АНАЛИЗ НОВОСТЕЙ НА РУССКОМ ЯЗЫКЕ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОКАЗАТЕЛЕЙ ФИНАНСОВЫХ АКТИВОВ

САЯПИН Артём Вячеславович
аспирант
ЯМАШКИН Станислав Анатольевич
кандидат технических наук, доцент
Мордовский государственный университет им. Н.П. Огарева
г. Саранск, Россия

В данной статье рассматриваются предпосылки использования методов обработки естественного языка с целью анализа финансовых новостей для дальнейшего прогнозирования биржевых характеристик активов. Произведен предварительный обзор литературы исследований в данной области, собран набор данных, произведена ручная разметка данных и произведено сравнение базовых моделей классификации новостей.

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

Введение. Интенсивное развитие коммуникационных технологий привело к формированию среды, в которой инвесторы могут мгновенно узнавать новости о различных событиях, заявлениях государственных лиц и представителей бизнеса. Эти новости оказывают влияние на мнение участников рынка и, в свою очередь, на предпринимаемые ими действия. В результате новостной фон может привести к изменению как стоимости, так и волатильности финансовых активов, путем влияния на оценку рыночной ситуации инвесторами. Подобное предположение позволяет поставить под сомнение гипотезу эффективного рынка [1], согласно которой вся доступная информация сразу же отражается в стоимости актива, и все агенты действуют рационально.

Для решения задачи числовой оценки новостного фона применяется анализ сентиментов. Он представляет из себя подраздел обработки естественных языков, направленный на классификацию текстов на основе анализа, содержащихся в тексте сентиментов [2]. Обычно анализ сентиментов принимает вид задачи бинарной или многоклассовой классификации.

В данном исследовании рассматривается использование больших языковых моделей (LLM, Large Language Models) для классификации сентимента новостей на русском языке и дальнейшее влияние полученных оценок новостного фона на моделирование доходности и волатильности для индекса МосБиржы. Целью исследования является оценка применимости больших языковых моделей для русского языка для задачи классификации сентимента для финансовых новостей и рассмотрение возможности дальнейшего использования результатов классификации для моделирования волатильности, цены акций и других биржевых характеристик и их динамики.

1. Обзор литературы.

1.1. Современные методы обработки естественных языков.

Одним из главных прорывов в сфере обработки естественных языков является применение архитектуры трансформеров [3]. Данная архитектура применяется для работы с последовательностями текстов. Она состоит из двух блоков: кодировщик и декодировщик. Ключевой особенностью этой архитектуры является механизм внимания, который позволяет вычислить схожесть каждого слова с другими словами в предложении. Для возможной оценки сходства по разным признакам в модели используются несколько механизмов внимания.

Применение модели BERT для конструирования индексов тональности на основе новостей за предыдущий период времени поз-

волило в 69% случаев верно спрогнозировать изменение индекса Доу-Джонса после дневного старта торгов [4].

1.2. Анализ тональностей текстов на русском языке.

В отличие от английского языка, для которого доступны множество словарей для определения тональности текстов для разных сфер, а также автоматических программных продуктов, в открытом доступе для русского языка находятся только словари РуСентиЛекс [5] и PolSentiLex [6].

Основной проблемой для анализа новостных текстов экономической направленности является отсутствие больших размеченных наборов данных для обучения моделей машинного обучения и тематических словарей тональности. Одной из немногих работ по применению анализа тональности текстов экономической направленности является статья [7], в которой используется классификатор новостей, обученный на размеченных вручную данных, оценки которого в дальнейшем применяются автором для построения индексов тональности для отдельных тематик и их дальнейшего использования в методе опорных векторов для прогнозирования индекса деловой активности России. Данный подход позволил снизить метрику MAE до 1 процентного пункта, что лучше стандартной авторегрессии (2.7 п.п.).

2. Методология исследования.

Обработка текстовых данных будет осуществляться с помощью больших языковых моделей архитектуры Encoder для русского языка, превращающих текст в векторные представления. Новостные данные за день переводятся в формат 2 токенов с помощью автоматического токенайзера модели после чего, они поступают на вход модели обработки естественного языка с добавленным выходным слоем классификации.

Набор новостных данных был получен в результате парсинга новостей с сайта новостного интернет-издания Лента.ру из разделов Экономика, Финансы и Бизнес по компаниям из индекса MOEX.

Итоговый набор данных включает 2457 новостей. Набор новостных данных вручную размечался авторами путем анализа новости и субъективной оценки её влияния на компанию, входящую в индекс MOEX.

Соотношение классов представлено в таблице 1.

Выборки\Классы	Негативный (0)	Положительный (1)
Обучающая	837	797
Тестовая	394	429
Итого	1231	1226

В анализе тональности новостей использовался подход, называемый тонкой настройкой (fine tuning) [8], который предполагает использование обученной языковой модели для прогнозирования замаскированного слова в тексте, и дальнейшее дообучение модели для решения другого типа задач.

Для сравнения качества оценки тональности новости была использована валидационная выборка новостей за период с 1 января 2018 г. до 1 июля 2019 г. Модели обучались на публикациях с 2013 по 2018 гг. со значениями классов (позитивный(1) и негативный(0) сентимент), полученными в результате ручной разметки. Было проведено сравнение трёх LLM-моделей (Sbert Large MT NLU RU, LABSE EN RU, Rubert-tiny2), которые были обучены для моделирования русского языка. Гиперпараметры обучения представлены в таблице 2.

Таблица 2

ГИПЕРПАРАМЕТРЫ ОБУЧЕНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ

Модели	Learning rate	Batch size	Epoch	Token size
SBERT	2e-5	16	2	128
LABSE	1e-5	32	7	256
Rubert	2e-5	16	5	256

В качестве метрик качества использовались accuracy, ROC-AUC, precision, recall, PR-AUC и F1-мера. Сравнение метрик бинарной классификации для моделей приведено в таблице 3.

Таблица 3

СРАВНЕНИЕ МОДЕЛЕЙ АНАЛИЗА СЕНТИМЕНТА НОВОСТЕЙ

Метрики\Модели	SBERT	LABSE	Rubert
Accuracy	0.64	0.63	0.60
Precision	0.66	0.63	0.63
Recall	0.64	0.66	0.55
F1	0.65	0.65	0.59
ROC-AUC	0.67	0.67	0.62
PR-AUC	0.69	0.68	0.65

В большинстве случаев модель Sbert Large показала лучшие значения метрик по сравнению с другими моделями. Матрица ошибок данной модели для валидационной выборки представлена на рисунке 1.

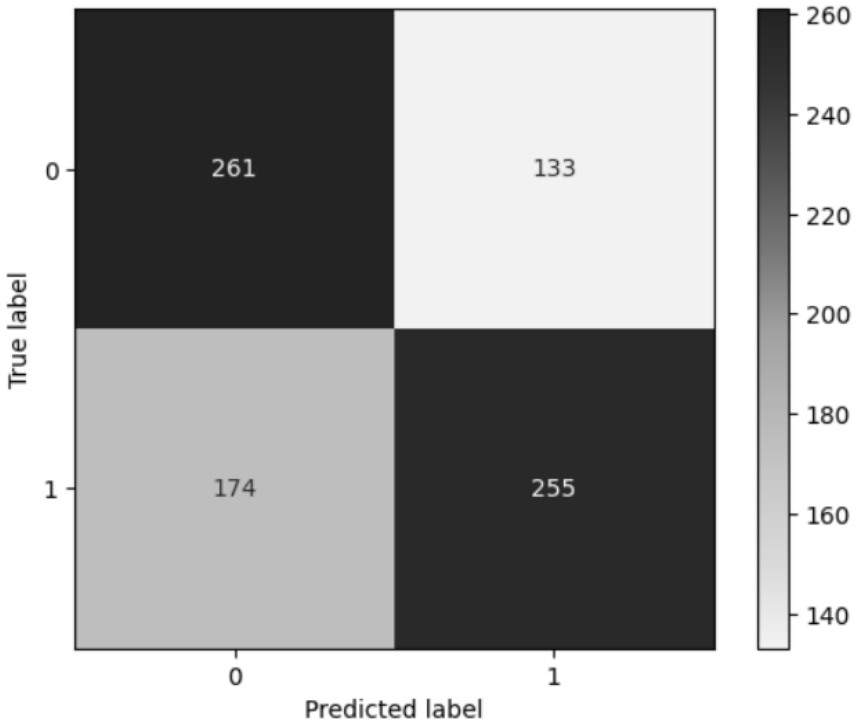

Рисунок 1

Модели показывают лучшие значения, чем случайное угадывание, но требуется дальнейшее расширение набора данных, что позволит добиться лучших показателей качества моделей.

Заключение. В данной статье была рассмотрена область анализа финансовых новостей для дальнейшего прогнозирования биржевых характеристик активов. В результате анализ литературы было выявлено, что данная область для недостаточно хорошо исследована для русского языка и российского рынка акций. Был собран набор данных финансовых новостей, произведена ручная разметка и произведено сравнение классификаторов на основе Больших языковых моделей для русского языка.

В последующем исследовании будет произведена оценка влияния полученных классов на изменения цены, волатильности и других характерстик путем использования их как дополнительных переменных в стандартных авторегрессионных моделях временных рядов.

В дальнейшем также предполагается опробовать автоматический подход для разметки данных, как для задач классификации направления изменения биржевых характеристик, так и для задач регрессии предсказания точного изменения мер активов.

СПИСОК ЛИТЕРАТУРЫ

1. *Koltsova O.Yu, Alexeeva S.V., Kolcov S.N.* An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Компьютерная лингвистика и интеллектуальные технологии. – 2016. – С. 277-287.

2. *Loukachevitch N., Levchik A.* Creating a General Russian Sentiment Lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1171-1176.

3. *Malkiel Burton G.* Efficient Market Hypothesis // Finance. 1989. P. 127-134. DOI: https://doi.org/10.1007/978-1-349-20213-3_13.

4. *Smetanin S.* The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access 8, 2020. P. 110693-110719. DOI: 10.1109/ACCESS.2020.3002215.

5. *Soniya Sandeep Paul, Lotika Singh* A review on advances in deep learning // 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI). 2015. P. 1-6. DOI: 10.1109/WCI.2015.7495514.

6. *Sousa Matheus и др.* BERT for Stock Market Sentiment Analysis // 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). 2019. P. 1597-1601. doi: 10.1109/ICTAI.2019.00231.

7. *Vaswani Ashish и др.* Attention is All you Need // Advances in Neural Information Processing Systems. Т. 30. 2017.

8. *Yakovleva K.* Text Mining-based Economic Activity Estimation // Russian Journal of Money and Finance 77.4, 2018. P. 26-41.

ANALYSIS OF NEWS IN RUSSIAN FOR FORECASTING FINANCIAL ASSETS MEASURES

SAYAPIN Artem Vyacheslavovich
Postgraduate Student
YAMASHKIN Stanislav Anatolievich
Candidate of Sciences in Technology, Associate Professor
N.P. Ogarev Mordovia State University
Saransk, Russia

This article discusses the prerequisites for using natural language processing methods to analyze financial news for further use in forecasting the exchange characteristics of assets. A preliminary review of the literature of research in this area was carried out, a dataset was collected, data was manually labeled, and basic news classification models were compared.
Keywords: sentiment analysis, natural language processing, large language models, text classification, financial news, machine learning.