УДК 004.852:51-77
DOI 10.17513/snt.40620

АНАЛИЗ СЕНТИМЕНТА ФИНАНСОВЫХ НОВОСТНЫХ ПУБЛИКАЦИЙ НА РУССКОМ ЯЗЫКЕ

Саяпин А.В. ORCID ID 0000-0003-3027-2915,
Ямашкин С.А. ORCID ID 0000-0002-7574-0981

Федеральное государственное бюджетное образовательное учреждение высшего образования «Мордовский государственный университет имени Н.П. Огарёва», Саранск, Российская Федерация, e-mail: mrzxfy@gmail.com

В данной статье рассматривается возможность применения моделей обработки естественного языка для задачи классификации финансовых новостей по содержащемуся в них сентименту, полученному с помощью автоматической разметки на основе направления изменения цены акции компании после публикации новости с упоминанием компании. Для автоматической разметки данных были предложены несколько подходов, учитывающих временные данные разной частотности. Целью исследования является разработка и оценка эффективности моделей обработки естественного языка для задачи анализа сентимента финансовых новостей на русском языке с использованием автоматической разметки сентимента на основе изменения цены в качестве прокси сентимента. Рассмотрены и предложены подходы к автоматической разметке сентимента на основе дневного и почасового изменения цен для оценки сентимента. Для анализа был собран набор новостных данных по 7 крупным российским компаниям, данные были автоматические размечены, и на них были обучены модели классификации сентимента. Полученные результаты свидетельствуют о том, что модель, обученная на распознавание сентимента, способна с удовлетворительным качеством определять содержащийся в тексте новости финансовый сентимент. Предложены направления дальнейшего исследования в данной области.

Ключевые слова: анализ сентиментов, обработка естественного языка, классификация текстов, финансовые новости, машинное обучение, фондовый рынок, большие языковые модели

SENTIMENT ANALYSIS OF FINANCIAL NEWS ARTICLES IN RUSSIAN LANGUAGE

Sayapin A.V. ORCID ID 0000-0003-3027-2915,
Yamashkin S.A. ORCID ID 0000-0002-7574-0981

Federal State Budgetary Educational Institution of Higher Education "Ogarev Mordovian State University", Saransk, Russian Federation, e-mail: mrzxfy@gmail.com

This paper examines the possibility of applying natural language processing models to the task of classifying financial news by the sentiment they contain, obtained using automatic labeling based on the direction of a company's share price change after the publication of the news item mentioning the company. Several approaches were proposed for automatic data labeling, taking into account time data of different frequencies. The aim of the study is to develop and evaluate the effectiveness of natural language processing models for the task of analyzing the sentiment of financial news in Russian using automatic sentiment labeling based on price change as a sentiment proxy. Approaches to automatic sentiment labeling based on daily and hourly price changes for sentiment assessment are considered and proposed. For the analysis, a set of news data on 7 large Russian companies was collected, the data was automatically labeled, and sentiment classification models were trained on them. The results indicate that the model trained to analyze sentiment is capable of determining the financial sentiment contained in the news text with satisfactory quality. Directions for further research in this area are suggested.

Keywords: sentiment analysis, natural language processing, text classification, financial news, machine learning, stock market, Large language models

Введение

На финансовых рынках участники используют различные источники информации для анализа компаний и их ценных бумаг и принятия решений на рынке. Одним из видов источников информации являются новости о компании, которые могут нести в себе как позитивную, так и негативную информацию. Содержащаяся в новостях информация может незамедлительно влиять на принимаемые инвесторами и игроками на рынке решения. В результате на краткосрочном периоде может отвергаться гипотеза эффективного рынка [1], согласно которой вся доступная информация сразу же отражается в стоимости актива, и все агенты действуют рационально. Применение методов анализа естественного языка может автоматизировать процесс ознакомления с новостной информацией и классифицировать новости как позитивные и негативные.

Анализ сентиментов – это подраздел обработки естественных языков, направленный на решение задачи классификации текстов на основе анализа содержащихся в тексте сентиментов, тональности [2]. Обычно ана-

лиз сентиментов принимает вид задачи бинарной или многоклассовой классификации.

Классические работы [3] по оценке влияния новостного фона на финансовые и биржевые показатели компаний демонстрируют, что новостная информация может использоваться для предсказания таких финансовых показателей, как прибыль, а также для прогнозирования цены акций после выхода новости. Данные работы измеряют тональность новости на основе словарей, по которым слова классифицируются как негативные или позитивные, а затем происходит расчёт агрегированного показателя для новости на основе полученных классов слов.

Применение моделей машинного обучения для анализа сентиментов финансовых новостей стало популярной темой для научных исследований за последние 10 лет. Большинство работ в данной области анализируют сентимент на основе данных, полученных из новостей [4] или социальных сетей [5]. В работах для финансовых текстов на английском языке большие языковые модели, дополнительно обученные на текстах финансовой тематики, такие как FinBERT [6], показывают наилучшие результаты по сравнению с другими моделями [7]. Также исследуется [8] дальнейшее применение полученных сентиментов новостей для задач предсказания различных мер акций, таких как цена и волатильность.

Помимо обучения моделей машинного обучения выполнять отдельные финансовые задачи, для финансовой области была разработана Большая языковая модель BloombergGPT [9], которая позволяет решать разнообразные задачи, связанные с анализом текстовой информации в финансах, такие как анализ сентиментов, распознавание именованных сущностей, ответы на вопросы.

Общедоступные наборы данных с размеченным сентиментом для текстов на русском языке предназначены для текстов общей направленности и были собраны на основе данных из социальных сетей [10]. В работах по анализу сентимента для текстов финансовой области на русском языке [11] применяется ручная разметка данных, а сами наборы данных не публикуются, что затрудняет проведение других исследований.

Возможным решением проблемы отсутствия качественных наборов данных для задачи определения финансового сентимента является использование подхода Cross-Lingual Transfer Learning для адаптации финансовых моделей для английского языка, таких как FinBERT и BloombergGPT, для русскоязычных текстов.

Целью исследования является разработка алгоритма для автоматической разметки сентимента на основе динамики цен в качестве прокси сентимента и оценка эффективности моделей обработки естественного языка, обученных для решения задачи анализа сентимента финансовых новостей на русском языке с использованием набора данных на основе автоматической разметки.

Материалы и методы исследования

Предыдущие работы по анализу сентимента финансовых публикаций используют ручную экспертную или пользовательскую разметку новостей для обучения моделей классификации. Недостатки данного подхода заключаются в больших временных и, если используется заказ разметки через краудсорсинговые сервисы, денежных затратах. Помимо этого, данный подход не гарантирует, что новость, размеченная как позитивная/негативная, привела к соответствующему изменению цен на акции компании после её публикации.

Для устранения приведенных выше недостатков в данном исследовании предлагается подход, основанный на автоматической разметке новостей с учётом изменения цены акций после её публикации. Разработаны 2 версии алгоритма разметки на основе дневных торговых данных и высокочастотных (час и меньше).

Сентимент на основе дневных данных котировок с учётом времени закрытия и открытия торгового дня оценивается на основе разницы дневных цен открытия или закрытия в зависимости от времени публикации новости в предыдущий день. Недостатком данного подхода является использование длительного периода для оценки эффекта новости, что может привести к зашумленности разметки, поскольку в этот период на цену акций могут влиять и другие факторы, не учитываемые алгоритмом.

Алгоритм разметки с использованием дневных данных представлен на рисунке 1.

Высокочастотный метод, который оценивает изменения цены через n коротких (час и менее) периодов после публикации новости, может позволить более точно оценить эффект от публикации новости за счёт подбора и использования меньшего временного периода для определения сентимента на основе динамики цен. Для расчёта сентимента используются следующие формулы:

$$R = (p_{t+n} - p_t) / p_t , \quad (1)$$

$$\text{Sentiment} = \text{sgn } R, \quad (2)$$

где p – цена закрытия, t – начальный период времени, n – число периодов, за которые определяется изменение цены, sgn – функция знака, R – доходность за период времени n.

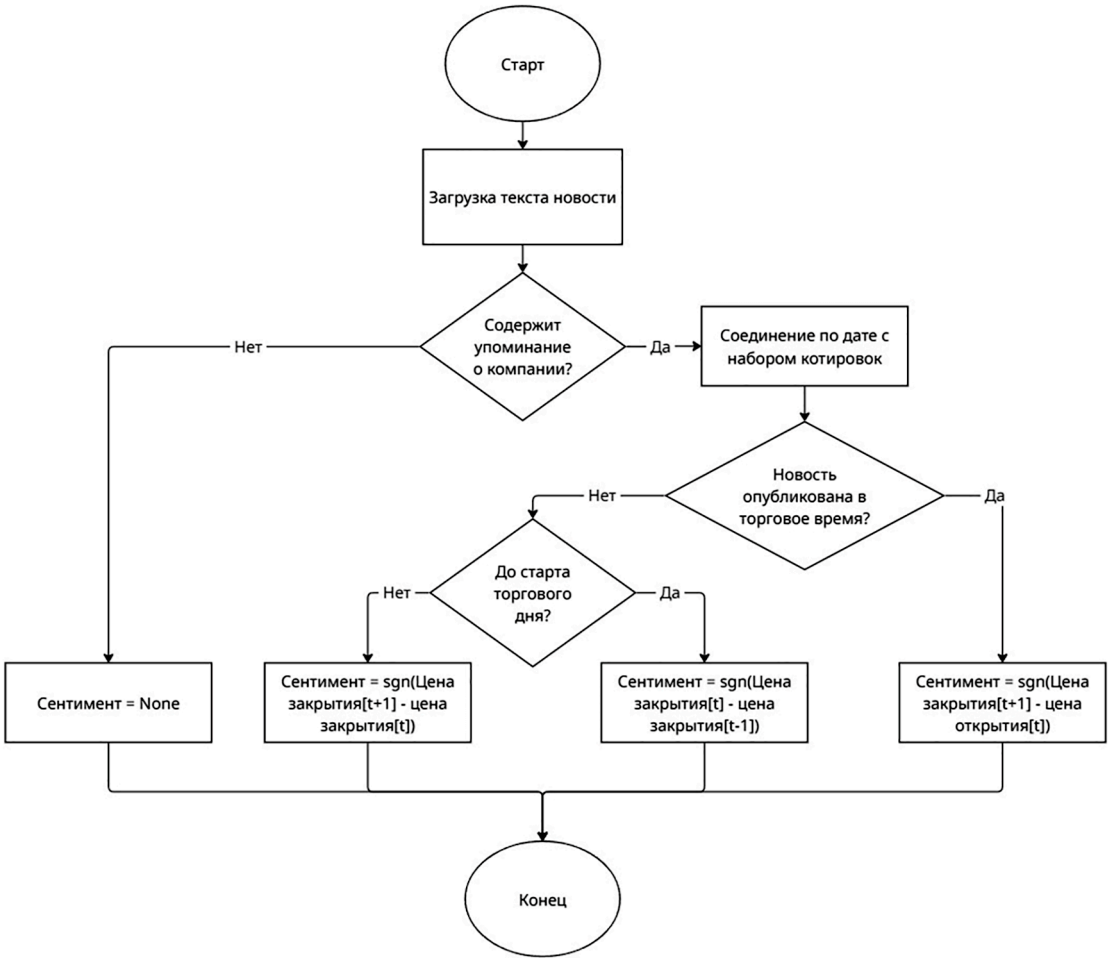

Рис. 1. Алгоритм разметки сентимента с помощью дневных данных с учётом времени закрытия и открытия торгового дня
Источник: составлено авторами

Новости, опубликованные в нерабочие часы и дни (праздники, выходные дни) биржи, используют цены за последний до выхода новости рабочий период биржи как цены начального периода времени *t*.

Для синхронизации новостных публикаций и рыночных котировок используются временные точки с одинаковым часовым поясом (UTC-3, Москва). Задержка публикации в выбранном источнике новостных данных по сравнению с другими новостными ресурсами не учитывается. Задержки в получении данных при выполнении запроса к серверам новостных ресурсов и брокерских компаний не учитываются в данном исследовании.

Для определения нейтрального класса в обоих подходах используется порог для значений доходности. Если значение доходности ниже порога, то класс новости определяется как нейтральный, если больше, то осуществляется разметка по алгоритмам, приведенным выше.

Недостатком обоих подходов является дилемма при необходимости выбирать между порогом для определения сентимента, уменьшение которого может ухудшить качество разметки, и размером итоговой выборки, поскольку увеличение порога сократит число новостей, для которых будет определен сентимент.

Для анализа были взяты новостные данные о компаниях Сбербанк, ВТБ, Яндекс, Газпром, Роснефть, МТС, Аэрофлот с сайта новостного издания Лента.ру за 2013–2021 гг. За тот же период были собраны данные дневных и часовых показателей цены обыкновенных акций данных организаций.

В рамках данного исследования оценка сентимента с учётом новостей о конкурентах компаний, новостей о макроэкономической ситуации на российском рынке не проводится, используются только новости о выбранных компаниях.

Рис. 2. Распределение дневной цены закрытия для обыкновенных акций компаний
Источник: составлено авторами

Распределение цены обыкновенных акций компаний Газпром, Роснефть, МТС, Аэрофлот, Яндекс, Сбербанк и ВТБ за выбранный временной период представлено на рисунке 2.

На данном временном отрезке видно, что цена закрытия акций достаточно волатильна.

Одним из базовых методов для обработки текстов на естественном языке является построение признаков TF-IDF [12] и их применение в моделях машинного обучения. Признаки при применении данного подхода формируются на основе расчёта следующей формулы для каждого слова в текстовом наборе данных:

$$TF - IDF_i = \frac{\text{количество вхождений слова i в документ}}{\text{количество слов в документе}} \times \log\left(\frac{N}{1+df_i}\right), \quad (3)$$

где N – количество документов в наборе данных, df – количество документов, в которых есть слово i.

В качестве моделей классификации текста в данном исследовании используется алгоритм RandomForest [13], градиентный бустинг (реализация XGBoost [14]), предобученные большие языковые модели на основе инфраструктуры Transformer [15].

Для анализа качества классификации новостей моделью в исследовании используются метрики:

Accuracy – доля правильно спрогнозированных моделью классов по отношению к общему числу прогнозов

$$Accuracy = (TP + TN) / (TP + FP + TN + FN), \quad (4)$$

где TP (True positive) – количество истинно положительных прогнозов, FP (False positive) – количество ложно положительных прогнозов, TN (True negative) – количество истинно отрицательных прогнозов, FN (False negative) – количество ложно отрицательных прогнозов.

Precision – точность, оценивает долю верно спрогнозированных ответов для положительного класса.

$$Precision = TP / (TP + FP), \quad (5)$$

Recall – полнота, оценивает, какую долю объектов положительного класса выделила модель.

$$Recall = TP / (TP + FN), \quad (6)$$

F1-мера представляет собой гармоническое среднее точности и полноты, что позволяет снижать значение метрики в случае, когда одна из метрик-аргументов значительно ниже другой.

$$F1 = 2 \cdot (precision \cdot recall) / (recall + precision). \quad (7)$$

ROC-AUC рассчитывается как площадь под ROC-кривой, данная метрика позволяет оценить качество предсказаний модели с учётом всех возможных порогов для определения класса.

Результаты исследования и их обсуждение

Для обучения моделей на основе признаков TF-IDF производилась обработка текстовых данных путем удаления стоп-слов, пунктуации и лемматизации. В качестве алгоритмов, которые обучались на признаках TF-IDF, были использованы RandomForest и градиентный бустинг (XGBoost).

Обучение больших языковых моделей проводилось без предварительной обработки текста, поскольку данный класс моделей способен работать с сырыми текстовыми данными. Для экспериментов использовались языковые модели ruSbert и LABSE. Для данных моделей был добавлен выходной слой-классификатор, обучение производилось только для данного слоя, остальные предварительно обученные веса модели были заморожены.

Разделение набора данных на обучающую и тестовую выборки осуществлялось с помощью подхода out-of-time. В обучающую выборку вошли данные за 01/01/2013–30/08/2021, в тестовую выборку за 01/09/2021–31/12/2021 год. Гиперпараметры для алгоритмов подбирались методом кросс-валидации на обучающей выборке. Параметры TF-IDF формировались на основе данных в обучающей выборке, чтобы избежать утечки тестовых данных при формировании лексикона TF-IDF.

Оценка качества моделей осуществлялась на данных, размеченных с использованием порогов = [0.01, 0.02], высокочастотный подход был оценен только при использовании порога 0.01. Выбор данных порогов основан на абсолютных значениях среднедневной волатильности на основе цен закрытия акций анализируемых компаний, значение 0.01 близко к медианному значению волатильности, 0.02 близко к значению 0.75 квантиля для дневного периода времени. Для часовых данных порог 0.01 соответствует значению 0.75 квантиля. Пороги используются в данном исследовании с целью удаления из набора данных тех новостей, которые не приводят к существенному изменению цены акций после их публикации.

Для оценки моделей применяются 2 подхода к классификации – с использованием нейтрального класса и без. Результаты с использования нейтрального класса представлены в таблице 1.

Метрики Precision, Recall, F1 оценивались взвешенным усреднением по классам, ROC-AUC оценивался подходом one-vs-rest. Усредненные значения метрик при решении задачи трёхклассовой классификации с нейтральным классом показывают худшие значения по сравнению с бинарной классификацией, также из-за дисбаланса классов из-за большего количества наблюдений с нейтральным классом модель Sbert переобучается на прогнозирование нейтрального класса.

Нейтральный класс, к которому новости определялись при отклонениях цены ниже порогового значения, был исключен из выборок из-за отмеченных выше результатов и не использовался при обучении моделей.

Таблица 1

Сравнение метрик для задачи бинарной и многоклассовой классификации

Алгоритм	Период	Классы	Порог	Accuracy	Precision	Recall	F1	ROC-AUC
TF-IDF+ Random Forest	Дневной	2	0.01	0.53	0.53	0.75	0.62	0.52
TF-IDF + XGBoost	Дневной	2	0.01	0.53	0.53	0.74	0.62	0.53
LABSE	Дневной	2	0.01	0.55	**0.56**	0.82	0.66	**0.55**
Sbert	Дневной	2	0.01	0.53	0.53	**0.96**	**0.68**	0.52
TF-IDF+ Random Forest	Дневной	3	0.01	0.34	0.11	0.34	0.17	0.52
TF-IDF + XGBoost	Дневной	3	0.01	0.33	0.30	0.33	0.24	0.51
LABSE	Дневной	3	0.01	0.34	0.42	0.34	0.18	0.50
Sbert	Дневной	3	0.01	**0.70**	0.48	0.69	0.57	0.54

Источник: составлено авторами.

Таблица 2

Статистика количества наблюдений в выборках

Выборка	Порог сентимента	Период данных	Количество негативных	Количество позитивных
Обучающая	0.01	Дневной	1423	1561
Обучающая	0.01	Высокочастотный	1234	1510
Обучающая	0.02	Дневной	639	789
Тестовая	0.01	Дневной	570	737
Тестовая	0.01	Высокочастотный	259	309
Тестовая	0.02	Дневной	224	262

Источник: составлено авторами.

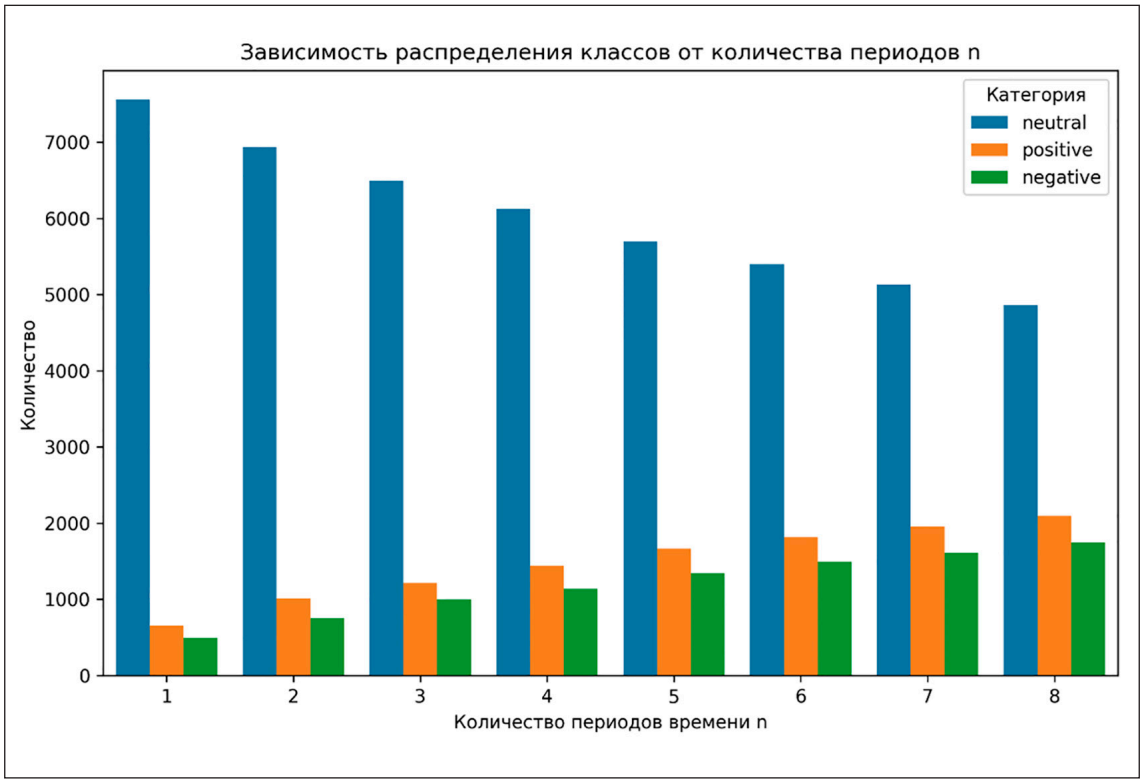

Рис. 3. Распределение классов в зависимости от количества периодов n
Источник: составлено авторами

Набор данных с ценами акций и код для обработки данных, разметки сентимента, обучения и анализа качества моделей, сравнения торговых стратегий доступны по адресу https://github.com/avsayapin/sentiment_analysis_russian_finance. Набор новостных данных доступен по адресу https://www.kaggle.com/datasets/spielmeister/economical-news-in-russian-lenta-ru2013-2019. Статистика по классам (позитивный=1, негативный=0) сентимента, полученная после разметки, представлена в таблице 2.

Распределение классов новостей на разных периодах и порогах близко к равномерному, дисбаланса классов нет. Распределения на тестовой и обучающей выборках схожи.

Для различных n в высокочастотном подходе оценивалось распределение классов в полученном наборе данных и значение метрик для модели RandomForest на основе признаков TF-IDF, что наглядно представлено на рисунке 3.

С увеличением числа периодов в высокочастотном методе разметки растёт и количество классов с позитивной или негативной разметкой с динамикой цены выше заданного порога (рис. 4).

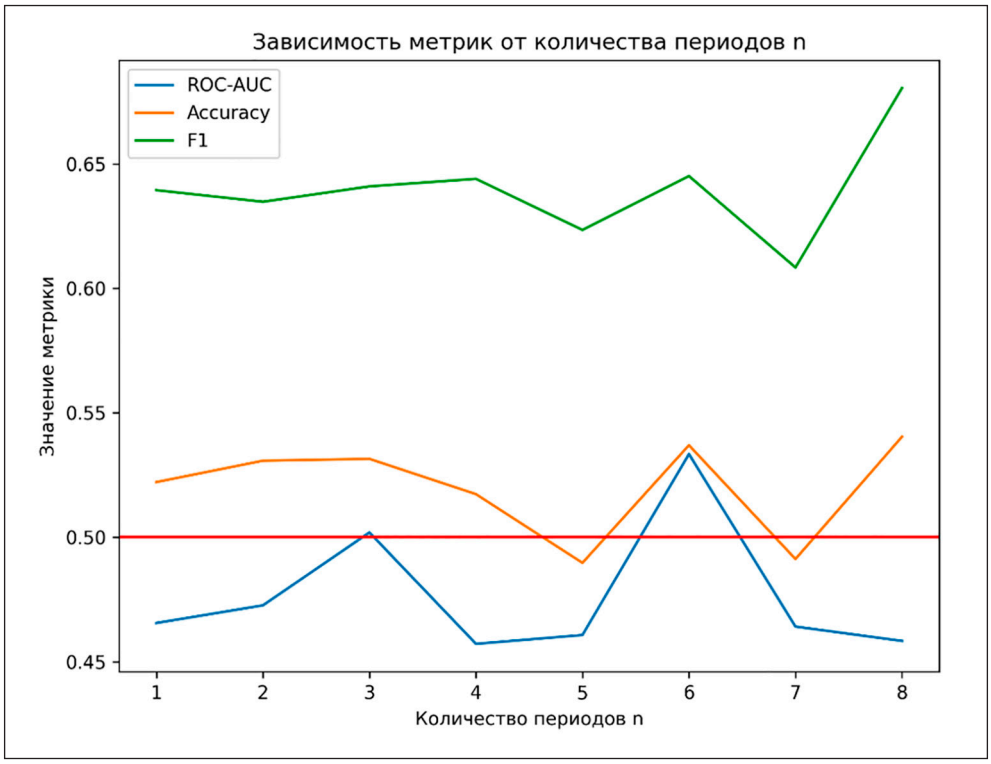

Рис. 4. График метрик качества модели в зависимости от количества периодов n
Источник: составлено авторами

Таблица 3

Метрики качества классификации на тестовой выборке

Алгоритм	Период	Порог	Accuracy	Precision	Recall	F1	ROC-AUC
TF-IDF+ RandomForest	Дневной	0.02	0.53	0.54	0.98	0.69	**0.57**
TF-IDF + XGBoost	Дневной	0.02	0.54	0.54	0.90	0.68	0.54
LABSE	Дневной	0.02	**0.56**	**0.56**	0.79	0.66	0.54
Sbert	Дневной	0.02	0.54	0.54	0.99	**0.7**	0.52
TF-IDF+ RandomForest	Дневной	0.01	0.53	0.53	0.75	0.62	0.52
TF-IDF + XGBoost	Дневной	0.01	0.53	0.53	0.74	0.62	0.53
LABSE	Дневной	0.01	0.55	0.56	0.82	0.66	0.55
Sbert	Дневной	0.01	0.53	0.53	0.96	0.68	0.52
TF-IDF+ RandomForest	Высокочастотный	0.01	0.53	0.55	0.75	0.63	0.50
TF-IDF + XGBoost	Высокочастотный	0.01	0.51	0.54	0.72	0.61	0.51
LABSE	Высокочастотный	0.01	0.51	0.54	0.76	0.63	0.52
Sbert	Высокочастотный	0.01	0.52	0.54	0.73	0.62	0.51

Источник: составлено авторами.

Модель показывает значения метрик выше 0.5 только для периодов 3 и 6, наилучшие же значения по всем трём метрикам получаются при использовании количества периодов, равного 6. Возможно, такое поведение можно объяснить тем, что по прошествии 6 периодов формируется наиболее точная интерпретация новости у инвесторов, что позволяет получить лучшую разметку и в итоге модель. В данном исследовании для дальнейших экспериментов применяется количество периодов, равное 6.

Результаты оценки качества моделей на тестовой выборке с помощью метрик классификации Accuracy, Precision, Recall, ROC_AUC, F1-мера представлены в таблице 3.

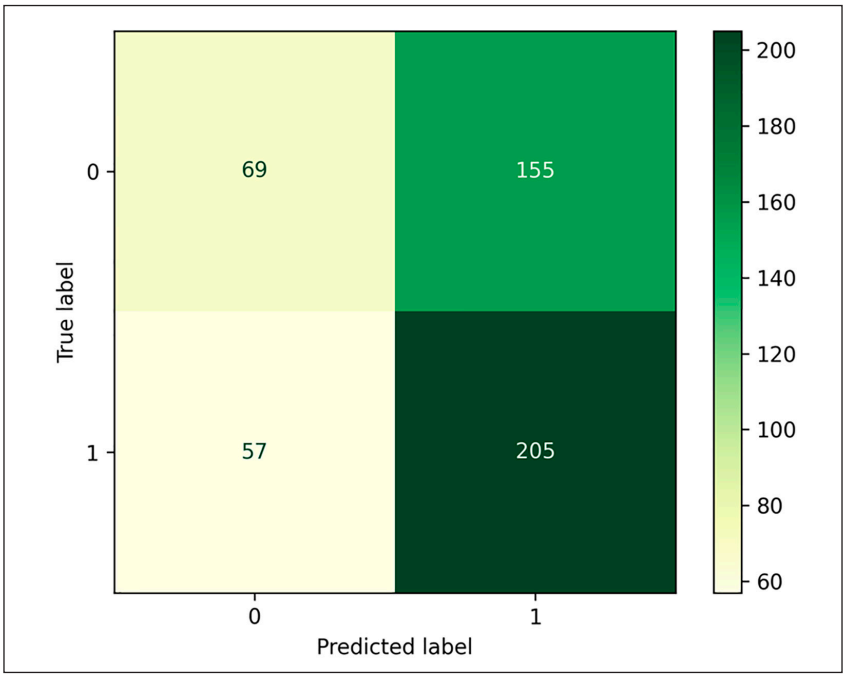

Рис. 5. Матрица ошибок модели LABSE на тестовой выборке
Источник: составлено авторами

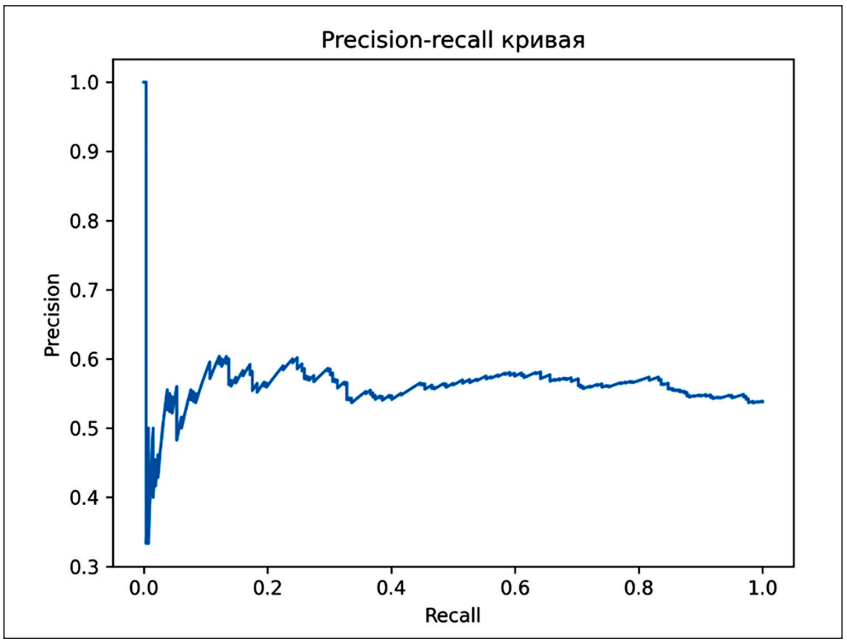

Рис. 6. PR-кривая на тестовых данных для модели LABSE
Источник: составлено авторами

Из результатов видно, что модели демонстрируют высокие значения по метрике Recall и низкие по Precision, подобное поведение может свидетельствовать о том, что модель почти не предсказывает негативный класс при стандартном пороге классификации – 0.5. ROC-AUC показывает наилучшие значения на моделях с использованием признаков TF-IDF.

Для наилучшей модели также была проанализирована матрица ошибок на дневных данных с использованием порога 0.02. Матрица наглядно представлена на рисунке 5.

Модель допускает достаточно большое количество ошибок на участке False Positive (предсказывает наблюдения, истинно относящиеся к классу 0, как наблюдения класса 1), что соответствует большому значению метрики Recall. Подбор порога отсечения, подбор весов для классов модели и расширение объема обучающих данных могут решить данную проблему в дальнейших исследованиях.

Дополнительно был произведен подбор порога отсечения для лучшей модели с целью найти такой порог, для которого значения precision и recall будут, для этого была построена PR-кривая, представленная на рисунке 6.

Порог отсечения подбирался через максимизацию F1-оценки, максимальное значение метрики было достигнуто при пороге 0.49. Показатели accuracy, precision, recall и f1 после изменения порога увеличились до 0.57, 0.57, 0.82, 0.67, что является небольшим улучшением по сравнению со стандартным порогом 0.5.

Для оценки практической значимости использования моделей анализа сентимента произведен анализ торговой стратегии на основе результатов классификации новостей с помощью модели LABSE на дневных торговых данных. Стратегия заключалась в покупке акций компаний и добавлении их в портфель, если новостной фон за день был положительный (класс 1) и акции компании уже не в портфеле, и продаже акций, если новостной фон отрицательный (класс 0) и акции находятся в портфеле. Поскольку в один день могут выйти несколько новостей о компании, то оценка новостного фона осуществлялась на основе расчёта средних значений предсказанных классов новостей по компании за день, порог отсечения для определения класса дневного новостного фона был подобран путем максимизации прибыльности стратегии и составил 0.56.

Сравнение стратегии осуществлялось с простой стратегией Buy and hold, при которой мы покупаем акции в начальный период времени и в дальнейшем не осуществляем с ними никаких действий. Поскольку стоимость акций существенно отличается, то для сравнения стратегий допускается предположение о возможности равномерного распределения количества акций в портфеле. Также для дополнительного сравнения используется индекс Мосбиржи, в который входят акции крупнейших компаний Российской Федерации. Сравнение осуществлялось на тестовой выборке на временном периоде 01.09.2021-31.12.2021. При расчете финансовых результатов по стратегии учитывались транзакционные комиссии в размере 0,05% от размера сделки.

Рис. 7. Сравнение торговых стратегий на тестовом периоде
Источник: составлено авторами

Таблица 4

Метрики эффективности торговых стратегий

Стратегия	Годовая доходность	Волатильность	Shapre ratio	Max drawdown	Turnover
Buy and hold	-2.42%	1.59	-1.59	-81.6%	326%
Новости	-1.23%	0.56	-1.35	-48.7%	0
Индекс Мосбиржи	-0.14%	0.2	-1.07	-15.6%	0

Источник: составлено авторами.

В данном исследовании используется допущение, что мы всегда можем осуществить лимитную заявку по цене конца/начала периода времени t и избежать проскальзывания. Результаты сравнения торговых стратегий представлены на рисунке 7.

Можно заметить, что на первоначальных периодах времени стратегия Buy and hold показывает лучшие результаты, чем стратегия на основе новостей, но в период падения цен на акции компании в ноябре-декабре 2021 года использование стратегии с классификацией новостей позволяет получить лучшую доходность за счёт оперативной продажи акции в случае негативного новостного фона и покупки при позитивном новостном фоне. В таблице 4 представлены основные метрики стратегий, для расчёта индекса Sharpe использовалась средняя ключевая ставка за данный период – 8%.

На основе полученных метрик можно заметить, что новостная стратегия показывает себя лучше, чем стратегия Buy and hold на основе тех же акций, применение модели анализа сентимента позволяет быстро реагировать на новостной фон и принимать решения на рынке. Все три стратегии показывают отрицательные значения по индексу Sharpe из-за предкризисного периода конца 2021 года. Индекс Мосбиржи показывает лучшие результаты за счёт диверсификации портфеля из-за расширенного состава акций.

Среднее время нахождения позиции в портфеле для стратегии составило 23 дня, больше всего покупок было с акциями Сбера – 11 покупок, наибольшая потеря при покупке 30 ноября и продаже 13 декабря акций Сбера – на уровне 10% от цены покупки.

Заключение

Разработанные модели анализа сентимента показали удовлетворительные результаты на данных, размеченных с помощью предложенного автоматического подхода. Для улучшения качества модели анализа сентиментов в дальнейшем предлагаются следующие шаги: увеличение объема данных для обучения; добавление экономических показателей компании, страны, отрасли; использование новостного контекста за определенный период времени для компании и её отрасли, включая новости конкурентов; обучение языковой модели на финансовых текстах на русском языке и дальнейшее её использование для дообучения дополнительного слоя для классификации сентимента.

Использование подобных моделей анализа финансового сентимента даст возможность компаниям и частным инвесторам автоматически оценивать влияние новостной информации на оценку компании и позволит увеличить точность и эффективность принимаемых решений агентами на фондовых рынках. Применение торговой стратегии на основе прогнозов обученной модели позволило показать лучшую доходность, чем стратегия Buy and hold и индекс Мосбиржи в тестовом периоде.

Предложенный автоматический подход для разметки сентимента финансовых новостей может ускорить и упростить сбор и обработку данных для обучения моделей как для российских текстов, так и для новостей на других языках. Возможность подбора порога делает данный метод гибким к оценке сентимента по компаниям, чьи акции исторически высоко волатильны.

Список литературы

1. Malkiel B.G. The efficient market hypothesis and its critics // Journal of economic perspectives. 2003. Т. 17. № 1. С. 59-82. URL: https://www.aeaweb.org/articles?id=10.1257/089533003321164958 (дата обращения: 10.07.2025).

2. Smetanin S. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives // IEEE Access. 2020. Т. 8. С. 110693-110719. URL: https://ieeexplore.ieee.org/document/9117010 (дата обращения: 12.07.2025).

3. Tetlock P.C., Saar-Tsechansky M., Macskassy S. More than words: Quantifying language to measure firms' fundamentals //The journal of finance. 2008. Т. 63. № 3. С. 1437–1467.

4. Sousa M.G. et al. BERT for stock market sentiment analysis //2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI). IEEE, 2019. С. 1597-1601. URL: https://www.researchgate.net/publication/339286476_BERT_

for_Stock_Market_Sentiment_Analysis (дата обращения: 22.07.2025).

5. Cicekyurt E., Bakal G. Enhancing sentiment analysis in stock market tweets through BERT-based knowledge transfer // Computational Economics. 2025. С. 1-23. URL: https://link.springer.com/article/10.1007/s10614-025-10901-8 (дата обращения: 22.07.2025).

6. Huang A.H., Wang H., Yang Y. FinBERT: A large language model for extracting information from financial text //Contemporary Accounting Research. 2023. Т. 40. № 2. С. 806-841. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1911-3846.12832 (дата обращения: 25.08.2025).

7. Shen Y., Zhang P. K. Financial sentiment analysis on news and reports using large language models and finbert // 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024. С. 717-721. URL: https://arxiv.org/pdf/2410.01987 (дата обращения: 03.08.2025).

8. Fazlija B., Harder P. Using financial news sentiment for stock price direction prediction // Mathematics. 2022. Т. 10. № 13. С. 2156. URL: https://www.mdpi.com/2227-7390/10/13/2156 (дата обращения: 05.08.2025).

9. Wu S. et al. Bloomberggpt: A large language model for finance // arXiv preprint arXiv:2303.17564. 2023. URL: https://arxiv.org/abs/2303.17564. (дата обращения: 30.08.2025).

10. Koltsova O.Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. 2016. Т. 2016. С. 277-287. URL: https://scila.hse.ru/data/2020/06/02/1603986481/koltsovaoyuetal.pdf (дата обращения: 29.07.2025).

11. Yakovleva K. Text mining-based economic activity estimation // Russian Journal of Money and Finance. 2018. Т. 77. № 4. С. 26-41. URL: https://rjmf.econs.online/en/2018/4/text-mining-based-economic-activity-estimation/ (дата обращения: 31.07.2025).

12. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF // Journal of documentation. 2004. Т. 60. № 5. С. 503-520. URL: https://www.researchgate.net/publication/238123710_Understanding_Inverse_Document_Frequency_On_Theoretical_Arguments_for_IDF (дата обращения: 02.08.2025).

13. Breiman L. Random forests // Machine learning. 2001. Т. 45. № 1. С. 5-32. URL: https://link.springer.com/article/10.1023/a:1010933404324 (дата обращения: 02.08.2025).

14. Chen T., Guestrin C. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. С. 785-794. URL: https://dl.acm.org/doi/abs/10.1145/2939672.2939785 (дата обращения: 02.08.2025).

15. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. 2017. Т. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 02.08.2025).

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.