Определение влияние финансовых новостей на изменение цен Bitcoin-a

Березин Даниил, Кривенко Михаил

Октябрь - Декабрь 2024

Постановка задачи

Целью данного исследования является анализ влияния финансовых новостей на изменения цен на *Bitcoin*. Задача состоит в создании модели машинного обучения, которая использует информацию из финансовых новостей для предсказания изменений стоимости *Bitcoin*. Основные этапы:

- Сбор и подготовка данных: сбор новостей и исторических данных о ценах *Bitcoin*.
- Предобработка текста: токенизация, лемматизация, векторизация с использованием *TF-IDF*.
- Построение модели: использование методов классификации (например, *CatBoost*, *BERT*) для предсказания изменения цен.
- Оценка результатов: использование метрик точности (*accuracy*, *F1-score*) для оценки качества модели.

Описание датасета

Данные о англоязычных новостях криптовалюты за более чем год *(2021-10-12 / 2023-12-19)* в структурированном формате, включающем заголовок, текст, источник, тему и анализ настроений. Данные взяты и из специализированных веб источников, включая *CoinTelegraph* (13,010 публикаций), *CryptoNews* (10,459 публикаций) и *CryptoPotato* (7,568 публикаций), что обеспечивает широкий охват актуальных событий и мнений в криптовалютной сфере.Ссылка на датасет на kaggle.

Источник	Count
CoinTelegraph	13,010
CryptoNews	10,459
CryptoPotato	7,568

Таблица 1: Распределение новостей по источникам

Тема	Count
Bitcoin	9,968
Altcoin	9,278
Blockchain	6,947
Ethereum	2,274
NFT	1,533
DeFi	1,037

Таблица 2: Темы новостей

В составе датасета представлены публикации, тематически связанные с ключевыми аспектами индустрии, такими как *Bitcoin* (9,968 упоминаний), *Altcoin* (9,278), *Blockchain*(6,947), а также более специфические направления, включая *Ethereum*, *NFT* и *DeFi*. Большой спектр тем позволяет исследовать не только прямое, но и косвенное влияние различных новостей на курс биткойна.

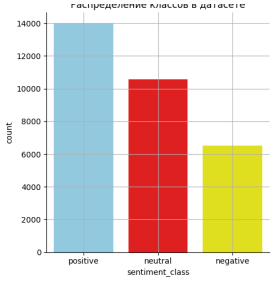

Рис. 1: Баланс классов

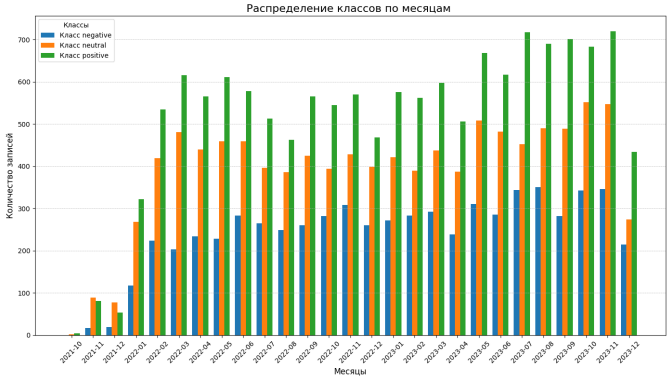

Рис. 2: Распределение классов по месяцам

Распределение классов в датасете следующее: *Positive* - 13,964 записей, *Neutral* - 10,555 записей, *Negative* - 6,518 записей. Данный дисбаланс классов показывает преобладание новостей с положительной или нейтральной тональностью, что может быть связано с общим оптимистичным настроем индустрии криптовалют.

Обзор литературы

1 Обзор использования NLP-моделей для предсказания поведения цен

В последние годы использование моделей обработки естественного языка (*NLP*) для прогнозирования поведения финансовых рынков привлекло значительное внимание. Одним из примеров является работа, описанная в статье *BloombergGPT: A Large Language Model for Finance*. В данном исследовании представлена языковая модель, специально адаптированная для задач финансового анализа.

Авторы модели *BloombergGPT* сделали акцент на использовании разнообразных источников данных. В рамках обучения модель была натренирована на двух группах данных:

- Финансовые наборы данных (363B токенов, что составляет 51,27% от общего объема): включают данные из веб-ресурсов (42,01%), новостей (5,31%), корпоративных отчетов (2,04%) и пресс-релизов (1,21%).

- Общедоступные наборы данных (345B токенов, 48,73% от общего объема): включают данные из открытых источников.

Целью авторов было построение модели, которая обеспечивает высокую точность в финансовых задачах, оставаясь конкурентоспособной по стандартам языковых моделей общего назначения (*LLM*). Однако стоит отметить, что при обучении использовалась архитектура декодера, что ограничивает способность модели учитывать контекст текста в обоих направлениях. Для повышения качества решения задач авторы использовали эффект переноса знаний, что позволило улучшить производительность на целевых задачах. Вместе с тем обучение крупной *LLM* потребовало решения проблемы ограниченности вычислительных ресурсов.

Для оценки производительности модели в финансовых задачах авторы применяли следующие тестовые наборы:

- **FPB (Financial Phrasebank)**: задача классификации настроений в предложениях из финансовых новостей (Malo et al., 2014).

- **FiQA SA**: задача прогнозирования настроений по конкретным аспектам в заголовках финансовых новостей и микроблогов (Maia et al., 2018).

- **Headline**: задача бинарной классификации содержания заголовков товарных новостей (Sinha and Khandait, 2020).
- **NER**: задача распознавания именованных сущностей на основе финансовых данных (Salinas Alvarado et al., 2015).
- **ConvFinQA**: задача ответа на вопросы, требующие численного анализа финансовых данных (Chen et al., 2022).

Наибольший интерес для нашей темы представляет задача **FPB**, где модель продемонстрировала результат **0.51 F1**, что можно считать достаточно высоким показателем для *LLM*.

В статье *Enhanced news sentiment analysis using deep learning methods* представлена модель, основанная на рекуррентной нейронной сети (*RNN*) с блоками долговременной кратковременной памяти (*LSTM*). Данная модель была адаптирована под конкретную задачу предсказания финансовых временных рядов. Средняя точность (**accuracy**) прогнозов модели составила около **0.76**. Однако авторы отметили, что в будущем следует рассмотреть применение подходов на основе механизма внимания (*attention*), что может повысить точность анализа.

Таким образом, использование NLP-моделей, как специализированных, так и адаптированных под конкретные задачи, демонстрирует значительные успехи в прогнозировании финансовых показателей, что подчеркивает актуальность данного направления.

Main

Для предсказания влияния новостей на *bitcoin* были использованы следующие подходы:

- Преобразование текстовых данных с помощью *TF-IDF* и автоматизированного подбора модели *AutoML*.
- Использование алгоритма *CatBoost* с автоматической обработкой текстовых признаков.
- Файнтюнинг моделей типа *BERT*.

1.1 Метод TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) — это популярный метод преобразования текстовых данных в числовые векторные пред-

ставления, который используется для оценки значимости слов в документах относительно корпуса текстов. *TF-IDF* состоит из двух основных компонентов:

- ***TF (Term Frequency)*** — частота термина, рассчитываемая как отношение количества появлений данного слова в документе к общему числу слов в документе:

$$TF(t,d) = \frac{t_d}{d}.$$

  Где t_d — количество вхождений термина t в документе d, d — общее число терминов в документе

- ***IDF (Inverse Document Frequency)*** — обратная частота документа, которая показывает редкость слова в наборе документов. Вычисляется по формуле:

$$IDF(t) = \log\left(\frac{N}{1+d_t}\right),$$

  N — общее количество документов в корпусе, а d_t — количество документов, содержащих термин t.

Результирующее значение *TF-IDF* для термина t в документе d представляет собой произведение *TF* и *IDF*:

$$\textit{TF-IDF}(t,d) = TF(t,d) \cdot IDF(t).$$

TF-IDF позволяет выделять важные слова, которые часто встречаются в конкретных документах, но редко — в остальном корпусе. Таким образом, метод эффективен для преобразования текстовых данных в числовую форму, пригодную для работы с моделями машинного обучения.

1.2 AutoML и модель классификации

AutoML (Automated Machine Learning) — это подход, направленный на автоматизацию разработки моделей машинного обучения. В данном случае использовался перебор моделей и подбор их гипер параметров, лучший результат показала модель градиентного бустинга `LGBMClassifier`. Настройки гиперпараметров для лучшей модели следующие:

Параметр	Значение
colsample_bytree	0.6649
learning_rate	0.1740
max_bin	255
min_child_samples	3
n_estimators	159
num_leaves	19
reg_alpha	0.00098
reg_lambda	0.00676

Таблица 3: Параметры модели

1.3 Результаты AutoML и tf-idf

Оценка точности модели проводилась на тестовой выборке, содержащей 6208 записей. Модель продемонстрировала следующие метрики качества классификации:

Класс	F1-score	Support
0 (Negative)	0.71	1350
1 (Neutral)	0.79	2077
2 (Positive)	0.82	2781
Итого	**0.79**	**6208**

Таблица 4: Отчёт Классификации модели LGBMClassifier

Видно, что модель наиболее эффективно распознаёт положительные классы новостей (F1-score = 0.82). Меньшая точность наблюдается для негативных классов, что может быть связано с их меньшей представленностью в данных.

2.1 CatBoost

2.2 Принцип работы бустинга и CatBoost

Бустинг представляет собой метод ансамблевого обучения, который комбинирует множество деревьев решений для создания одной сильной модели. Основной принцип заключается в последовательном обучении моделей, где каждая следующая модель старается исправить ошибки, допущенные предыдущими. Итоговый прогноз формируется как взвешенная сумма предсказаний всех моделей.

В данной задаче используется алгоритм *CatBoost*, Основные особенности которого:

- **Эффективная обработка категориальных признаков:** CatBoost автоматически преобразует категориальные данные в числовую форму, используя метод *mean target encoding*, что позволяет избежать проблемы переобучения.
- **Обработка текстовых признаков:** Для задачи предсказания влияния новостей на рынок криптовалют CatBoost был обучен на текстовых признаках, преобразованных в числовую форму.
- **Высокая производительность:** CatBoost оптимизирован для быстрого обучения с использованием GPU и работы с большими объемами данных.

2.3 Обучение

При обучении использовался метод проверки на валидационной выборке на каждой итерации для отслеживания переобучения. Обучение происходило на исходном 70% от исходного датасета. Для обучения использовался не только текст новости но и ее заголовок, длинна строки в символах и в словах, а так же обработанный текст (все слова приведены в нормальную форму, удалены "стоп-слова"и т.д.)

Ниже представленны параметры модели:

Параметр	**Значение**
`iterations`	7000
`learning_rate`	0.05
`task_type`	`GPU`

Таблица 5: Параметры модели CatBoost

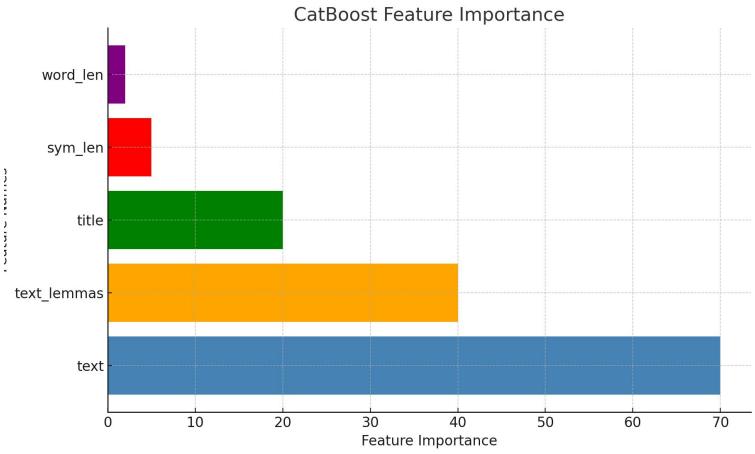

Рис. 3: Важность признаков, используемых в модели CatBoost.

Наибольший вклад в предсказание целевого значения вносит признак *text*, *text_lemmas*, отражающий лемматизированную версию текста оказался менее полезен. *title* (заголовок новости), *sym_len* (длина текста в символах) и *word_len* (длина текста в словах) оказались еще менее значимыми, что кажется достаточно логичным.

2.4 Результаты CatBoost

Оценка точности модели проводилась на тестовой выборке, содержащей 6208 записей(30% исходного датасета). Баланс классов повторяет исходное распределение.

Метрика	**Значение**
`F1 (macro average)`	0.8465
`Accuracy`	0.8589

Таблица 6: Результаты модели CatBoost на тестовых данных

3.1 BERT

Для данной задачи использовались модель *BERT* и *mBERT*

3.2 Принцип работы алгоритма BERT

Алгоритм *BERT* (*Bidirectional Encoder Representations from Transformers*) представляет собой модель на основе трансформеров, которая обучается с учетом контекста слов с обеих сторон (слева и справа) в предложении. Это позволяет модели понимать смысл слов в зависимости от их окружения, что делает её более эффективной для обработки текстов по сравнению с *tf-idf* и *CatBoost*, которые использовали только частоты слова или просто его наличие в новости.

Алгоритм обучения *BERT* состоит из двух основных этапов:

- **Предобучение**: на этом этапе модель обучается на большом корпусе текстов для предсказания пропущенных слов в предложении (маскированное предсказание), а также для предсказания отношения между двумя предложениями (задача *Next Sentence Prediction*).
- **Fine-tuning**: после предобучения модель адаптируется под конкретную задачу, такую как классификация текста, анализ настроений или другие задачи, с использованием небольшого набора данных с метками.

Таким образом, *BERT* является мощным инструментом для решения широкого круга задач обработки естественного языка, благодаря своей способности учитывать двусторонний контекст слов в тексте.

3.3 Обучение

Для обучения использовалось 70% от исходного датасета. В качестве текста в модель передавалась лемматизированная пара: новость - заголовок

Лучший результат на обучении показал *BERT* в отличии от *mBERT* он был предобучен только на английском кормусе.

Parameter	Value
`learning_rate`	5e-5
`max_steps`	5000
`num_train_epochs`	4
`batch_size`	16

Таблица 7: BERT Training Parameters

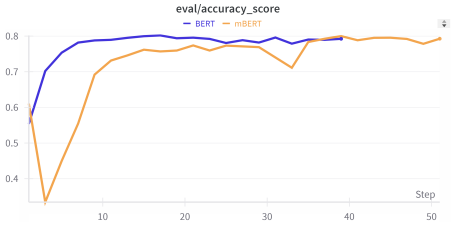

Рис. 4: Сравнение *accuracy* на валидации для *mBERT* и *BERT*

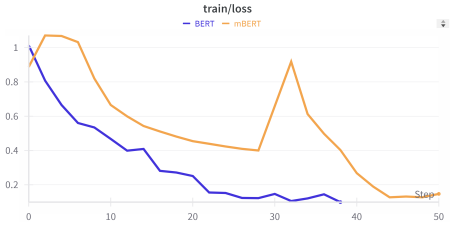

Рис. 5: Сравнение loss-а на обучении для *mBERT* и *BERT*

3.4 Результаты BERT

Метрика	Значение
`Accuracy`	0.7989
`F1-Score`	0.7989

Таблица 8: Результаты модели BERT

3.5 Интерпретируемость

LIME (Local Interpretable Model-agnostic Explanations) — это метод интерпретации моделей машинного обучения, который объясняет предсказания, основываясь на локальных аппроксимациях сложных моделей.

Принцип работы *LIME* заключается в следующем:

- Для выбранного предсказания *LIME* генерирует множество искусственно модифицированных примеров, изменяя значения признаков в окрестности исходного примера(в данном случае изменение/удаление добавления слов в текст новости).
- Для каждого искусственного примера вычисляется предсказание исходной модели.
- Эти примеры используются для построения простой интерпретируемой модели (например, линейной регрессии), которая аппроксимирует поведение сложной модели вблизи выбранного наблюдения.
- Признаки, которые наиболее сильно влияют на локальное предсказание, интерпретируются как основные факторы, определяющие решение модели.

Таким образом, *LIME* позволяет объяснять решения сложных моделей на уровне отдельных наблюдений, оставаясь независимым от конкретного типа модели.

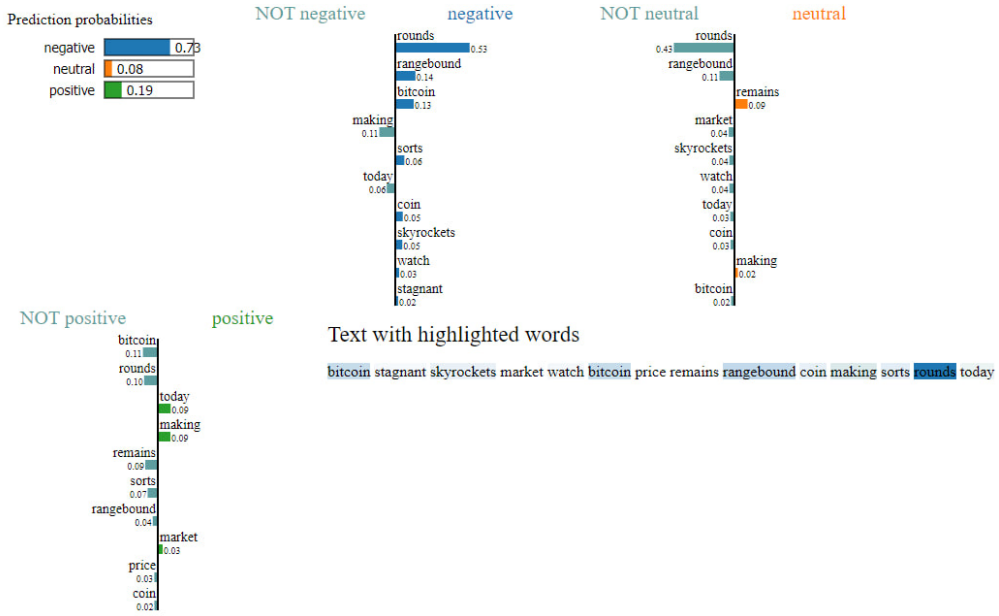

Рис. 6: Визуализация важности слов для модели *BERT* в случае *negative* класса

На рисунке 6 представлена визуализация работы модели анализа тональности текста.

- **Вероятности классов**: Модель предсказывает три вероятности: **negative** , **neutral** и **positive**. Видно, что наиболее вероятной модель считает отрицательную тональность (0.73), что является верным предсказанием.
- **Важность слов**: Графики показывают, как отдельные слова текста влияют на вероятность каждого класса:
  - Для **negative** доминируют слова `rounds` и `rangebound`.
  - Для **positive** ключевыми являются `bitcoin` и `today`.
  - Для **neutral** важны слова `remains` и `rangebound`.
- **Текст с выделенными словами**: В нижней части рисунка представлен текст, в котором выделены важные для модели слова, их вклад в итоговое предсказание классификации обозначен цветами.

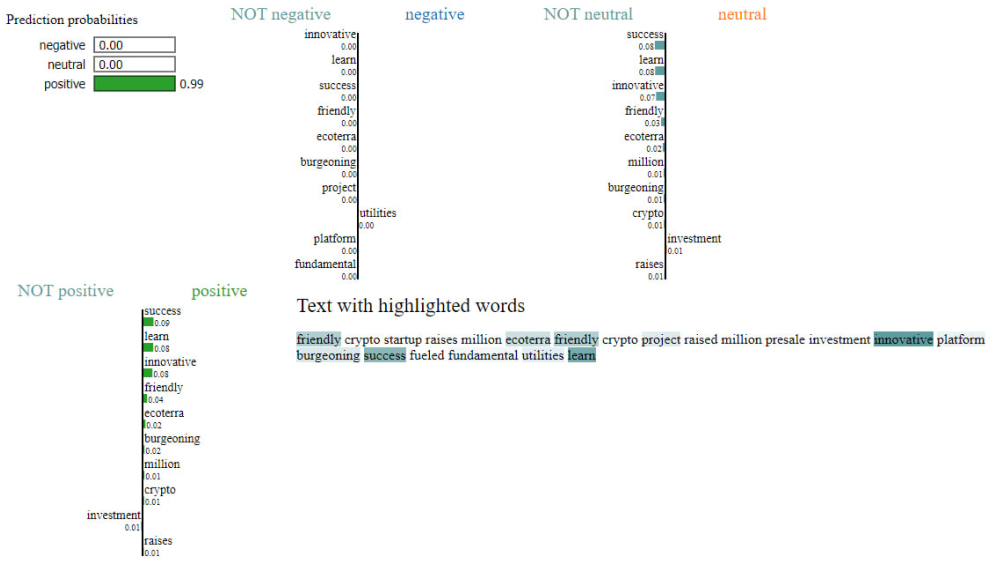

Рис. 7: Визуализация важности слов для модели *BERT* в случае *positive* класса

На рисунке 7 представлена визуализация результатов анализа тональности текста.

- **Вероятности классов**: Модель предсказала высокую вероятность для класса **positive** (0.99), в то время как вероятности классов **negative** и **neutral** равны 0. Это указывает на высокую уверенность модели в положительной тональности текста.

- **Важность слов**: В нижней части визуализации представлен исходный текст , слова `friendly`, `success`, `innovative`, и `learn` выделены как ключевые для положительной тональности текста.

Данная визуализация помогает понять, почему модель классифицировала текст как положительный, демонстрируя её интерпретируемость и связь с конкретными словами.

4.1 Результаты классификации

Тестирование *BloombergGPT* проходило на **FPB** задаче, которая смежна с нашей, однако на другом датасете - собранном в компании *Bloomberg*.

Модель	Accuracy	F1-Score
CatBoost	0.8589	**0.8465**
BERT	0.7989	0.7989
TF-IDF + AutoML	-	0.79
BloombergGPT	-	0.51*
RNN + LSTM	0.76	-

Таблица 9: Результаты обучения для разных моделей

Выводы

В ходе работы была решена задача определения влияния финансовых новостей на изменения цен *Bitcoin*. Для этого использовались различные модели машинного обучения, такие как *CatBoost*, *BERT* и методы на основе *TF-IDF*. Модели были обучены на исторических данных о ценах *Bitcoin* и новостях, после чего их эффективность была оценена с использованием метрик *F1-score* и *accuracy*. Программно были реализованы модели *CatBoost*, *BERT* и *AutoML*.

Результаты показали, что наилучшие показатели были достигнуты с использованием модели *CatBoost*, однако при помощи модели *BERT* удалось оценить влияние конкретных слов в новости на принятие итогового решения.Были побиты результаты полученные в статьях *BloombergGPT: A Large Language Model for Finance* и *Enhanced news sentiment analysis using deep learning methods*.