic-classification. https://huggingface.co/nickmuchi/finbert-tone-finetuned-finance-topic-classification. Accessed:2023-05-17
12. finbert-tone. https://huggingface.co/yiyanghkust/finbert-tone. Accessed: 2023-05-17
13. *Lea C., Flynn M., Vidal R., Reiter A., Hager G.* Temporal convolutional networks for action segmentation and detection. 2016.
14. *Zeng A., Chen M., Zhang L., Xu Q.* Are Transformers Effective for Time Series Forecasting? https://doi.org/10.48550/arXiv.2205.13504
15. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.* Attention Is All You Need (2023).
16. *Lim B., Arik S., Loeff N., Pfister T.* Temporal fusion transformers for inter-pretable multi-horizon time series forecasting. International Journal of Forecasting. 2021. V. 37. https://doi.org/10.1016/j.ijforecast.2021.03.012
17. Temporal Fusion Transformer. https://unit8-co.github.io/darts/generated_api/darts.models.forecasting.tft_model.html Accessed: 2023-05-17
18. Historical financial news archive. https://www.kaggle.com/gennadiyr/ us-equities-news-data/tasks. Accessed: 2023-05-17

DO WE BENEFIT FROM THE CATEGORIZATION OF THE NEWS FLOW IN THE STOCK PRICE PREDICTION PROBLEM?

T. D. Kulikova[a], E. Y. Kovtun[b], and S. A. Budennyy[c]

[a]*National Research University Higher School of Economics, Moscow, Russian Federation*

[b]*Sber AI Lab, Moscow, Russian Federation*

[c]*Artificial Intelligence Research Institute (AIRI), Moscow, Russian Federation*

Presented by Academician of the RAS A.A. Shananin

The power of machine learning is widely leveraged in the task of company stock price prediction. It is essential to incorporate historical stock prices and relevant external world information for constructing a more accurate predictive model. The sentiments of the financial news connected with the company can become such valuable knowledge. However, financial news has different topics, such as Macro, Markets, or Product news. The adoption of such categorization is usually out of scope in a market research. In this work, we aim to close this gap and explore the effect of capturing the news topic differentiation in the stock price prediction problem. Initially, we classify the financial news stream into 20 pre-defined topics with the pre-trained model. Then, we get sentiments and explore the topic of news group sentiment labeling. Moreover, we conduct the experiments with the several well-proved models for time series forecasting, including the Temporal Convolutional Network (TCN), the D-Linear, the Transformer, and the Temporal Fusion Transformer (TFT). In the results of our research, utilizing the information from separate topic groups contributes to a better performance of deep learning models compared to the approach when we consider all news sentiments without any division.

Keywords: Financial news, Stock market, BERT, Topic classification, Sentiment analysis, Time-series forecasting, Deep learning, External data