ИТОГОВЫЙ ПРОЕКТ

АНАЛИЗ СЕНТИМЕНТА РЫНКА

на примере: криптовалюты TON
сайта bitcointalk

Трусковская Д.Р. МФР232

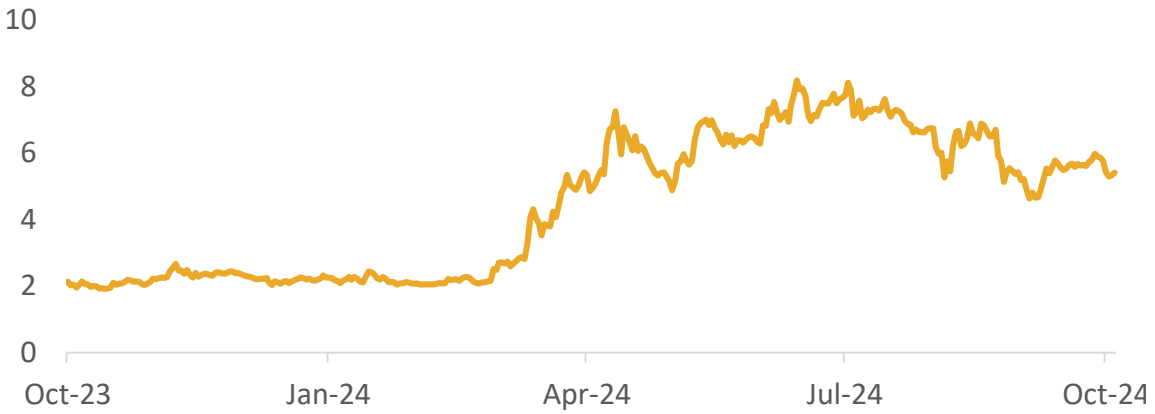
13.12.2024

TON + BITCOINTALK

Почему криптовалюта TON?

TON – довольно известная криптовалюта в СНГ-сегменте (и не только). Выпущена создателями месенджера Telegram. Инфраструктуру TON также могут использовать валюты HMSTR и NOT и другие, что также может влиять на курс TON.

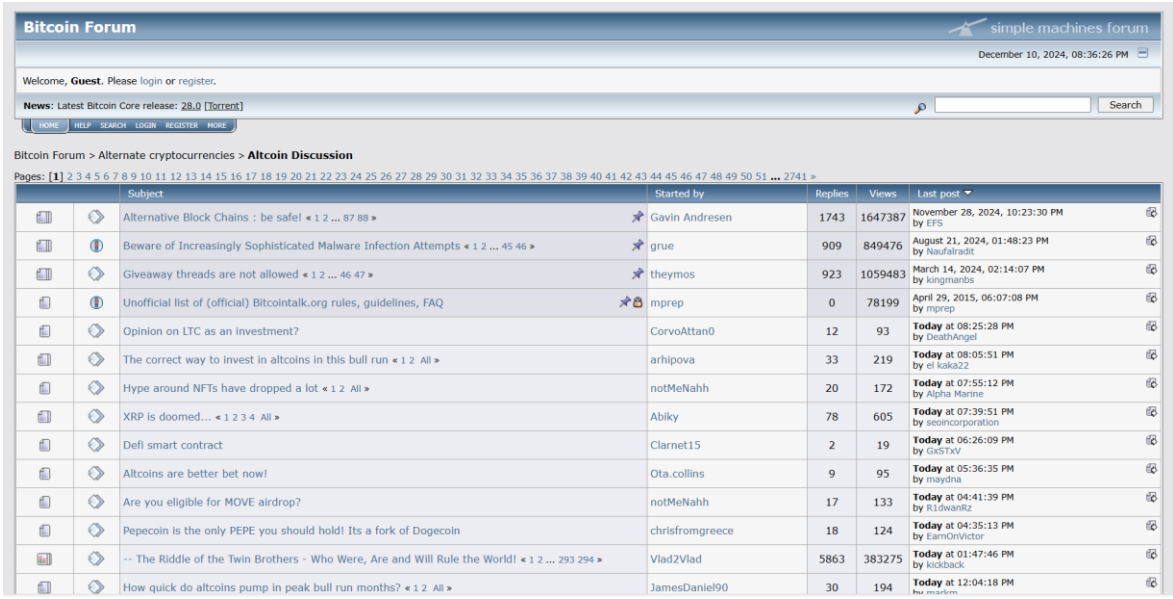
Дневные котировки скачены с Investing за период с 20.09.2023 до 04.11.2024



Источник: Investing, bitcointalk

Почему сайт bitcointalk.org?

TON не такая популярная валюта, как BTC и ETH, и не является одной из популярных мем-койнов, поэтому кроме телеграмм-каналов сложно найти сайт с активным обсуждением TON. Теоретически на «древнем» форуме, посвященном всем криптовалютам должно быть достаточное обсуждение и TON



ПАРСИНГ КОММЕНТАРИЕВ [ДЗ 1]

С каждой ссылки скачиваются: номер темы, номер поста, дата и время поста, автор, ранг автора на сайте, сам пост (на данном форуме любой комментарий в обсуждении является постом) и цитаты, если есть (цитаты = части предыдущего сообщения в дереве обсуждений)

Фильтр по ключевым словам "ton", "toncoin" без учета регистра + фильтр по датам с 20.09.2023 до 04.11.2024

Концепция:
1. Собрать все ссылки за 1 год 1 месяц с Обсуждения альткойнов на BT
2. Парсить все страницы и добавлять в общий датафрейм
3. Применить фильтр по ключевым словам

Парсинг в основном по XPath

Итоговый df
1523 комм

	topic_number	post_number	post_date	post_time	author	rank	content	quotes
0	5516468	26	2024-11-04	05:39:56 PM	Z_MBFM	Sr. Member	Bitcoin price is moving abnormally because of ...	Quote from: RaraAvis on November 02, 2024, 10:...
1	5516588	1	2024-11-03	07:09:35 PM	AHOYBRAUSE	Hero Member	So I just logged into my atomic wallet on desk...	NaN
2	5516588	2	2024-11-03	07:39:59 PM	Stalker22	Legendary	How do you still feel comfortable using Atomic...	Quote from: AHOYBRAUSE on November 03, 2024, 0...
3	5516588	3	2024-11-03	08:40:31 PM	Charles-Tim	Legendary	It could be a hackers means to do something ju...	NaN
4	5516588	7	2024-11-04	10:40:45 AM	GxSTxV	Hero Member	I Personally have seen similar case with MetaM...	NaN
...	...	...	...	...	...	...	...	...

Источник: bitcointalk

ОБУЧЕНИЕ МОДЕЛИ [ДЗ 2]

Предоставленный для обучения датасет нужно было разделить на 9 датасетов и обучить модель, которая бы подошла для всех датасетов и показала результат на тестовой выборке «Accuracy > 0.71»

Для начала было создано 3 массива данных:

- текст после базовой очистки (базовый)
- текст, в котором слова приведены к своим основам (стемминг)
- текст, где слова приведены к нормальной форме (леммация)

Далее подбирались параметры для каждого массива. Подобранные параметры использовались при токенизации каждого отдельного массива

Далее я обучала модель FCNN (Fully Connected Neural Network) на всех датасетах. Тестовая выборка 20%. Цель – обучить модель предсказывать тональность комментариев. Результат точности предсказания на тестовой выборке представлен в таблице ниже. В ходе выполнения следующей части проектной работы использовала датасет, в котором слова приведены к своим основам и представленных методом TF-IDF

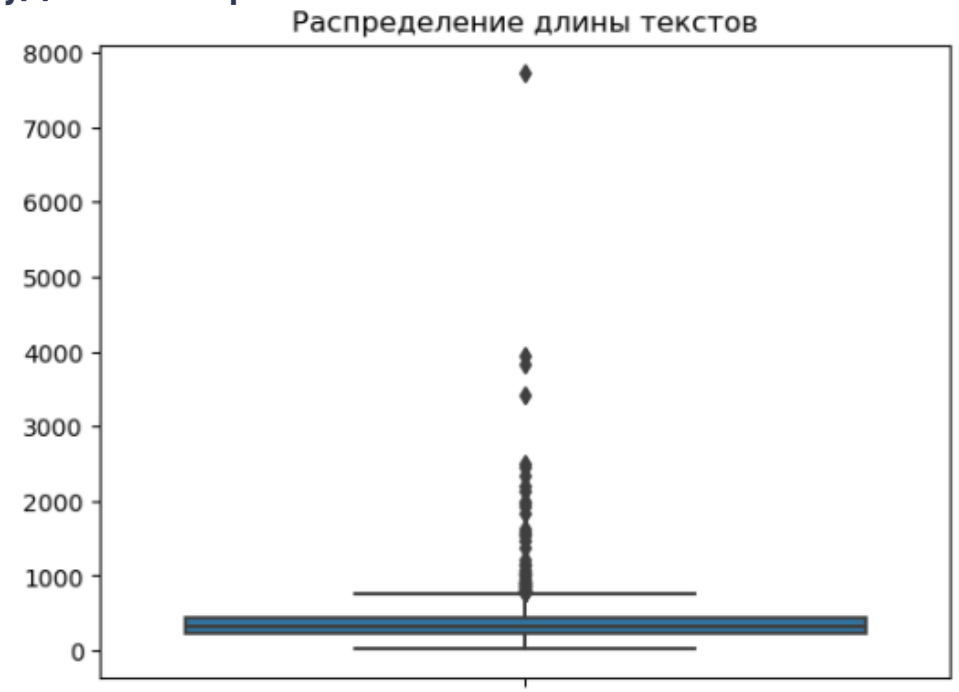
	BAG OF WORDS			TF-IDF			WORD2VEC		
	Базовая	Стемминг	Леммация	Базовая	Стемминг	Леммация	Базовая	Стемминг	Леммация
Accuracy	0.8753	0.8918	0.9054	0.9491	0.9503	0.9523	0.7937	0.7298	0.7403

Источник: собственные расчеты

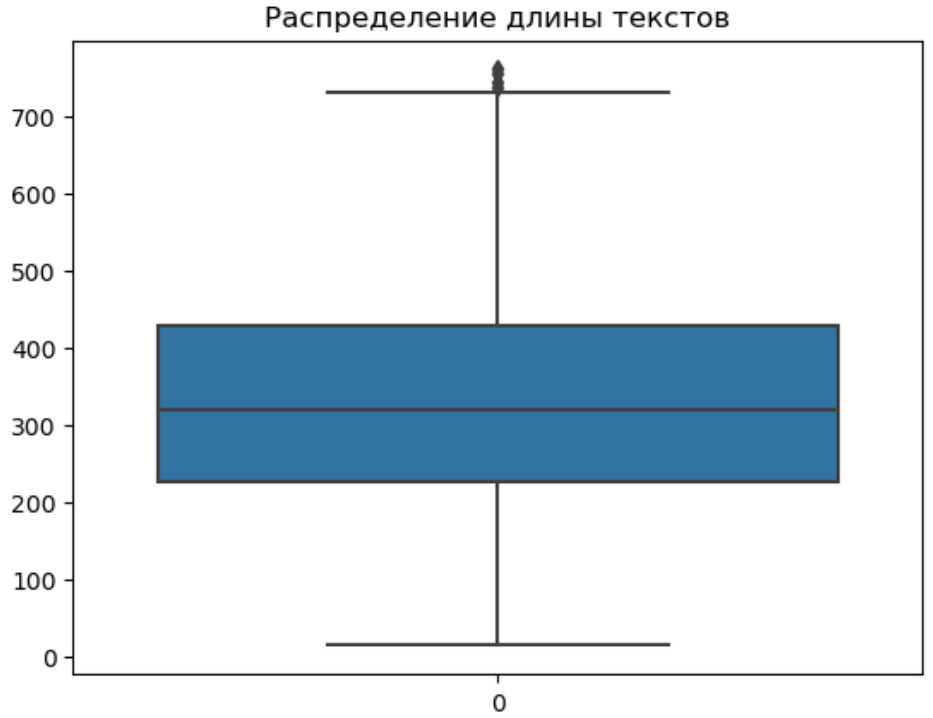
ПОДГОТОВКА ДАТАСЕТА [1/2]

1. Для начала были исследованы выбросы: слишком длинные/короткие комментарии

До удаления выбросов:



После удаления выбросов:



Данные сильно ассиметричны, поэтому выбросы выделялись и впоследствии удалялись методом межквартильного размаха (IQR - Interquartile Range)

$$\text{IQR} = Q_3 - Q_1$$

Q_1 (Первый квартиль) – это значение, ниже которого лежит **25%** данных.

Q_3 (Третий квартиль) – это значение, ниже которого лежит **75%** данных.

Нижняя граница выбросов:

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR}$$

Верхняя граница выбросов:

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR}$$

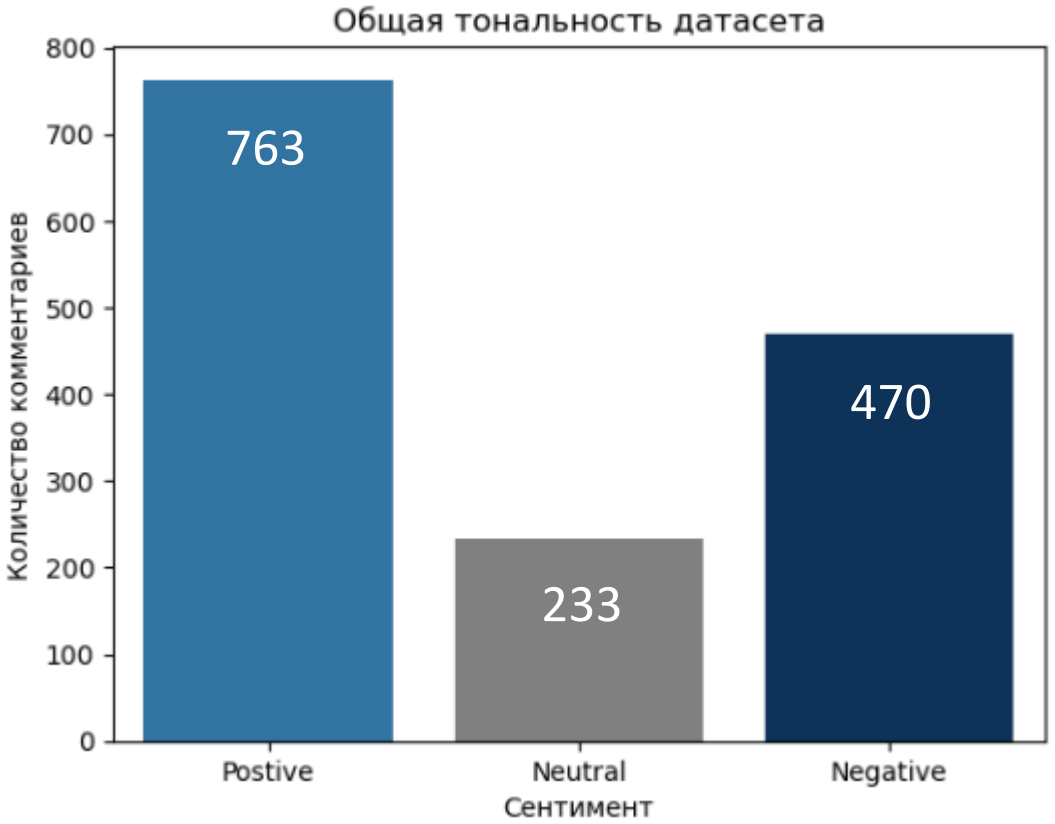
Итоговый df

1466 комм

Источник: собственные расчеты

ПОДГОТОВКА ДАТАСЕТА [2/2]

2. Модель по предсказанию тональности текста обучалась на датасете, в котором слова приведены к своим основам (путем стемминга) и представленных методом TF-IDF. Соответственно наш датафрейм с комментариями необходимо было привести к такому же виду, чтобы модель грамотно определила тональность каждого комментария



Источник: собственные расчеты

3. В процессе парсинга собирался «ранг» авторов. Предполагается использовать его в процессе анализа сентимента по опытности авторов

Распределение авторов комментариев по рангам на сайте

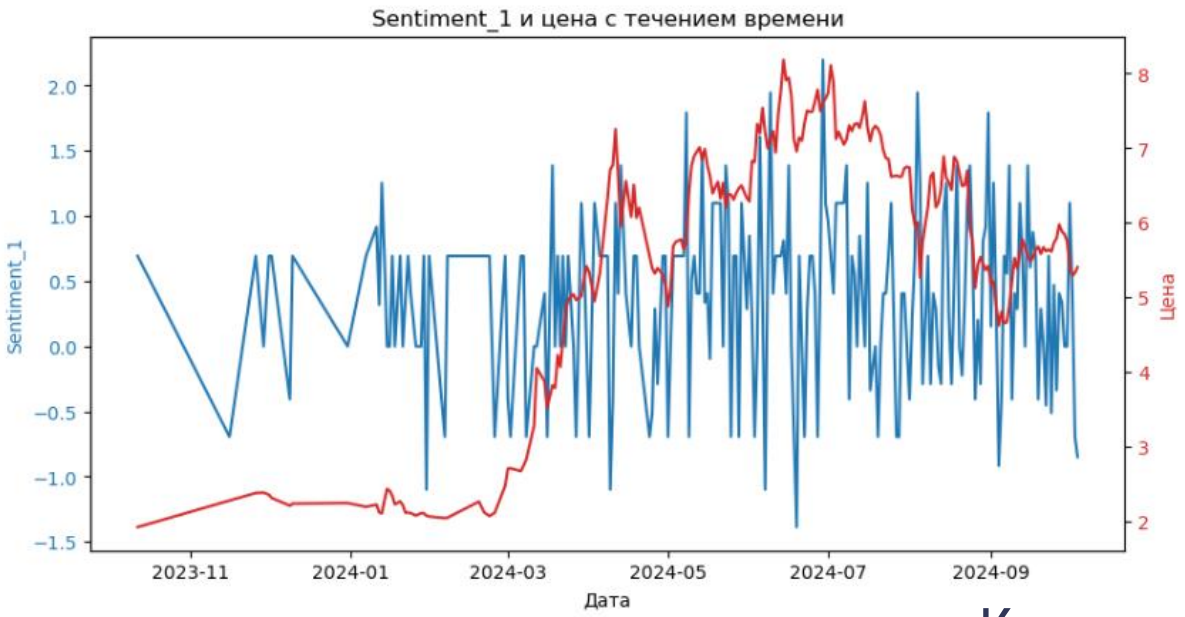
Ранг на сайте	Количество	Обобщенный ранг
Newbie	84	Low rank
Jr. Member	95	
Copper Member	135	
Member	95	Mid rank
Full Member	122	
Sr. Member	245	High rang
Legendary	312	
Hero Member	375	
Staff	2	
Donator	1	
Сумма комм	1466	

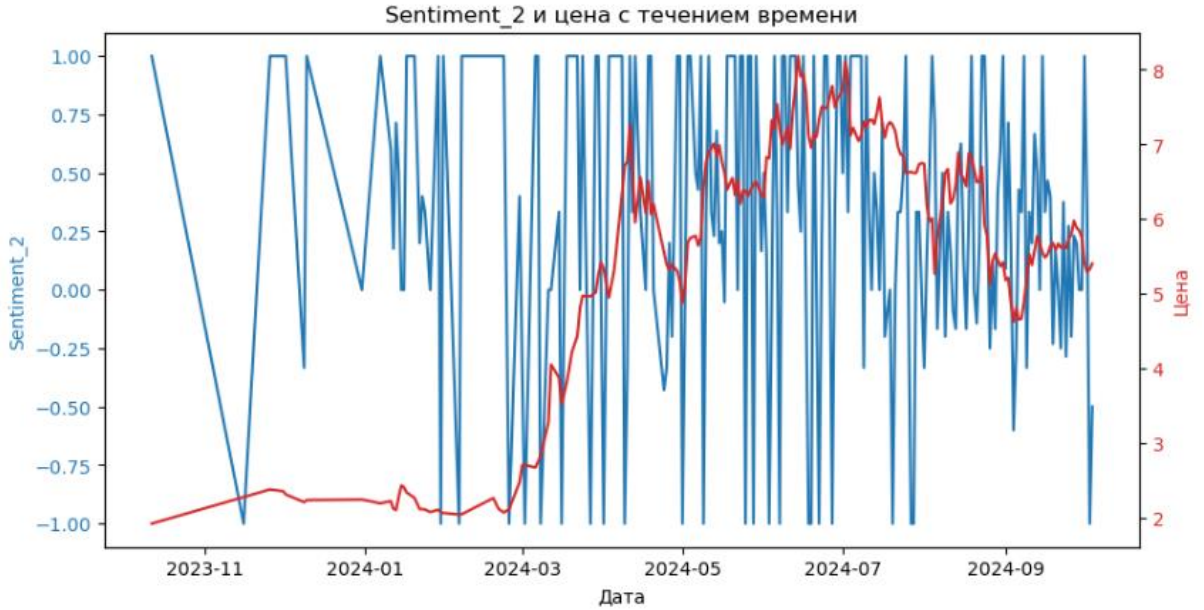
МЕТРИКА СЕНТИМЕНТА (ОБЩАЯ) [1/3]

Получены 2 метрики сентимента: логарифмическая шкала (S_1) и нормализованная разность (S_2)

$$\text{sentiment\_1} = \log\left(\frac{\text{Количество\_положительных} + 1}{\text{Количество\_отрицательных} + 1}\right)$$

$$\text{sentiment\_2} = \frac{\text{Количество\_положительных} - \text{Количество\_отрицательных}}{\text{Количество\_положительных} + \text{Количество\_отрицательных}}$$





Корреляционные матрицы

	Sentiment 1	Close return
Sentiment 1	1	-0.0278
Price	0.0912	1

	Sentiment 2	Close return
Sentiment 2	1	-0.0046
Price	-0.0063	1

‘Sentiment_1’ выглядит более стабильным для дальнейшего исследования, тогда как ‘Sentiment_2’ показывает резкие скачки между -1 и 1, что может усложнить анализ

Источник: собственные расчеты

МЕТРИКА СЕНТИМЕНТА (ОБЩАЯ) [2/3]

Можно заметить, что количество комментариев в целом резко увеличивается с ростом котировок. Отдельно был построен график активности комментаторов в сравнении с котировками

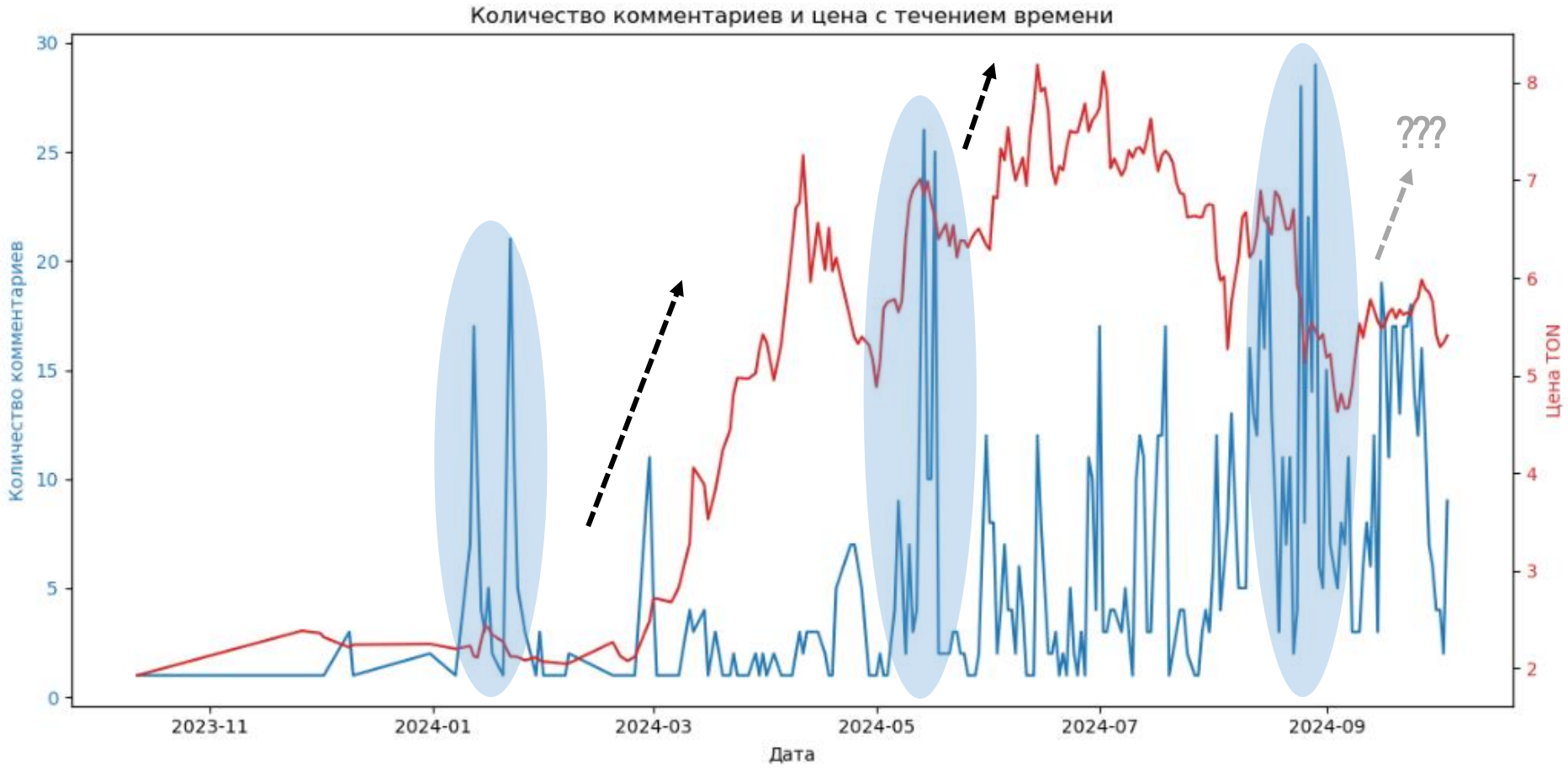
Корреляционные матрицы:

	Comments count	Close return
Comments count	1	-0.1164
Close return	-0.1164	1

Обратная зависимость:
рост количества комментариев сопровождается снижением доходности

	Comments count	Price
Comments count	1	0.1365
Price (lag30)	0.3282	1

Прямая зависимость:
рост количества комментариев сопровождается ростом котировок



Источник: собственные расчеты

МЕТРИКА СЕНТИМЕНТА (ОБЩАЯ) [3/3]

Тест Грейнджера на причинность:

Причина - временной ряд Общий индекса сентимента. Следствие - временной ряд котировок TON

'Sentiment_1'

Лаги	F-test	p-value	Вывод
1	0.2963	0.5867	Нет причинно-следственной связи
2	0.3373	0.7141	Нет причинно-следственной связи
3	0.8742	0.4552	Нет причинно-следственной связи
4	0.6587	0.6214	Нет причинно-следственной связи
5	0.6101	0.6923	Нет причинно-следственной связи

'Sentiment_2'

Лаги	F-test	p-value	Вывод
1	0.0139	0.9062	Нет причинно-следственной связи
2	0.7330	0.4817	Нет причинно-следственной связи
3	1.1795	0.3185	Нет причинно-следственной связи
4	0.8798	0.4769	Нет причинно-следственной связи
5	0.7008	0.6234	Нет причинно-следственной связи

Согласно полученным результатам, 'Sentiment_1' и 'Sentiment_2' не оказывают причинно-следственного влияния на курс TON, т к во всех тестах с разными лагами значение тестовой статистики превышает уровень значимости (p-value > 0.05).

Тест на коинтеграцию:

Не проводится, т к разные порядки интегрирования

	Sentiment	Price
ADF-тест	0.0000	0.2581

Источник: собственные расчеты

МЕТРИКА СЕНТИМЕНТА (ПО РАНГАМ АВТОРОВ) [1/2]

Логарифмическая шкала (S_1)
Нормализованная разность (S_2)

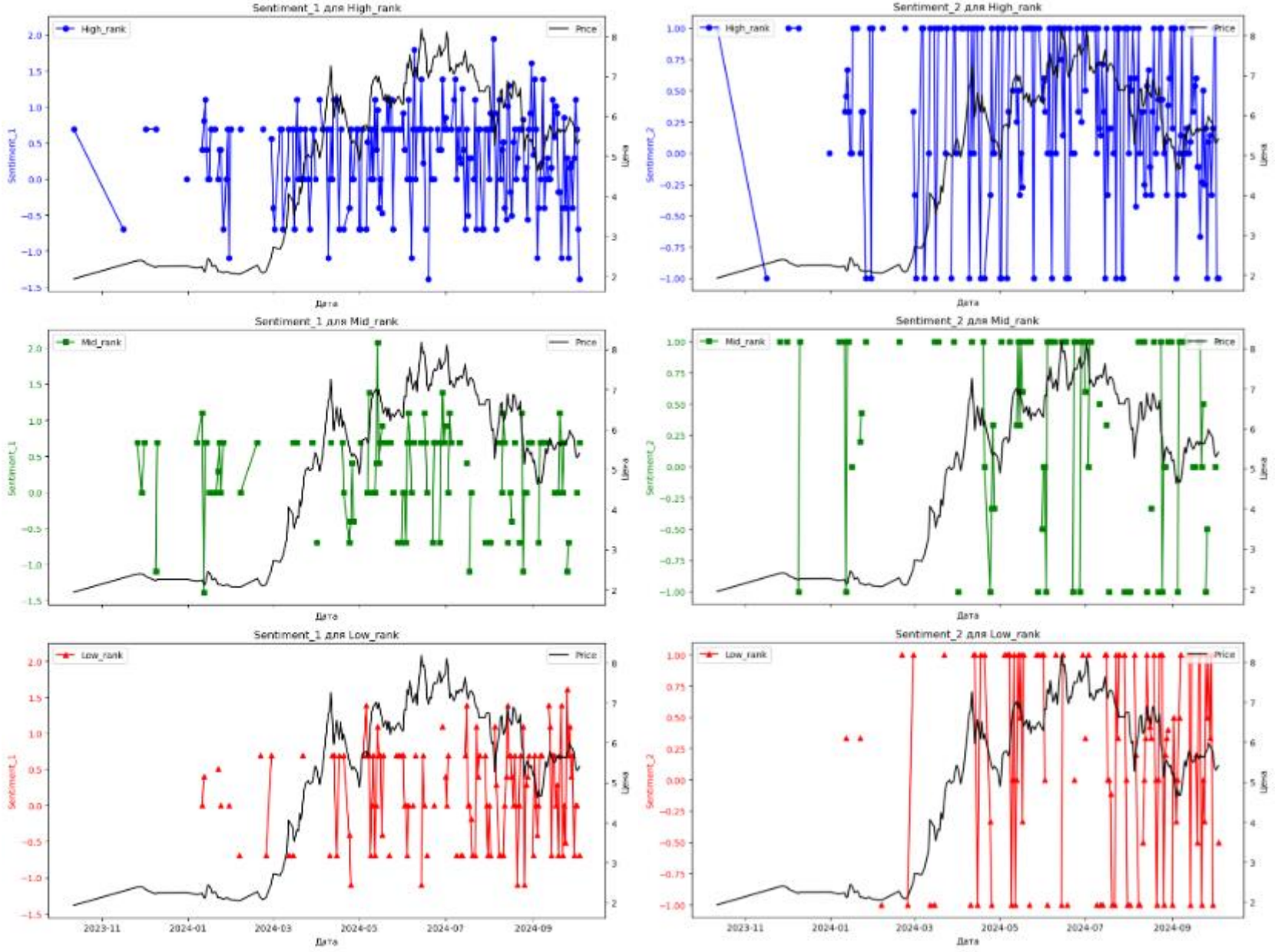
Корреляционная матрица

‘Sentiment_2’ \ ‘Sentiment_1’	High rank	Low rank	Mid rank	Close return	Price
High rank	1	-0.0494	0.1100	-0.0076	0.0847
Low rank	-0.1038	1	0.0030	-0.0497	0.0092
Mid rank	0.0936	-0.0287	1	0.0738	0.0375
Close return	0.0294	-0.0979	0.1246	1	-0.0485
Price	0.0508	0.0097	-0.0593	-0.0685	1

Для ‘Sentiment_1’ и ‘Sentiment_2’ результаты корреляционного анализа схожи. Во всех случаях корреляция между сентиментом и доходностью слабая.

Наибольшая корреляция наблюдается между Close return и Mid rank. Группа Low Rank показывает отрицательные корреляции и с Close Return, и с рангами авторов. Возможно, комментарии в этой группе более критичны и чаще негативные.

Источник: собственные расчеты



МЕТРИКА СЕНТИМЕНТА (ПО РАНГАМ АВТОРОВ) [2/2]

Тест Грейнджера на причинность:

Причина - временной ряд индекс сентимента по рангам (S_1 и S_2). Следствие - временной ряд котировок TON

'Sentiment_1'

Группа ранга	ADF-тест (p-value)	Лаги	F-test	p-value	Вывод
High rank	0.0000	1	0.1798	0.6720	Нет причинно-следственной связи
		2	0.1260	0.8817	Нет причинно-следственной связи
		3	0.3459	0.7922	Нет причинно-следственной связи
		4	0.3277	0.8592	Нет причинно-следственной связи
		5	0.3333	0.8924	Нет причинно-следственной связи
Mid rank	0.0000	1	4.0096	0.0465	Есть причинно-следственная связь на уровне значимости 5%
		2	3.0364	0.0501	Нет причинно-следственной связи
		3	1.9997	0.1151	Нет причинно-следственной связи
		4	1.6895	0.1537	Нет причинно-следственной связи
		5	1.4058	0.2235	Нет причинно-следственной связи
Low rank	0.0000	1	0.4577	0.4994	Нет причинно-следственной связи
		2	0.2714	0.7626	Нет причинно-следственной связи
		3	1.6700	0.1745	Нет причинно-следственной связи
		4	1.2811	0.2785	Нет причинно-следственной связи
		5	1.0405	0.3949	Нет причинно-следственной связи

'Sentiment_2'

Группа ранга	ADF-тест (p-value)	Лаги	F-test	p-value	Вывод
High rank	0.0000	1	0.0417	0.8384	Нет причинно-следственной связи
		2	0.0602	0.9416	Нет причинно-следственной связи
		3	0.1755	0.9129	Нет причинно-следственной связи
		4	0.2403	0.9153	Нет причинно-следственной связи
		5	0.2242	0.9518	Нет причинно-следственной связи
Mid rank	0.0000	1	2.8549	0.0925	Возможна слабая причинно-следственная связь на уровне 10%
		2	1.5396	0.2168	Нет причинно-следственной связи
		3	1.0457	0.3733	Нет причинно-следственной связи
		4	0.9456	0.4386	Нет причинно-следственной связи
		5	0.7721	0.5709	Нет причинно-следственной связи
Low rank	0.0000	1	0.1992	0.6558	Нет причинно-следственной связи
		2	0.1015	0.9035	Нет причинно-следственной связи
		3	1.5904	0.1927	Нет причинно-следственной связи
		4	1.1615	0.3289	Нет причинно-следственной связи
		5	0.9214	0.4681	Нет причинно-следственной связи

Тест на коинтеграцию

Не проводится, т к разные порядки интегрирования

	High rank	Mid rank	Low rank	Price
ADF-тест	0.0000	0.0000	0.0000	0.2581

Источник: собственные расчеты

ЧТО МОЖНО УЛУЧШИТЬ?...

СЛОЖНОСТИ / ВЫЗОВЫ	ИДЕИ ПО УЛУЧШЕНИЮ
Ограниченный временной период	Расширить временной диапазон данных для более глубокого анализа долгосрочных тенденций и зависимостей
Небольшой датасет (1466 комм), неравномерное распределение комментариев и много пропусков в данных	Можно собирать комментарии не только на сайте BT, но и на Reddit, в официальном телеграмм-канале, иных популярных телеграмм-каналах и сайтах с обсуждениями. А также, возможно, исследовать не дневные, а недельные котировки. При решении вопроса пропусков можно использовать не интерполяцию, а иные методы (ближайших соседей или предсказанными)
Низкая корреляция между сентиментом и котировками	Использовать более тонкие метрики сентимента. Например, Word2Vec, который учитывает семантические связи между словами
Использовалась простая модель классификации	Разработка и обучение более продвинутых моделей. Например, CNN + LSTM. Слои CNN хорошо подходят для извлечения временных паттернов из текстовых данных, потому что они могут выявлять соседние связи между токенами (словами или признаками). Слои LSTM обрабатывают временные или последовательные данные и учитывают долгосрочные зависимости. Модель CNN-LSTM может объединить сильные стороны данных типов архитектур
Не учитывался лаг между комментарием и его влиянием на котировку	Вместо фиксированного лага протестировать разные временные окна (14-30-60-90 дней) и выявить наиболее релевантный
Не учитывалось влияние общего новостного фона	Включить в модель анализ новостного фона для учета влияния важных событий (экономических кризисов, политических новостей), а также интегрировать макроэкономические факторы, такие как индексы страха и жадности, курс доллара, индексы (криптовалют)


СПАСИБО ЗА
ВНИМАНИЕ!)