Санкт-Петербургский государственный университет

КАРТАШОВС Романс

Выпускная квалификационная работа

Извлечение лингвистической информации из социальных медиа для предсказания трендов на рынке криптовалют

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805.2020. «Компьюьютерная и прикладная лингвстика»

Научный руководитель:
доцент, Кафедра математической лингвистики,
Митрофанова Ольга Александровна

Рецензент:
к.т.н, Карельский научный центр РАН РАН
Крижановский Андрей Анатольевич

Санкт-Петербург
2022

Содержание

Введение

С каждым годом всё больше людей увлекаются криптовалютами, а 2021 год был богат на новостные заголовки о многократном взлёте и падении цен на крипторынках. Люди обсуждают и делятся опытом покупки нового платёжного средства. В социальной сети Reddit даже существуют отдельные сообщества, состоящие из миллионов активных пользователей, где люди разделяют свои интересы и оказывают поддержку новой технологии.

Актуальность работы обусловлена стремительно растущим интересом к технологиям блокчейн и, в частности, криптовалютам. После скачка цен на акции компании GameStop из-за пользователей соцсети Reddit стало очевидно, что социальные сети и деятельность пользователей в них может оказывать влияние на фондовые рынки.

Объектом исследования является лингвистическая информация, ассоциированная с криптовалютами и их ролью на финансовом рынке. **Предметом** исследования является тональная оценка текстов, в которых обсуждаются криптовалюты, общая тематика публикаций о документах и взаимосвязи тем в статике и динамике.

Новизна исследования заключается в комплексном рассмотрении существования взаимосвязи между публикациями в социальной сети Reddit и трендами на рынке криптовалют.

Цель работы заключается в том, чтобы выявить взаимосвязь между пользовательской активностью в тематических группах в соцсетях и трендом на рынке криптовалют, а также определить варьируется ли сила этой связи в зависимости от конкретной криптовалюты.

Поставленная цель предполагает выполнение следующих **задач**:

1) изучить методы анализа тональности текста,

2) изучить методы тематического моделирования,

3) собрать корпус, состоящий из текстов постов и комментариев пользователей криптосообществ в Reddit за 2021 г.,

4) построить модель анализа тональности текстов корпуса,

5) построить стандартную тематическую модель корпуса,

6) построить динамическую тематическую модель корпуса,

7) сравнить полученные результаты и обобщить их.

Материалом для данного исследования стали англоязычные тексты публикаций в социальной сети Reddit.

В исследовании применяются **методы** квантитативной и корпусной лингвистики, машинного обучения, приемы лингвистического анализа, тематического моделирования и тонального анализа.

Теоретическая значимость исследования заключается в установлении связей между лингвистическими данными о тональности и тематике текстов с экстралингвистическими параметрами описываемого в них явления (рынок криптовалют).

Практическое применение результатов исследования возможно в различных областях финансовой деятельности, комбинирующей приемы лингвистической обработки текстовой информации и работу с экономическими данными, предполагающей операции с криптовалютами и прогнозирование их рынка.

Глава 1. Анализ лингвистической информации для предсказания тренда на рынке криптовалют

1.1. Лингвистические данные

В последние годы растет роль обработки текстовых данных для принятия решений в финансовой сфере: разрабатываются экспертные системы для бизнеса и коммерции, виртуальные агенты (чатботы) для обслуживания клиентов банков, рекомендательные системы для принятия решений в сфере инвестиций, кредитования, оценки финансовых рисков. Создаются специализированные лингвистические ресурсы и инструменты для лингвистических исследований в финансовой сфере: словари терминов, тональные словари, корпусы текстов, обучаются модели распределенных векторов и т.д. Многие компании уже оценили пользу применение методов автоматической обработки языка. Такие гиганты как Deloitte, Ernst & Young используют методы NLP в своём документообороте [Senyuk 2021]. Банковские чатботы при общении с клиентами проверяют активность клиента в социальных сетях для высчитывания общего кредитного рейтинга [Tsarouva 2020]. Источниками текстовых данных при этом являются финансовая документация, новостные сообщения, материалы социальных сетей и т.д.

Многие исследователи и специалисты-практики прибегают к анализу лингвистических данных для предсказания цен на фондовой бирже или валютном рынке [Heeyoung 2014, Chuluunsaikhan 2020, Wooley 2019, Chahat Tandon 2021, Ramon, 2021]. Сейчас внимание многих авторов привлечено к особенной проблеме, каковой является предсказание движения цен на рынке криптовалют. Криптовалюта - это разновидность цифровой валюты, учёт

внутренних расчётных единиц которой обеспечивает децентрализованная платёжная система или блокчейн[1].

В исследовании [Heeyoung 2014] учёные анализировали отчёты компаний, так называемую форму 8-к, которую публичные компании США предоставляют Комиссии по ценным бумагам и биржам всякий раз, когда в компании происходят важные деловые события, включая банкротства, увольнения, выборы директора, изменение в финансировании и т.д.

Статья [Heeyoung 2014] доказывает важность анализа лингвистических данных для предсказания цены следующего дня – показано, что при использовании финансовых признаков вместе с признаками полученными неотрицательной матричной факторизацией в показатель был улучшен на 10%. Однако, в исследовании уточняется, что данная техника не может составлять базу для прибыльной стратегии торговли, так как следует принимать во внимание различные экстралингвистические аспекты, связанные с операциями на биржах.

Одним из источников лингвистических данных может быть новостной корпус текстов. Так, в работе [Arora 2019] для определения движения цены Индекса Доу-Джонса исследователи анализировали заголовки новостных статей, причем рассматривались и тексты, напрямую связанные с финансовой тематикой, но и из других сфер жизни общества. Авторы работы пришли к выводу, что Индекс Доу-Джонса, возможно, не имеет достаточно выраженную зависимость от новостей, как например, фондовый индекс S&P 500.

Помимо новостных статей внимание исследователей привлекают социальные сети, поскольку известно, что общественное мнение и

---

[1] https://ru.wikipedia.org/wiki/%D0%9A%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%B2%D0%B0%D0%BB%D1%8E%D1%82%D0%B0 (дата обращения 31.05.2022)

настроения, формируемые в сетевых сообществах, косвенно оказывают влияние на цены акций и криптовалют. Например, в работе [Chahat Tandon 2021] исследуется вопрос влияния публикаций Илона Маска в Twitter на цены биткоина (BTC) и доджкоина (DOGE). Несмотря на то, что некоторые публикации Илона Маска предшествовали изменению цены, они лишь совпали по тональности с рыночными индикаторами, по которым можно было предугадать резкое изменение цены.

В исследовании [Wooley 2019] учёные обращаются к публикациям из Reddit для выявления тренда на рынке криптовалют. В работе делаются выводы, что данные из Reddit действительно можно использовать для повышения эффективности предсказательных моделей. Однако, в другом исследовании также посвящённом предсказанию цен на криптовалюты [Ramon 2021] заключается, что данных из социальных сетей недостаточно.

В нашем исследовании мы исходим из предположения о влиянии информации социальных медиа на курсы криптовалют, при этом допускаем, что для выявления связей между процессами в сетевых источниках и финансовой сфере необходимо располагать как текстовыми данными, так и алгоритмами и инструментами их обработки.

1.2. Подходы к анализу лингвистических данных

Всё множество работ, посвящённых предсказанию цен на акции или криптовалюты, отличается между собой разными подходами к решению поставленной задачи. Цены на любые активы представляют из себя временной ряд, то есть некие статистические данные, изменяющиеся во времени. Существуют модели предсказывающие изменение временных рядов такие как, например, модель ARIMA, иногда называемая модель Бокса-Дженкинса, искусственные нейронные сети, стохастический алгоритм

геометрического Броуновского движения [Islam 2020]. Для предсказания будущего движения цен такие модели используют исторические ценовые данные для обучения.

На изменение цен могут влиять не только различные биржевые показатели, но и внешние события: будь то политические, экономические или внутри определённой компании. Для определения таких событий при анализе текстовых данных большой популярностью пользуются методы тематического и тонального анализа. Тематический анализ помогает выделить группы событий, потенциально имеющих связь с изменением цен, а анализ тональности помогает оценивать тексты, выражающие мнения, что также может иметь значение при анализе событий.

Лингвистические данные выступают признаками для различных моделей машинного обучения, так например, в исследовании [Shynkevich, 2015] используется метод многоядерного обучения с с различными комбинациями линейных, полиномиальных и гауссовских ядер, в работе [Saloni 2019] используется рекуррентная нейронная сеть в комбинации со свёрточной нейронной сетью, в [Kalyani 2016] используется метод опорных векторов. Во всех приведённых моделях лингвистические признаки извлекаются из текстов при помощи алгоритмов анализа тональности. В исследовании [Chuluunsaikhan 2020], в отличие от многих других работ, используется тематическое моделирование как источник лингвистических признаков для обучения сети долгой краткосрочной памяти.

Следует отметить, что модели машинного обучения, которые разрабатываются на основе лингвистических данных, связанных с информацией о ценах на фондовых или криптобиржах, могут иметь разную направленность. В некоторых работах, как например в [Heeyoung 2014], преследуется цель улучшить показатели предсказывающей модели, которая

использует рыночные метрики для предсказания цены. В других, как например в [Kim 2019], анализ тональности используется лишь для предсказания направления тренда на фондовом рынке.

Итак, существующие исследования доказывают корреляцию между содержанием новостных статей или пользовательской активностью в социальной сетях и ценой на фондовом рынке или рынке криптовалют. Социальная сеть Reddit представляется наиболее удобным источником для настоящего исследования ввиду следующих причин:

1) публикации уже распределены по тематическим группам;

2) пользователи обсуждают всё, что связано с темой группы, то есть, с большой долей вероятности, в публикациях и комментариях будет отражена вся информация, касающаяся темы группы – не только личное мнение участников сообществ, но и новости или высказывания известных личностей в Twitter.

В нынешнем исследовании мы также будем исследовать публикации в Reddit и комментарии к ним при помощи алгоритмов анализа тональности, а также тематического моделирования для выявления ключевых событий в сфере рынка криптовалют.

1.3. Анализ тональности

Мы выяснили, что анализ тональности является распространённым методом работы с лингвистическими данными, содержащими информацию о финансовых рынках. Анализ тональности представляет из себя целую отдельную область в системе проблем NLP. Анализ тональности важен, например, при оценивании объективного мнения экспертов в той или иной области, но с ещё более сложной задачей исследователи сталкиваются при

оценивании тональности текстов публикуемых в социальных сетях обычными людьми. В России долгое время проводился Российский семинар по Оценке Методов Информационного Поиска, в рамках которого учёные принимали участие в соревнованиях алгоритмов анализа тональности, также подобные семинары регулярно проводятся и за рубежом, например, проходящий в этом году SemEval-2022.

Проанализируем опыт, накопленный учеными в области анализа тональности, чтобы принять обоснованное решение о выборе алгоритма, соответствующего цели нашего исследования.

Среди подходов к анализу тональности текстов можно выделить прежде всего подходы, основанные на словарях и правилах (инженерно-лингвистические), и подходы на основе машинного обучения [Большакова 2017]. Последние в свою очередь можно разделить на методы машинного обучения с учителем и методы машинного обучения без учителя. Существует и третий подход, гибридный, который использует преимущества как методов на правилах, так и методов машинного обучения.

1.3.1. Методы определения тональности с использованием словарных ресурсов и правил

Поскольку анализ тональности опирается на лексические данные, для такого типа анализа существуют специальные словари оценочной лексики. Однако, подобные словари имеют сильную зависимость от предметной области, что делает невозможным создание универсального словаря лексики без специальных уточнений. Также различные словари могут использовать разные шкалы оценивания лексики. Лексика может расцениваться как положительная, нейтральная или отрицательная, либо высчитываться по определённым форумам и принимать значение, например, в промежутке от -1

до 1. Тональная оценка каждого слова в предложении суммируются, чтобы высчитывать общую оценку тональности для предложения. Разработка шкалы оценок является ответственностью авторов таких словарей.

Не удивительно, что большинство словарей оценочной лексики создано для английского языка. Одним из первых таких словарей General Inquirer появился ещё в 1962 г. [Philip 1962]. Словарь содержит распределение слов по тональности (положительная, негативная), по силе тональности (сильная, слабая), по категории ощущений (удовольствие, боль, моральные оценки) [Большакова 2017]. Среди других известных словарей оценочной лексики английского языка можно назвать: MPQA, SentiWordNet, ANEW [Большакова, 2017]. Поскольку в рамках данного исследования рассматриваются специализированные по тематике тексты, затрагивающие ситуацию на финансовом рынке, рассмотрим источники, пригодные для проведения наших экспериментов.

К словарям оценочной лексики английского языка, собранным специально для финансовой области, относится NTUSD-Fin [Chen 2018]. Словарь собран тайваньскими учеными на основе публикаций в социальной сети StockTwits, которая является аналогом Twitter и является платформой, где трейдеры делятся своими прогнозами. В словаре содержатся 8331 уникальных слов, а также данный словарь примечателен тем, что регистрирует тональную оценку специальных символов, таких как хештеги и эмоджи. Каждое значение в словаре имеет 8 тегов. Словосочетание “бычий рынок” означает “растущий в цене рынок”, “медвежий рынок” означает “падающий в цене рынок” [Чернова, 2022]. Теги выглядят следующим образом:

- “token”: слово, хэштег или эмоджи

- “bull_freq”: частота встречаемости в “бычьих” постах
- “bear_freq”: частота встречаемости в “медвежьих” постах
- “bull_cfidf”: collection frequency в “бычих” постах
- “bear_cfidf”: collection frequency в “медвежьих” постах
- “chi_squared”: результат теста хи-квадрат для токена
- “market_sentiment”: разность взаимной информации
- “word_vec”: 300-размерное векторное представление модели word2vec

Данный словарь содержит множество слов, уникальных для тематики криптовалют, например, такие слова и сокращения как: *hodl, moon, ath.* Также важную роль играют эмоджи, например, эмоджи с изображением луны 🌙 или ракеты, летящей к луне 🚀, что в критосообществе является выражением уверенности или надежды на то, что криптовалюта в скором времени покажет многократный рост в цене. Такие эмоджи имеют соответствующую положительную оценку в словаре. Любопытно, что фамилия Илона Маска, видного сторонника криптовалют и известного своими попытками манипулировать ценами в Twitter, имеет крайне отрицательную оценку в словаре. Однако, данный словарь предоставляет оценки лишь для специфичных слов, относящихся к финансовой сфере и не может служить универсальным источником информации о тональности слов в корпусе.

Подходы на основе правил используют вручную прописанные правила классификации и эмоционально размеченные словари. Несмотря на прекрасную эффективность в текстах из какой-то определенной тематики, методы на основе правил плохо способны обобщать. Кроме того, они крайне трудоёмки в создании, особенно когда нет доступа к подходящему словарю настроений. Последнее особенно характерно для русского языка, потому что

на нём не так много источников, как на английском, особенно в сфере анализа тональности [Smetanin 2020].

1.3.2. Методы определения тональности с использованием машинного обучения

Существует большое количество традиционных алгоритмов машинного обучения для оценки тональности текстов. В последнее время наблюдается рост популярности методов глубокого обучения, в частности, искусственных нейронных сетей. Глубокое обучение представляет собой набор алгоритмов машинного обучения, которые моделируют высокоуровневые абстракции в данных, используя архитектуры, состоящие из множества нелинейных трансформаций и выделяя из данных «скрытые признаки» [Bengio 2009].

В исследовании [Самигулин 2021] проводилась оценка эффективности как традиционных моделей машинного обучения, так и искусственных нейронных сетей в задаче определения тональности новостных статей. Оценки данного исследования окажутся полезными в настоящей работе при выборе алгоритма анализа тональности публикаций социальных медиа.

Рассмотрим традиционные модели машинного обучения, применимые к анализу тональности в нашем случае [Хобсон 2020, Самигулин 2021].

Наивный байесовский классификатор представляет вероятностный подход к решению задачи классификации. Наивная байесовская модель вычисляет условную вероятность отнесения объекта к классу на основе распределения слов в документе. Данная модель основана на теореме Байеса с предположением о том, что все признаки классифицируемых объектов являются независимыми (благодаря этому допущению она получила название

“*наивный* байесовский классификатор”). Обычно при работе с текстовыми документами предположение о независимости признаков не подтверждается, что делает данный алгоритм малоэффективным.

Метод максимума энтропии так же, как и наивный байесовский классификатор, является вероятностным классификатором. Данный метод основан на идее о том, что наиболее характерным распределением вероятностей неопределенной среды являются распределения, которые максимизируют выбранную меру неопределенности при заданной информации о поведении среды. В отличие от наивного байесовского классификатора, метод максимума энтропии не делает предположения о независимости признаков, что позволяет добиться лучших результатов.

Деревья решений – это логический классификатор, который наглядно представляется в виде древовидной структуры, где в узлах регистрируются атрибуты, от которых зависит распределение вероятностей классов, а в листьях указываются значения вероятностей классов. Данный метод прост в интерпретации и требует минимальной предобработки данных. Но сами по себе деревья решений используются редко, так как они слишком зависимы от обучающих данных. При небольших изменениях в обучающей выборке мы получаем разнящиеся результаты на тестовых данных.

Случайный лес – ансамбль решающих деревьев. При создании случайного леса строится большое число “слабых учеников” – решающих деревьев разной глубины на разных обучающих данных с разным выбором признаков. Деревья строятся до тех пор, пока в каждом листе не окажется малое количество объектов, это позволяет избегать переобучения. Затем все деревья объединяются, и мы получаем одного “сильного ученика”, эффективный классификатор, у которого сокращены или отсутствуют

недостатки отдельных решающих деревьев. Но это вызывает некоторые проблемы, если признаков очень много, то этот подход работает не очень хорошо: деревья будут очень глубокими, на их построение будет уходить слишком много времени.

Градиентный бустинг – ансамблевый метод машинного обучения для классификации и регрессии. Этот метод обучается поэтапно, улучшая на каждом следующем шаге модель, которая получилась на прошлом этапе. В качестве базовых используются очень простые алгоритмы, например неглубокие решающие деревья. При использовании градиентного бустинга решающие деревья, в отличие от случайного леса, имеют незначительную глубину, однако это не вызывает проблем. Каждое дерево может учесть лишь небольшое подмножество признаков, в то время как зачастую ответ зависит от комбинации большого количества слов в тексте.

Логистическая регрессия является линейным методом классификации, оценивающим вероятность принадлежности объектов к классу путем сравнения с логистической кривой по значениям множества признаков. Логистическая регрессия применима как для регрессионного анализа, так и для классификации. Логистическая регрессия – это один из самых популярных методов классификации, обученная таким образом модель показывает очень хорошие результаты. Из недостатков логистической регрессии можно выделить то, что необходима качественная предобработка данных и тщательный отбор признаков для обучения.

Метод опорных векторов (SVM) относится к линейным алгоритмам машинного обучения. Цель SVM заключается в нахождении среди всех возможных гиперплоскостей, отделяющих два класса обучающих примеров друг от друга, такой гиперплоскости, расстояния от которой до ближайших

(опорных) векторов обоих классов равны (оптимальная разделяющая гиперплоскость). Как и другие линейные классификаторы, метод SVM применим в случае линейно разделимых данных.

В упомянутом выше исследовании [Самигулин 2021] были показаны результаты лишь для алгоритмов, показавших наилучший результат. В качестве метрики эффективности метода использовалась AUC – площадь под ROC-кривой (кривой ошибок), были получены следующие результаты:

- логистическая регрессия: тестовая выборка – 0.93445.
- дерево принятий решений: тестовая выборка – 0.6500.
- случайный лес: тестовая выборка – 0.84000.
- метод опорных векторов: тестовая выборка – 0.86167.

Здесь же приводятся оценки из другой работы, где авторы в качестве метрики эффективности использовали соотношение правильно предсказанных объектов к общему количеству объектов в наборе данных [Ahmad 2017]:

- метод максимума энтропии – точность 72.60%
- случайный лес – точность 88.39%
- наивный байесовский классификатор – точность 75.50%
- метод опорных векторов – точность 91.15%

На примере исследований [Самигулин 2021] и [Ahmad 2017] видно, что лучший результат дают линейные модели: логистическая регрессия и метод опорных векторов. Однако, эффективность традиционных методов сильно зависит от объема и качества обучающих данных. Кроме того, на результат влияет выбор признаков, что значительно усложняет процесс обучения. В рассмотренных работах также говорится, что эффективность данных

алгоритмов значительно падает при анализе текстов социальных сетей, в которых зачастую не соблюдаются правила грамматики языка.

Иначе дело обстоит с искусственными нейронными сетями, которые сами производят отбор признаков в данных без участия человека. Искусственная нейронная сеть – это математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей. В процессе обучения искусственная нейронная сеть выявляет сложные зависимости между входными данными и выходными. Это значит, что в случае успешного обучения сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искаженных данных. Еще одним преимуществом искусственной нейронной сети является ее способность адаптироваться к разным вариантам постановки задачи с небольшими изменениями в системе, выполняющей анализ тональности [Пескишева 2017].

В обсуждавшейся работе [Самигулин 2021] рассматриваются следующие модели искусственных нейронных сетей.

- **Сверточные нейронные сети (CNN).** В сверточных нейронных сетях используется операция свертки, когда каждый фрагмент данных умножается на матрицу (ядро) свертки поэлементно, после чего результат суммируется и записывается в аналогичную позицию выходных данных. Поскольку свертки происходят на соседних словах, модель может уловить отрицания или n-граммы, которые несут новую информацию о настроении / мнении автора текста.

- **Рекуррентные нейронные сети (RNN).** Особенности рекуррентных нейронных сетей — это наличие обратных связей, связь от более удаленного элемента к менее удаленному. Это позволяет запоминать и воспроизводить последовательности реакций на один стимул. Значение весов сети зависит как от текущих, так и от предыдущих входных данных, благодаря чему вес каждого слова влияет на веса остальных слоев в предложении.

В качестве метрики оценивания использовалось соотношение правильно предсказанных объектов к общему количеству объектов в наборе данных (точность). Сравнение указанных моделей дало следующий результат:

- сверточная нейронная сеть: точность – 86.8%.
- рекуррентная нейронная сеть: точность – 87.5%.

В исследовании [László 2021] приводится сравнение других методов анализа тональности, среди которых:

- **Рекуррентные нейронные сети (RNN)** (см. выше).
- **BERT** – нейронная сеть архитектуры Трансформер, разработанная в компании Google. В отличие от традиционных языковых моделей, использующих однонаправленный подход, то есть чтение текста либо слева направо, либо справа налево, BERT считывает сразу всю последовательность слов. BERT использует преобразователь, который по сути представляет собой механизм для построения отношений между словами в наборе данных. В своей простейшей форме BERT состоит из двух моделей

обработки — кодировщика и декодера. Кодировщик считывает входной текст, а декодер выдает предсказания. Поскольку основной целью BERT является создание обученной модели, кодировщик имеет приоритет над декодером.

Помимо нейросетевых моделей в работе анализируются некоторые инструменты.

- TextBlob — это библиотека автоматической обработки текста для языка Python, построенная на основе NLTK. Этот инструмент можно использовать для выполнения различных задач автоматической обработки текстов, начиная от маркировки частей речи и заканчивая анализом тональности[2] .
- SentimentIntensityAnalyzer пакета NLTK вместе с модулем VADER. VADER (Valence Aware Dictionary and sEntiment Reasoner) — это основанный на словаре и правилах инструмент анализа тональности, специально настроенный на задачу извлечения мнений из текстов, которые встречаются в социальных сетях, и так же хорошо работающий с текстами из других областей[3].

В данном исследовании BERT использовался как основной метод, все остальные сравнивались с ним. В качестве входных данных служили заголовки новостных статей. Рекуррентные нейронные сети показали результат, схожий с результатом работы BERT. В качестве преимущества указывалось отсутствие нейтральной тональности у заголовков, что также было выявлено в экспериментах с BERT. TextBlob и NLTK VADER помимо

[2] URL: https://textblob.readthedocs.io (дата обращения 31.05.2022 г.)
[3] URL: https://www.nltk.org/howto/sentiment.html (дата обращения 31.05.2022 г.)

положительных и отрицательных тональностей присваивали нейтральные оценки заголовкам. В целом, VADER также показал хороший результат, гораздо лучший чем у TextBlob.

Итак, мы рассмотрели два разных подхода анализа тональности: инженерно-лингвистический и с использованием машинного обучения. Использование словарей даёт простой и эффективный способ анализа тональности, но основной проблемой словарей является их строгая привязка к конкретной сфере, что является существенной преградой в их повсеместном использовании. Различные модели машинного обучения являются мощным инструментом в решении данной задачи, но и они не лишены своих недостатков, таких например как качество и объём обучаемых данных.

Из всех методов машинного обучения наиболее эффективными являются нейронные сети и, в частности, рекуррентные нейронные сети. Однако, в исследовании [László 2021] мы увидели, что и другие подходы могут давать схожий результат, как например, BERT или VADER из пакета NLTK. Замечания авторов исследования к VADER были справедливы: действительно, заголовки новостей компаний не должны иметь нейтральную оценку тональности, иначе подобные новости не было бы смысла публиковать. Но тексты, на которых мы предполагаем проводить исследования, могут содержать нейтральный текст, никак не отражающий отношения пользователя Reddit к событиям на рынке криптовалют, поэтому мы не можем утверждать, что данное замечание является достаточным основанием не принимать во внимание этот инструмент. Кроме того, VADER уже имеет свою полноценную реализацию в пакете NLTK и изначально

создавался для работы с текстами социальных сетей, что делает этот метод подходящим инструментом в решении нашей задачи.

1.4. Тематическое моделирование

Тематическое моделирование может оказаться полезной процедурой в выявлении тренда на рынке криптовалют. Существуют работы, в которых описываются примеры таких попыток предсказания цены [Chuluunsaikhan 2020], однако они не многочисленны. Непопулярность подхода применительно к цели нашего исследования не гарантирует, что данный подход является неэффективным.

При тематическом моделировании текстовых документов определяется принадлежность каждого документа к разным темам, генерируется список слов, из которых образована каждая тема. По своей природе тематические модели сходны с нечеткой кластеризацией, потому что каждое слово в корпусе может быть связано с несколькими темами и каждый документ может быть описан несколькими темами. На данный момент существует обширное множество тематических моделей. Их дифференциация по типам определяется тем, как строится промежуточное представление корпуса (матричное, векторное представление), как происходит сокращение размерности модели, какими закономерностями задается соотношение слов, тем и документов, какие параметры лежат в основе тематической модели (авторство, метаданные, хронологические рамки корпуса и т.д.), что дополнительно вводится в модель (иерархия тем, метки тем, генерация тем по ключевым словам и т.д.) [Daud et al. 2010]. Как правило, противопоставляются следующие два типа моделей: алгебраические модели (например, Vector Space Model (VSM), Latent Semantic Indexing (LSI)),

Non-negative Matrix Factorization (NMF) и т.д.) и вероятностные модели (например, Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) и т.д.) [Воронцов 2013]. Наиболее популярной в исследованиях является латентное размещение Дирихле (LDA) [Blei et al. 2003], в котором распределение слов по темам и тем по документам описывается сопряженным к мультиномиальному распределением Дирихле с гиперпараметрами α и β (см. схему 1).

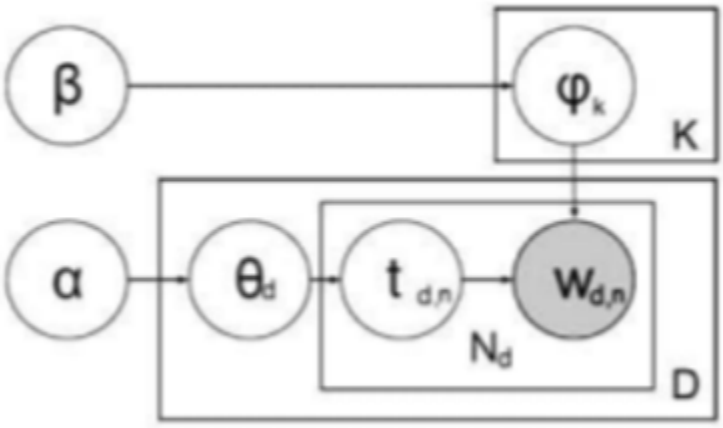

Схема 1. *Архитектура тематической модели LDA*

Определение тем, на которые люди общаются в социальных сетях, является очень полезной задачей не только для научного исследования, но и для бизнеса, политических исследований и т.д. Однако, большую пользу принесёт исследование того, как отношение людей менялось по отношению к тем темам, которые они обсуждают, для этого и созданы алгоритмы динамического тематического моделирования (DTM).

В модели DTM [Blei 2006] эволюция тем прослеживается внутри корпуса текстов, организованного в виде последовательности. Время предполагается дискретным, поэтому мелкие изменения игнорируются, но выявляются интервалы повышения и понижения популярности слов,

описывающих темы.

Альтернативой DTM является многомасштабная томографическая модель [Nallapati et al. 2007], более естественная для анализа последовательных корпусов. В ней используются сопряжённые распределения и одновременно несколько различных масштабов времени. Всё это обеспечивает в результате лучшую интерпретируемость тем [Daud et al. 2010].

Динамическая тематическая модель с непрерывным временем [Wang, Blei, Heckerman 2008] основана на использовании законов броуновского движения [Uhlenbeck, Ornstein 1930] для моделирования эволюции тем в непрерывном времени, что позволяет также снять некоторые ограничения по памяти и производительности, присущие DTM.

В модели тематики во времени [Wang, McCallum 2006] каждая тема связана с распределением по временным меткам. Каждой теме ставится в соответствие бета-распределение, зависящее от времени, так что тема порождает одновременно и слово, и отметку времени. Однако, игнорируются, временные паттерны и возможные взаимосвязи между темами, например, сходство или вложенность (отношение тема–подтема).

Среди алгебраических моделей Неотрицательная матричная факторизация (NMF) также эффективна. В работе [Saha, Sindhwani 2012] предлагается система с использованием NMF для извлечения тем из потокового контента социальных сетей путем разделения потоков на короткие скользящие временные окна. Неотрицательная матричная факторизация является популярной моделью на данный момент.

Моделью, показавшей свою эффективность, а также оказавшейся простой в настройке обучении в ряде предварительных экспериментов данного исследования является модель BERTopic [Grootendorst 2022],

комбинирующая контекстуализированные языковые модели распределённых векторов и алгоритмы снижения размерности[4]. На схеме 2 представлена архитектура модели. BERTopic генерирует представление тем в три этапа. Сперва, каждый документ преобразуется в векторное представление с использованием предварительно обученной языковой модели. Затем перед кластеризацией размерность полученных векторов уменьшается для оптимизации процесса кластеризации. Наконец, из кластеров документов извлекаются тематические представления с использованием модифицированного варианта TF-IDF на основе классов. Преимуществами BERTopic являются высокая обобщающая способность модели, ее многофункциональность, минимальная процедурная поддержка на этапе предобработки корпуса, и т.д.

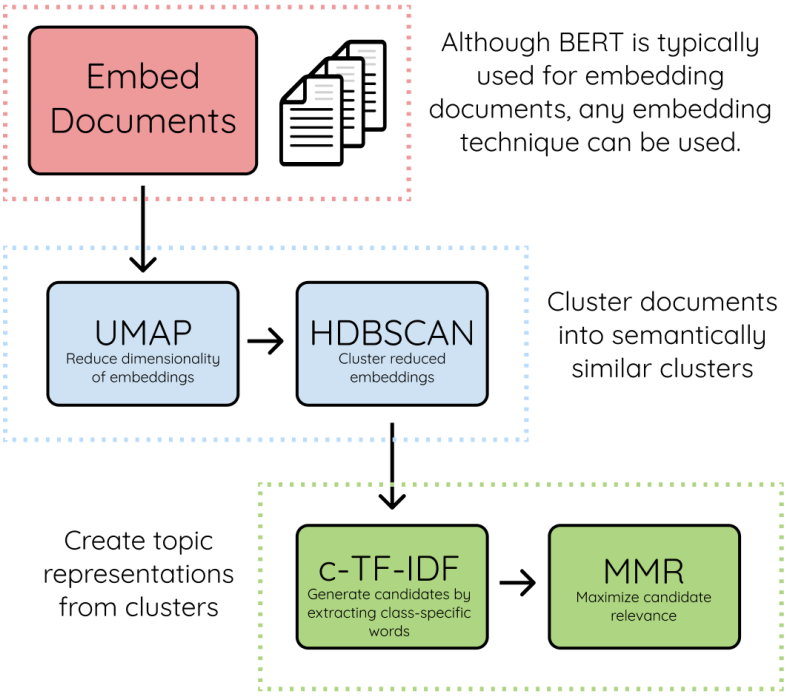

Схема 2. *Архитектура тематической модели BERTopic*

[4] URL: https://github.com/MaartenGr/BERTopic (дата обращения 31.05.2022 г.)

Выводы к Главе 1

Итак, на основании описанного в данной главе можно сделать следующие выводы:

1. Доказана корреляция между новостными статьями или пользовательской активностью в социальных сетях и ценой на рынке криптовалют. Социальная сеть Reddit представляется удобным источником для нашего исследования ввиду следующих причин:
    - Публикации уже распределены по тематическим группам;
    - Пользователи обсуждают всё, что связано с темой группы, то есть, с большой долей вероятности, в публикациях и комментариях будет отражена вся информация, касающаяся темы группы – не только личное мнение участников сообществ, но и новости или высказывания известных личностей в Twitter.
2. Из всех методов машинного обучения в решении задачи анализа тональности наиболее эффективными являются нейронные сети и, в частности, рекуррентные нейронные сети. Однако, в задачи тонального анализа текстов социальны сетей с их уровенем эффектиности можно сопоставить более простую модель, основанную на правилах VADER и словари финансовой лексики.
3. Существует большое количество алгоритмов тематического и динамического тематического анализов. Наиболее популярной в современных исследованиях является модель латентного размещения Дерихле. Также языковая модель BERTopic является крайне перспективным решением проблемы тематического моделирования.

Глава 2. Эксперименты по выявление трендов на рынке криптовалют на материале англоязычного корпуса текстов социальной сети Reddit

2.1. Крпитовалюты и источники лингвистической информации о них

В первой главе мы определили источники данных и инструментарий, которые будут использованы в экспериментальном исследовании. В фокусе эксперимента находятся так называемые альткоины и, в частности, мемкоины. Феномен мемкоинов интересен тем, что они не имеют практического применения в отличие от многих других криптовалют (хотя, к примеу, биткоин используется только как платёжное средство). В 2021 г. большинство криптовалют значительно прибавили в цене и обновили исторические максимумы, мемкоины не стали исключением. Самыми яркими примерами криптовалют, цены на которые возросли, явились доджкоин и шиба ину. В нашей работе мы будем анализировать следующие криптовалюты: биткоин (BTC), эфириум (ETH), доджкоин (DOGE) и шиба ину (SHIB). Биткоин является “золотым стандартом” в мире криптовалют, и изменение его цены часто влияет на цены альткоинов. Эфириум является валютой, или токеном, блокчейна Эфир, инфраструктуры, на которой строятся приложения и в которых данный применяется для разных задач, не только как платёжное средство. Блокчейн Эфир имеет множество практических реализаций и является одним из крупнейших проектов на данный момент, поэтому нам также будет важно учитывать его.

В качестве источника текстов для построения корпуса мы выбрали социальную сеть Reddit[5]. У данного выбора есть ряд преимуществ —

[5] https://www.reddit.com/ (дата обращения: 01.04.2022)

тематическая специализация сообществ, количество пользователей равно 36 миллионам, посещаемость превышает один миллиард в месяц. Также у пользователей Reddit имеются истории успешного манипулирования рынком ценных бумаг. Английский язык является лингва-франка для подобных социальных сетей, поэтому англоязычные тексты преобладают над текстами других языков социальной сети Reddit. Исходя из этого мы будем рассматривать только англоязычные тексты.

Для извлечения лингвистической информации, релевантной для предсказания ситуации на рынке криптовалют, мы проведём анализ тональности текстов и процедуру тематического моделирования. Мы не будем использовать полученные результаты в качестве признаков для предсказательных моделей временных рядов, так как данные модели применительно к выбранным нами криптовалют кажутся бессмысленными. Никакая модель не сможет предсказать изменение цены в тысячи раз по данным, которые на протяжении долгого периода времени оставались крайне мало изменичивыми и без ярко выраженного тренда, как например доджкоин, который существует и торгуется на биржах с 2013 г., а взрывной рост показал только в начале 2021 г. Для имеющих некоторый успех алгоритмов скальпинга (торговли на крайне малых временных отрезках) лингвистические данные малопригодны, так как торговля осуществляется за период гораздо меньший, чем тот, который понадобится для обработки лингвистической информации и передачи, ввиду сетевых задержек. Именно поэтому мы будем пытаться выявить корреляцию лингвистических данных с крупными сезонными трендами.

2.2. Построение англоязычного корпуса текстов социальной сети Reddit

На данный момент API Reddit не предоставляет доступ к историческим данным публикаций, поэтому мы использовали PushshiftAPI программной библиотеки для языка Python psaw[6], который предоставляет доступ к историческим данным социальной сети Reddit в базах ресурса pushshift.io. В данном пакете имеется два метода search_submissions() для поиска публикаций и search_comments() для поиска комментариев. Максимальный временной интервал, который мы можем указать в запросе к базе, является один месяц, поэтому для каждого сообщества сделано 12 запросов. Запрос возвращает список публикаций за указанный период, однако, он не содержит тексты комментариев: для отдельного поиска комментариев мы использовали числовой идентификатор публикации, и затем в корпус были добавлены тексты комментариев. Работа данного алгоритма заняла одну неделю непрерывного сбора текстов и комментариев.

Из текстов публикаций и комментариев к ним в Reddit за 2021 г. был собран корпус, состоящий из 10 отдельных рубрик (сабреддитов): Bitcoin, BTC, Crypto, Ethereum, ETH, Shibainucoin, shib, SHIBArmy, Dogecoin, DOGE. Рубрики представляют собой сообщества, в которых пользователи ведут дискуссию и делятся личным опытом по поводу темы сообщества. Десять сообществ имеют следующую тематическую направленность:

1) Bitcoin, BTC, Crypto – посвящены Bitcoin и рынку криптовалют в целом,
2) Ethereum, ETH – посвящены Ethereum,
3) Shibainucoin, shib, SHIBArmy – посвящены Shiba Inu,
4) Dogecoin, DOGE – посвящены Dogecoin.

[6] https://pypi.org/project/psaw/ (дата обращения: 01.04.2022)

На данном этапе мы произвели предобработку текстов: удалили гиперссылки, стопслова, избыточные пробелы, спецсимволы, а также никнеймы пользователей. После предобработки корпус содержит 1 177 235 уникальных токенов. По группам они распределены следующим образом:

1) группа текстов Bitcoin содержит 823 556 уникальных токенов,

2) группа текстов Ethereum содержит 154 384 уникальных токенов,

3) группа текстов Shiba Inu содержит 190 860 уникальных токена,

4) группа текстов Doge содержит 242 480 уникальных токена.

Корпус представляет из себя файл в формате json. Ключами верхнего уровня являются названия сообществ, внутри которых сохранены объекты с ключами-месяцами. Каждый такой объект содержит массив публикаций с полями id (числовой идетификатор публикации), date (дата публикации), comments (массив текстов комментариев) и text (текст публикации). Даты сохранены в наносекундах с 1970 года, что является общепринятой практикой в хранении данных. Структура выглядит следующим образом:


```
{
    "DOGE": [
        {
            "id": "kvm514",
            "date": 1610421818.0,
            "comments": [
            "1 DOGE = 1 DOGE. Doge trust. Moon!"
            ],
            "text": "I'm hunting different Doges trying collect different doges order write Holy Dogle, story sort mock religion elements religions characters doge asking type comments different names doge variations put link meme
```

```
                using templateDO SEND PHOTOSHOPPED DOGEI want memes like Karen doge kid doge, please put different photos Shiba Inu"
        },
        {
                "id": "kxesao",
                "date": 1610647578.0,
                "comments": 0,
            "text": "Hello everyone, I recently started making memes and stuff on this subreddit for the last 3 days, and I noticed that many people don’t post. Yet they upvote. I ask everyone to start making more memes and keep this community alive. Thanks for coming to my Ted-Talk."
        }
        ]
}
```


Количество публикаций в Reddit с учетом их тематики и времени их размещения представлено на графике (рис. 1). Мы можем видеть, что пользовательская активность в сообществах варьировалась на протяжении года. Однако, в сообществах относящихся к ETH пользовательская активность оставалась на одном уровне на протяжении года.

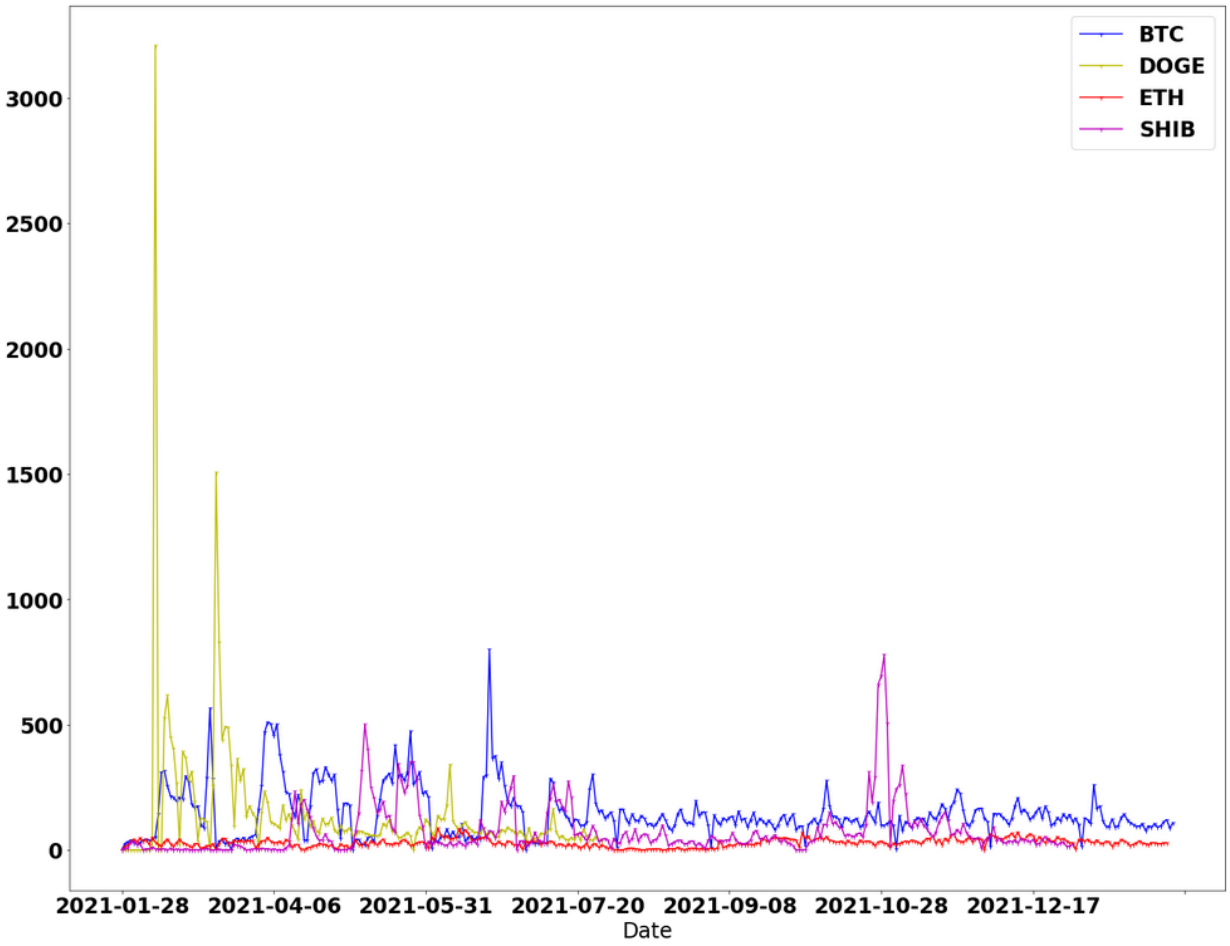

Рис. 1. *Количество публикаций в корпусе Reddit за 2021 год* с учетом их тематики и времени их размещения

Если мы взглянем на график (рис. 2), показывающий количество комментариев, то увидим, что в некоторые дни велось активное обсуждение в комментариях к отдельным публикациям. Например, если бы мы рассматривали исключительно тексты публикаций, то мы бы упустили резкий скачок в активности в сообществах, посвящённых шиба ину, который произошел 30 января 2021 г. Исходя из данных рассуждений, для дальнейшего удобства мы суммируем публикации и комментарии к ним.

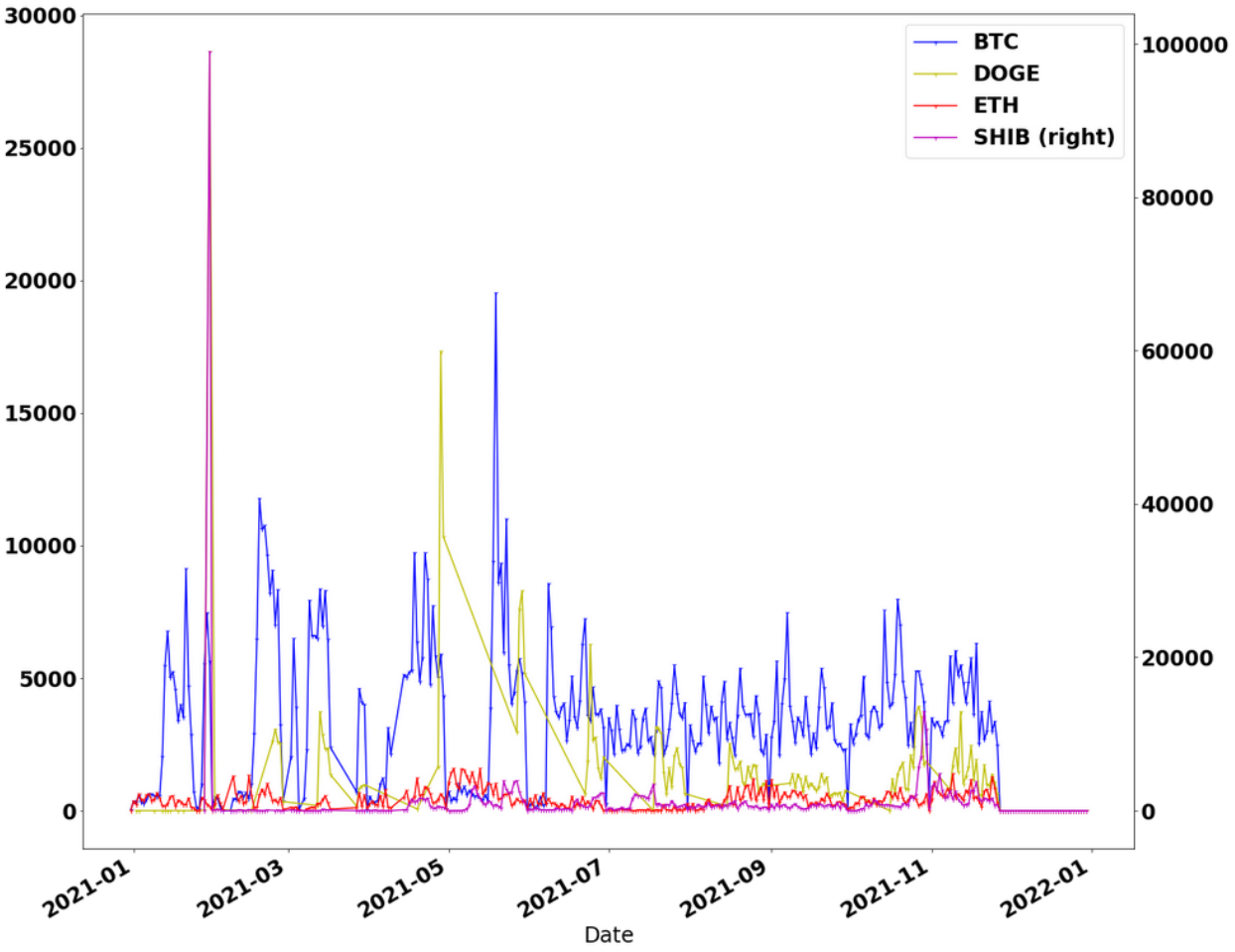

Рис. 2. Количество комментариев в корпусе Reddit за 2021 год с учетом их тематики и времени их размещения (2)

После изменений (объединение информации о постах и комментариях) график принял следующий вид (рис. 3).

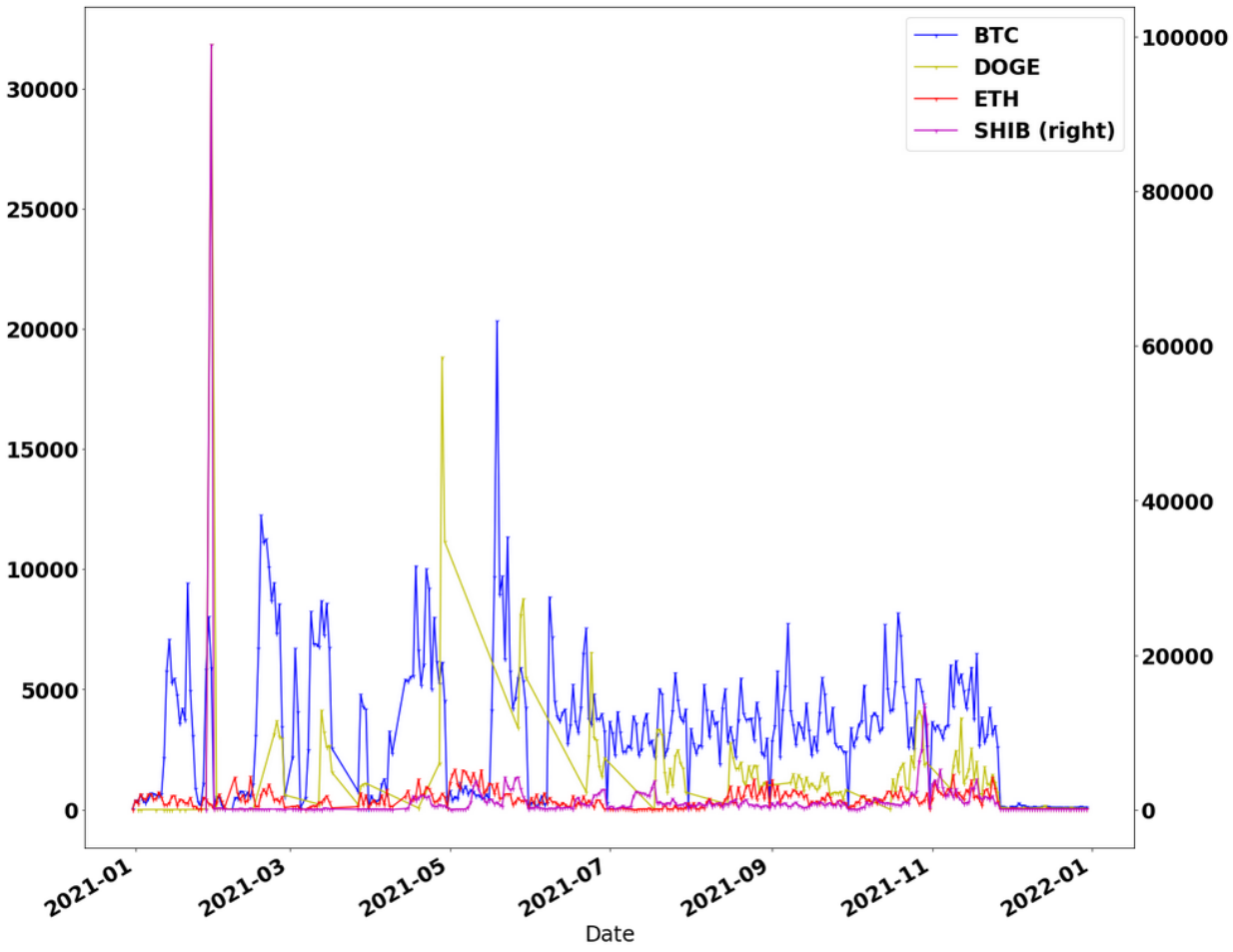

Рис. 3. *Количество публикаций и комментариев за 2021 год* с учетом их тематики и времени их размещения *(3)*

Теперь нам необходимо сравнить каждую из групп по отдельности с ценой на соответствующую криптовалюту и посчитать корреляцию количества публикаций и трёх показателей:

1) объём торгов,
2) изменение цены – высчитывается процентное соотношение разницы цены открытия и цены закрытия с ценой открытия,
3) волатильность – высчитывается разница наивысшей цены в день торгов и наименьшей цены в день торгов.

В качестве метода подсчёта корреляции мы будем использовать коэффициент корреляции Пирсона. Чтобы найти значение коэффициента

корреляции Пирсона, нам нужно убедиться, что данные по количеству публикаций соответствуют нормальному распределению, для этого мы используем тест Дагостина-Пирсона, который дал нам следующие результаты p-критерия:

1) группа сообществ BTC – $6.754924327483769 * 10^{-22}$
2) группа сообществ ETH – $1.1013141976106628 * 10^{-13}$
3) группа сообществ DOGE – $1.517576631836189 * 10^{-52}$
4) группа сообществ SHIB – $1.1611881778995035 * 10^{-152}$

Как видно из приведённых значений p-критерия данные распределены нормально, и мы можем рассчитывать коэффициент корреляции Пирсона.

2.1.1 Биткоин

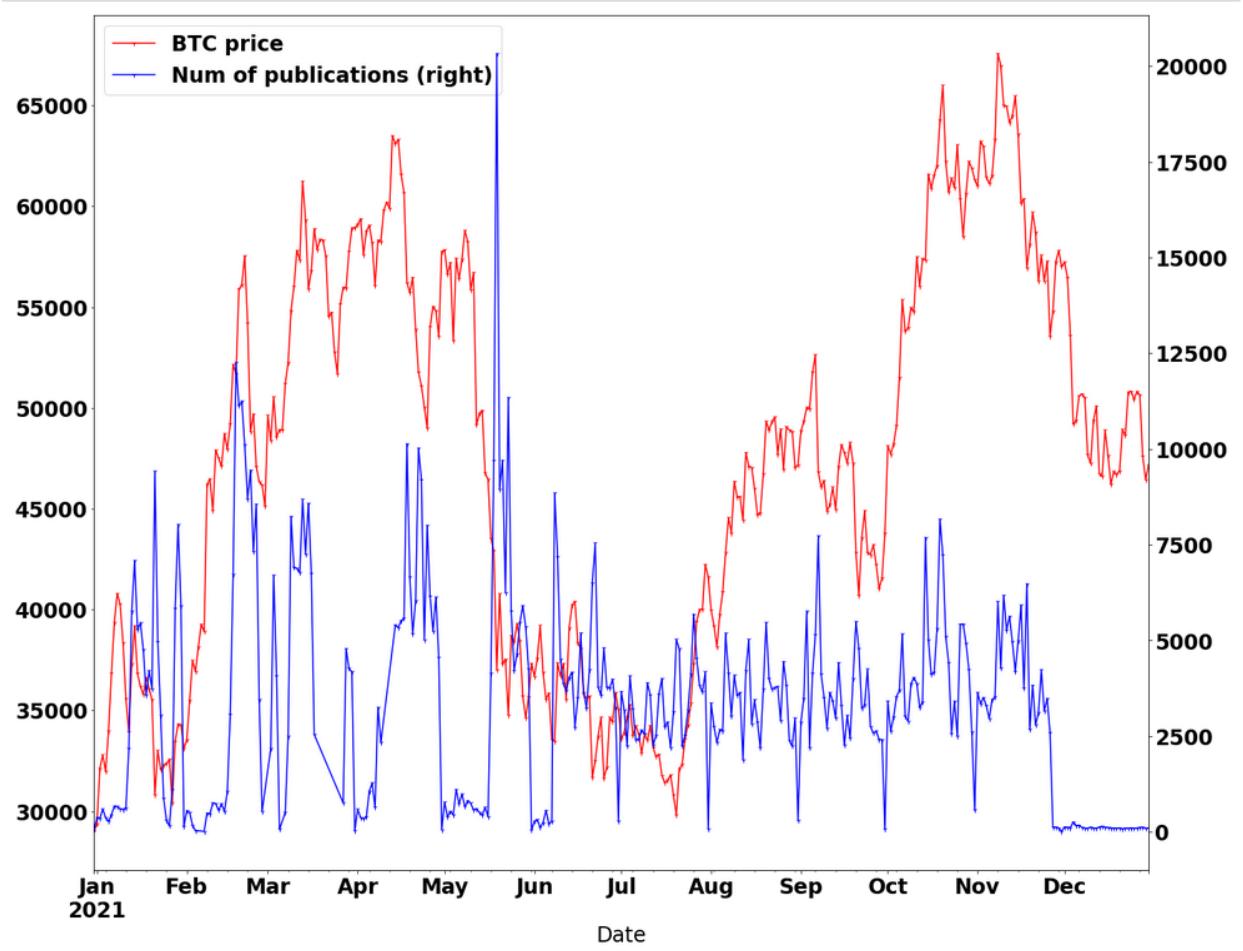

Рис. 4. Цена на биткоин и количество публикаций (4)

На графике (см. рис. 4) мы видим, что количество публикаций имело сильную волатильность в первую половину 2021 года до середины мая, далее количество публикаций уже не имеет таких резких скачков. Значения корреляций имеет следующим вид:

1) Сравнение с изменением объёма дало результат 0.174293
2) Сравнение с изменением цены дало результат -0.099005
3) Сравнение с волатильностью дало результат 0.374305

2.1.2 Эфириум

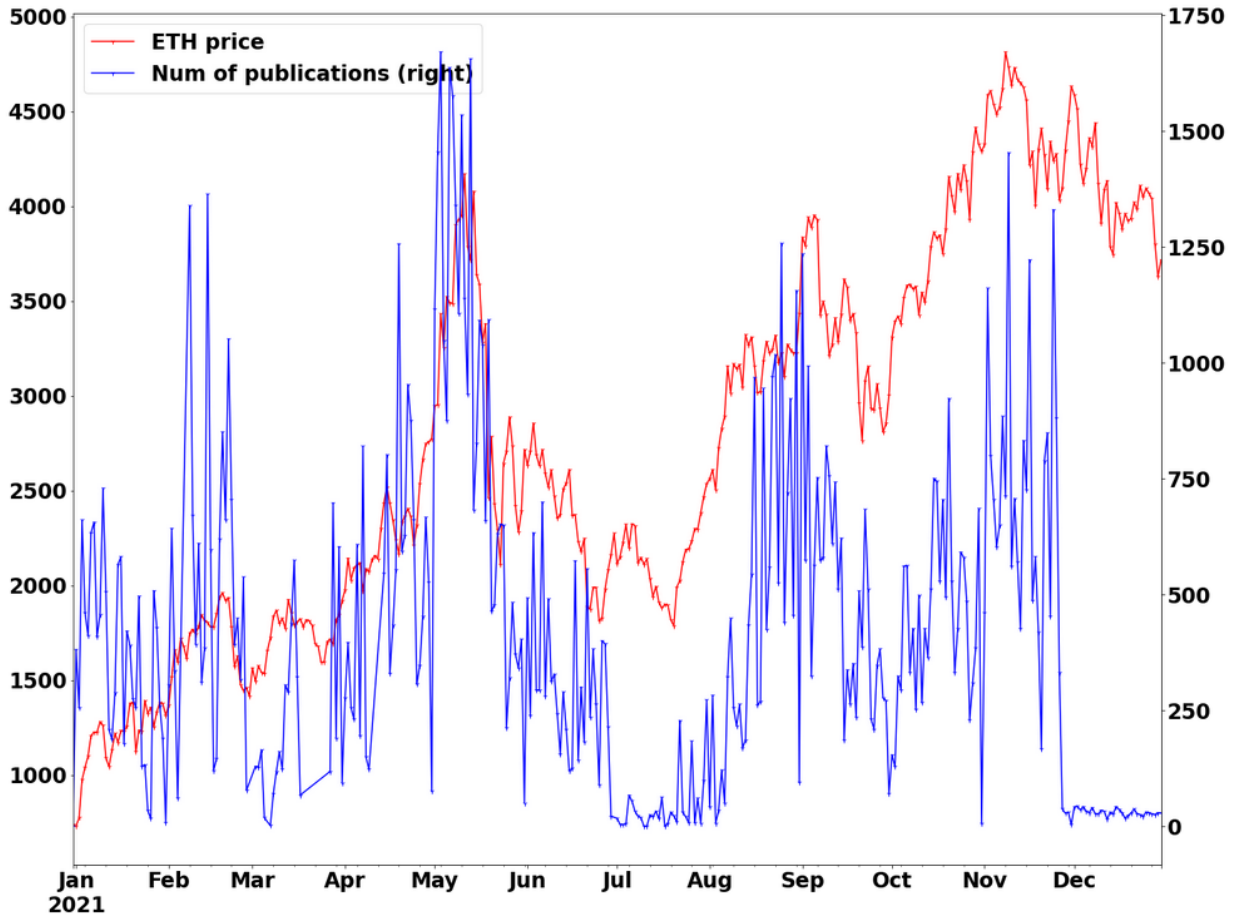

Рис. 5. *Цена на эфириум и количество публикаций (5)*

Несмотря на наименьшую пользовательскую активность в группах посвящённых данной криптовалюте, мы наблюдаем большее визуальное соответствие количества публикаций движению цены (ср. рис. 4 и 5), нежели в группах, посвященным биткиону - данное соответствие сохраняется на протяжении всего 2021 года. Пользовательская активность достигала максимальных значений вплоть до пика в мае 2021 года и падала вместе с

ценой и вновь росла вплоть до исторического максимума данной криптовалюты. Значения корреляций имеет следующим вид:

1) Сравнение с изменением объёма дало результат 0.362661
2) Сравнение с изменением цены дало результат 0.004127
3) Сравнение с волатильностью дало результат 0.285909

2.1.3 Доджкоин

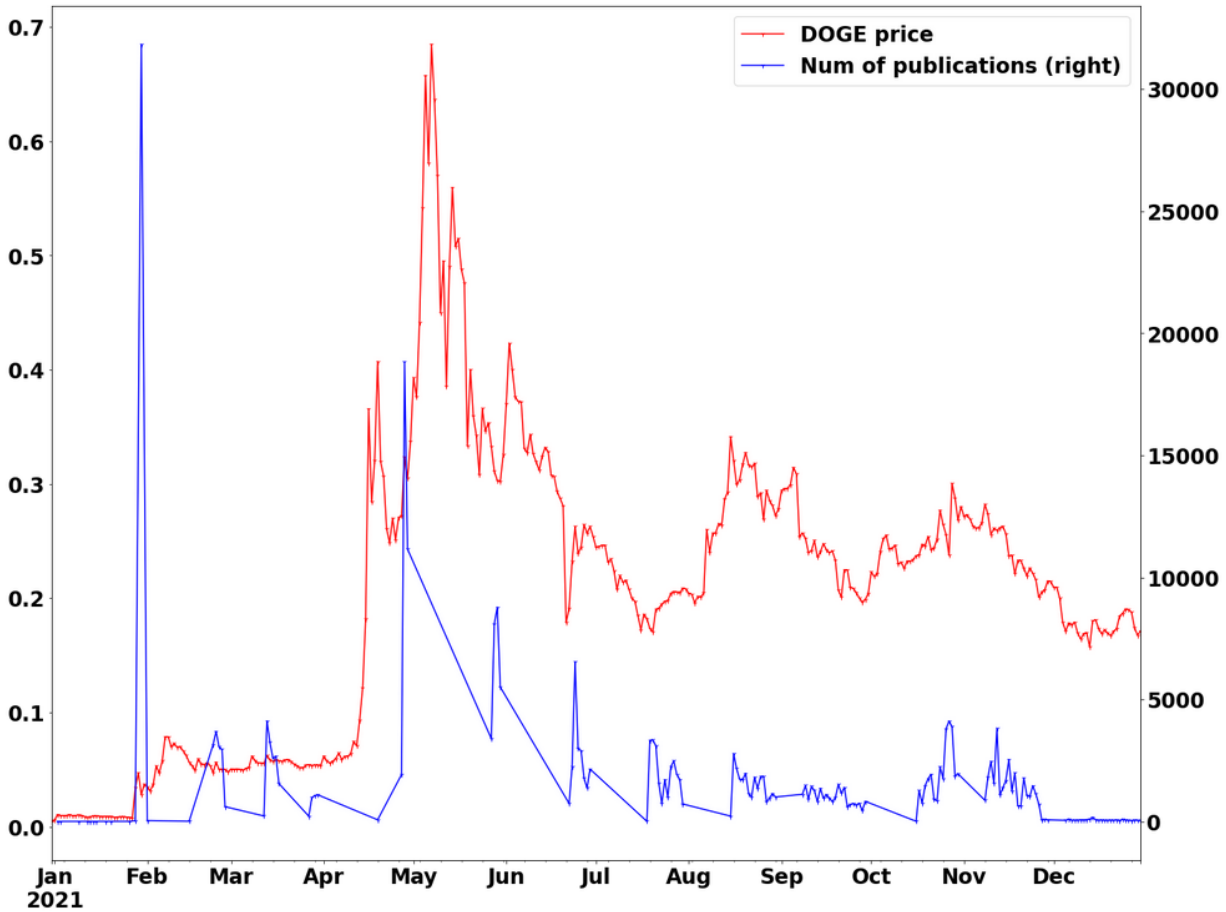

Рис. 6. Цена на доджкоин и количество публикаций (6)

На графике (см. рис. 6) мы видим, что резкий рост публикаций наблюдался в первой половине 2021 года. Пик пришёлся на 30 января, когда

за два дня до этого цена на доджкоин впервые в истории превысила 1 цент, что явилось поворотным моментом в истории данной криптовалюты. Следующий пик активности 28 апреля предшествовал историческому максимуму данной криптовалюты 8 мая. Значения корреляций имеет следующим вид:

1) Сравнение с изменением объёма дало результат 0.016995
2) Сравнение с изменением цены дало результат -0.081652
3) Сравнение с волатильностью дало результат 0.011723

2.1.4 Шиба ину

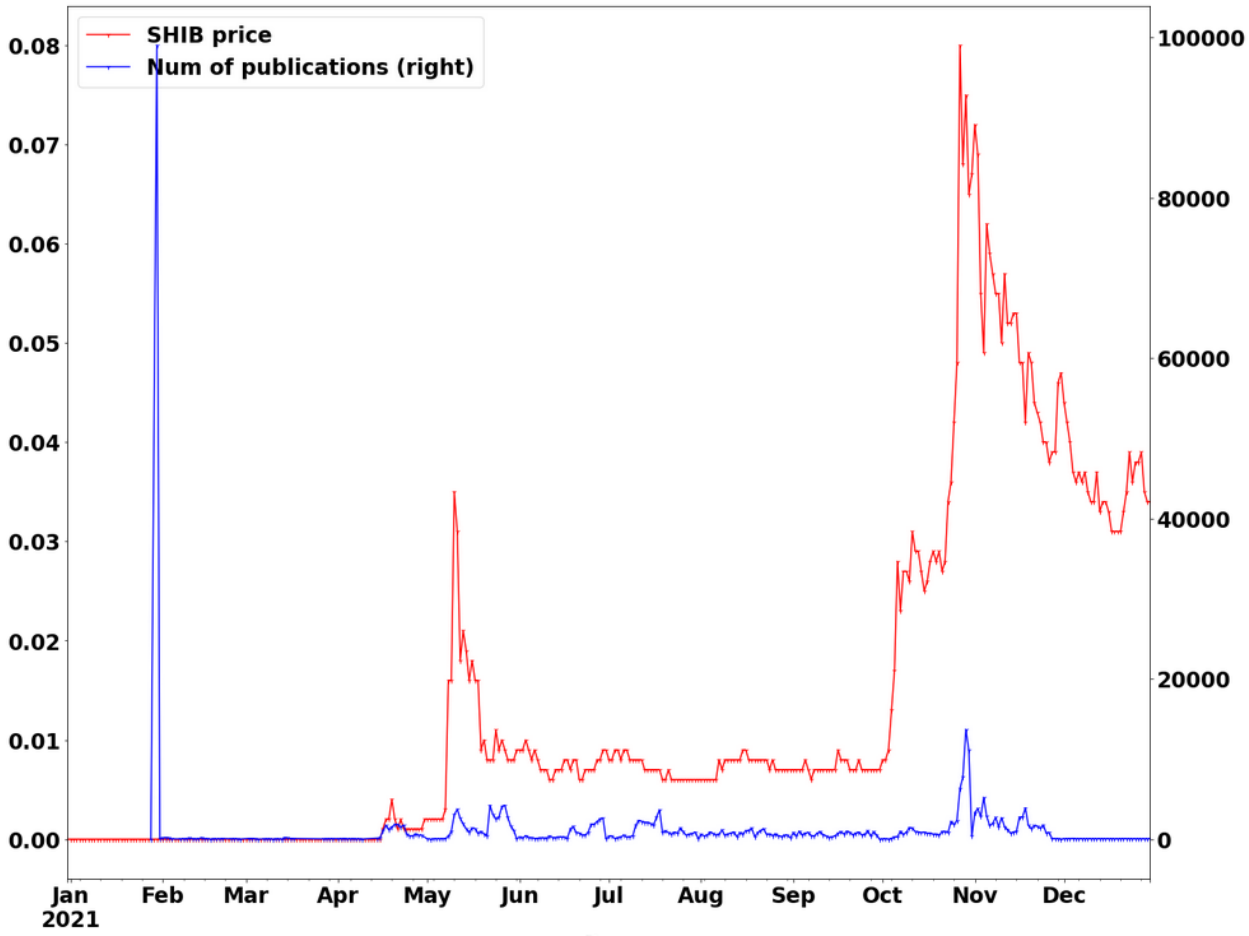

Рис.7. *Цена на шиба ину и количество публикаций (7)*

К сожалению, данных о цене на коин шиба ину имеется только с 1 мая, так как данная криптовалюта не имела реальной цены и не торговалась на официальных криптобиржах. Ради удобства визуализации графика (см. рис. 7) цена шиба ину была умножена на 1000. Мы видим, что пользователи вели обсуждения в сообществах Reddit ещё до официального листинга данной криптовалюты на популярных криптобиржах. Движение цены обнаруживает

малое визуальное соответствие изменению количества публикаций. Значения корреляций имеет следующим вид:

1) Сравнение с изменением объёма дало результат 0.016995
2) Сравнение с изменением цены дало результат -0.015153
3) Сравнение с волатильностью дало результат 0.011723

Итак, наш анализ пользовательской активности показал, что данный показатель не имеет корреляции с ценой на криптоактив. Однако, группа сообществ BTC и ETH имели результаты близкие к слабой корреляции для изменения объёма и волатильности: у BTC корреляция с волатильностью имеет значение 0.374305, а у ETH корреляция корреляция с изменением объёма имеет значение 0.362661 и с волатильностью 0.285909.

Выводы по главе 2

Из проведённого нами статистического анализа англоязычных публикаций социальной сети Reddit можно сделать такой вывод: количество публикаций в тематических сообществах социальной сети Reddit не имеет никакой корреляции с ценой на соответствующую криптовалюту.

Однако, как видно из примера с шиба ину, пользовательская поддержка криптовалюты может быть сильной несмотря на то, что данный криптоактив ещё не торгуется на криптобиржах и не имеет реальной стоимости в фиатной валюте.

Глава 3. Анализ тональности публикаций англоязычного корпуса текстов социальной сети Reddit

3.1. Алгоритм анализа тональности публикаций англоязычного корпуса текстов социальной сети Reddit

На основании выводов, сделанных в главе 1, для анализа тональности мы будем использовать гибридный подход, состоящий в следующем: мы совместим обращение к словарю тональности финансовой лексики NTUSD-fin[7] и инструменту VADER[8] программной библиотеки для языка Python NLTK. VADER рассчитает общую тональность текстов, а словарь мы будем использовать для корректировки оценок, так как специфичную для криптовалют лексику VADER расценивает как нейтральную: слова *moon*, *bullish*, *pump* и т.д. Также как и для анализа пользовательской активности, мы будем использовать коэффициент корреляции Пирсона.

Наш алгоритм анализа тональности состоит из двух шагов:

1) Получение оценок с помощью VADER.
2) Дополнительная оценка для финансовой лексики и эмоджи с помощью словаря NTUSD-fin.

Перед запуском алгоритма необходимо поменять структуру корпуса. Мы объедим отдельные сообщества в группы, чтобы сохранить их в отдельные файлы json. Каждый файл содержит джейсон объект с ключами-датами и в качестве значения тексты всех публикаций и комментариев за указанную дату.

[7] http://nlg.csie.ntu.edu.tw/nlpresource/NTUSD-Fin/
[8] https://www.nltk.org/_modules/nltk/sentiment/vader.html

На первом шаге алгоритма мы разбиваем текст за каждую дату на отдельные предложения используя метод sent_tokenize() пакета NLTK. Далее, алгоритм в цикле проходит по каждому предложению за указанную дату и передаёт предложение в метод polarity_scores() класса VADER. Данный метод возвращает положительные и отрицательные оценки, данного предложения, а также сумму этих оценок. Мы берём данные сумму и прибавляем её к общему значению тональности за обрабатываемую дату. Результат работы сохраняется в файл формата csv.

На втором шаге алгоритма мы повторно проходим в цикле по предложениям текста, токенизируя их. Затем в цикле проходим по каждому токену предложения, проверяя его наличие в словаре NTUSD-fin. При нахождении слова или эмиоджи в словаре, мы добавляем оценку словаря к общем оценке тональности предложения, как это было сделано на предыдущем шаге. Отдельно следует отметить, что оценки VADER и оценки NTUSD-fin значительно отличаются друг друга, если первый выдаёт результат в диапозоне от -1 до 1, то оценки NTUSD-fin могут иметь значения превышающие 1 или -1. Для приведения в соответствие оценок мы нормализуем оценку словаря. Нормализация происходит по следущей формуле:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

где X - текущая оценка токена, X_{min} - минимальная оценка в словаре, X_{max} - максимальная оценка в словаре.

По аналогии с предыдущим шагом мы высчитываем общую оценку для предложения, используя только токены найденные в словаре, и обновляем

оценку в ранее сохранённом csv файле. Файл состоит из трёх колонок - дата, оценка и соответсвующая криптовалюта:

```
Date,sentiment_score,Subreddit
2021-01-12,199,BTC
2021-01-13,487,BTC
2021-01-14,479,BTC
2021-01-15,387,BTC
2021-01-16,335,BTC
2021-01-17,321,BTC
2021-01-18,297,BTC
2021-01-19,340,BTC
2021-01-20,314,BTC
2021-01-21,422,BTC
```

3.2. Результаты работы алгоритма анализа тональности публикаций англоязычного корпуса текстов социальной сети Reddit

Вышеописанный алгоритм дал нам следующие результаты, которые мы проиллюстрируем в виде графика (см. рис. 8):

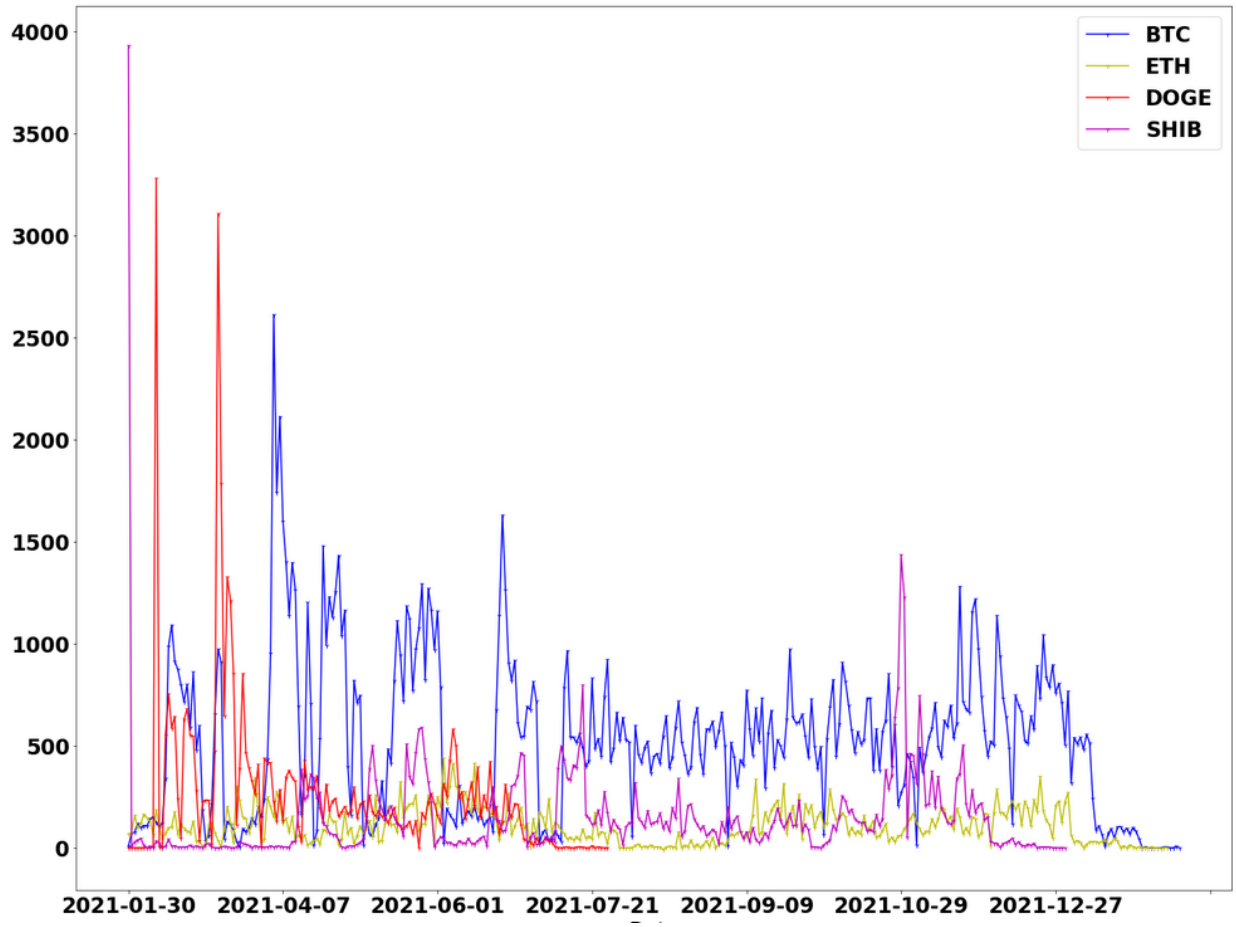

Рис. 8. *Сравнение тональности публикаций публикаций англоязычного корпуса текстов социальной сети Reddit (8)*

В целом, можно заключить, что криптосообщество было настроено крайне оптимистично: на протяжении всего года не наблюдается ни одного дня с отрицательной тональностью публикаций. Необходимо отметить, что тональность сильно варьировалась на протяжении года. Видно, что

доджкоины и шиба ину имели сильную поддержку пользователей в начале года, пользовательская активность была больше, чем в сообществах посвящённых биткоину и эфириуму.

Рассмотрим отдельные примеры, чтобы понять, как отработал наш алгоритм. Возьмём текст публикации в группе DOGE за 30 января:

"lets make dogecoin next bitcoin 🚀 buy tweet#doge#dogecoin#elondoge#elonbuydogecoin"

Наш алгоритм присудил оценку 4.5 данному сообщению. Оно очевидно положительное и даже содержит конкретный маркер - эмиоджи ракеты 🚀.

Рассмотрим другой пример из группы текстов BTC за 16 января:

“bear market always pretty bitchy time around crypto space matter bull markets toxicity ie. sensationalism, exuberance lost sense reality”

Наш алгоритм успешно определил негативное отношение пользователя и присудил оценку -6.81.

Прежде, чем приступить к сопоставлению тональности и цены наших криптовалют необходимо проверить полученные данные на нормальность распределения:

1) p-критерий для BTC имеет значение $4.473340393729717 * 10^{-16}$
2) p-критерий для ETH имеет значение $9.7413856052746 * 10^{-10}$
3) p-критерий для DOGE имеет значение $2.7523135246882634 * 10^{-42}$
4) p-критерий для SHIB имеет значение $6.595441892199395 * 10^{-109}$

Из приведенных выше значений мы делаем вывод, что данные распределены нормально и мы можем рассчитывать коэффициент корреляции Пирсона. Аналогично анализу количества публикаций мы будем

рассчитывать корреляцию тональности с объёмом торгов, изменением цены и волатильностью.

3.2.1. Биткоин

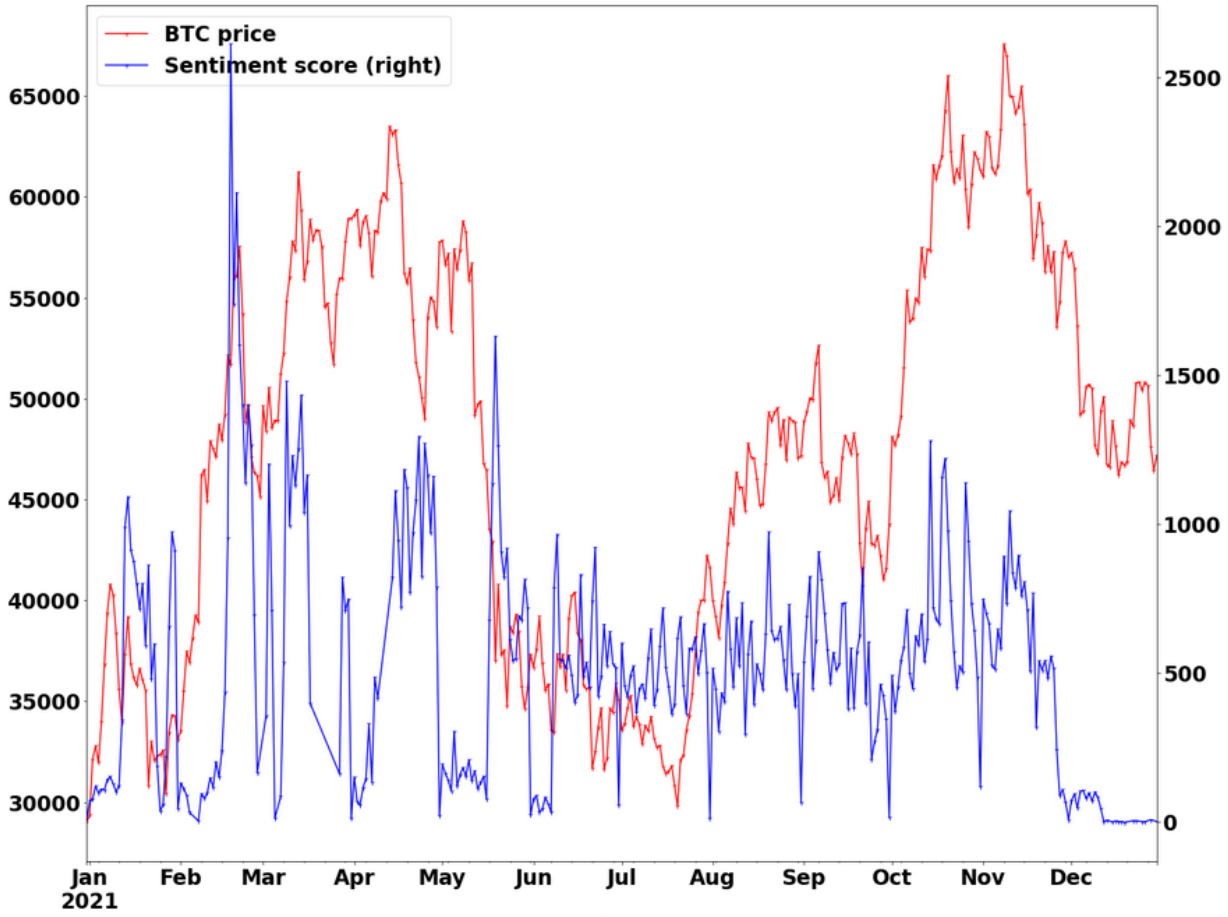

Рис. 9. Оценки тональности и цены биткоина (9)

Как видно на графике (см. рис. 9) явная корреляция между тональностью публикаций и цены биткоина не обнаруживается. В период с января по февраль тональность проявляется как реакция на изменение цены. Далее, имеет слабую реакцию на февральский рост цены и предшествует первому пику цены 2021 года. Однако, далее уже не наблюдается какой-либо

зависимости между тональностью и ценой. Значения корреляций имеет следующим вид:

1. Сравнение с изменением объёма дало результат 0.124706
2. Сравнение с изменением цены дало результат -0.018386
3. Сравнение с волатильностью дало результат 0.224478

3.2.2 Эфириум

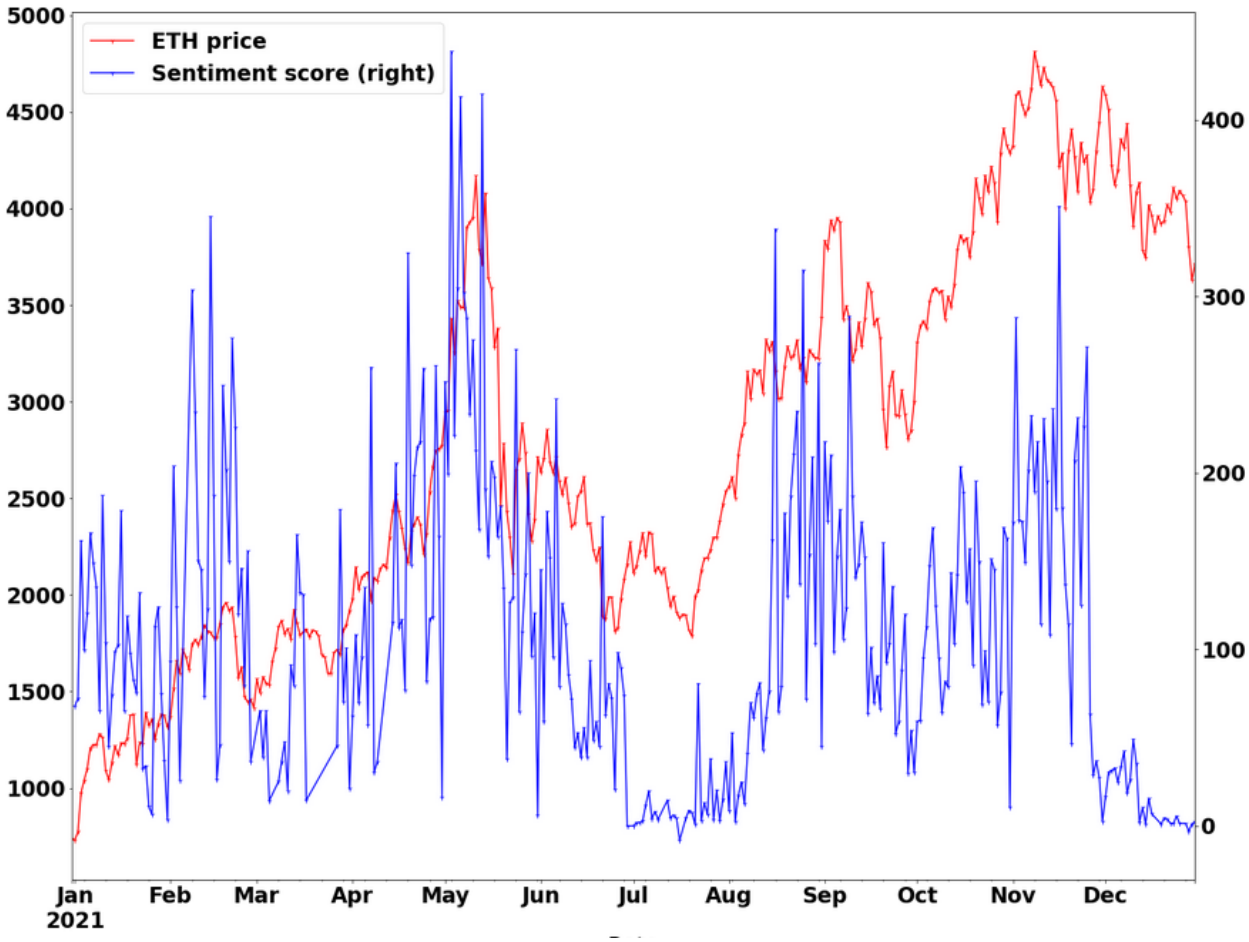

Рис. 10. *Оценки тональности и цены эфириума (10)*

График (см. рис. 10) показывает совсем иное поведение кривой тональности нежели у биткоина. На графике присутствуют множественные хаотичные всплески изредка прерывающиеся практически нейтральными месяцами, как в период с июль по август или в декабре месяце. Самые

большие оценки тональности присутствуют в период с начала мая по середину мая, когда цена впервые в истории превысила отметку в 4000 тысячи долларов за монету. Однако, когда цена обновляет исторический максимум в ноябре, кривая тональности не повторяет предыдущего всплеска. Значения корреляций имеет следующим вид:

1. Сравнение с изменением объёма дало результат 0.338299
2. Сравнение с изменением цены дало результат 0.036591
3. Сравнение с волатильностью дало результат 0.237855

3.2.3. Доджкоин

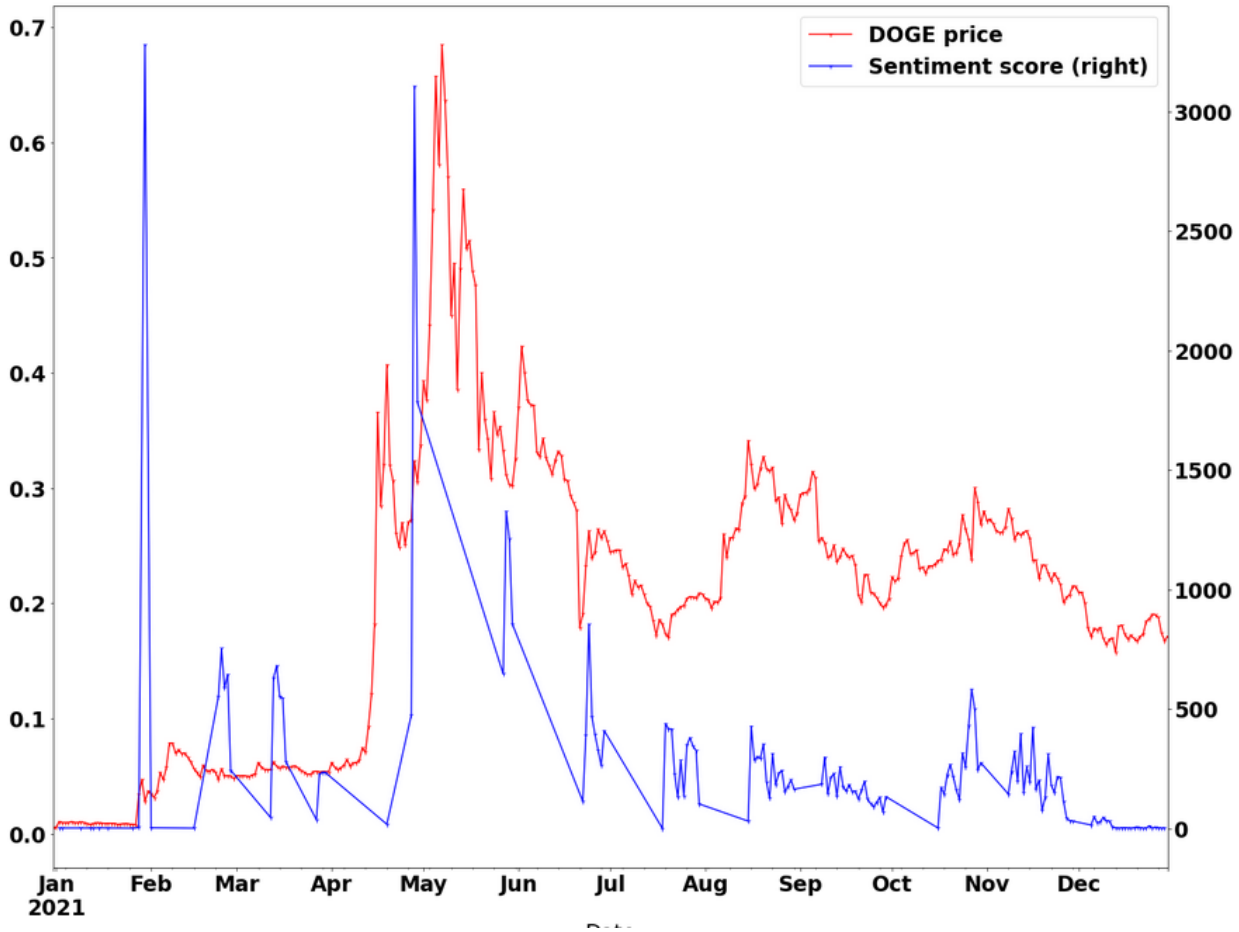

Рис. 11. Оценки тональности и цены доджкоина (11)

График оценки тональности (10) имеет сильное сходство с графиком количества публикаций. Здесь сложно разглядеть какую-либо зависимость - в первом полугодии данные кажутся запоздавшей реакцией на изменение цены и далее не показывают закономерных всплесков. Значения корреляций имеет следующим вид:

1. Сравнение с изменением объёма дало результат 0.006683
2. Сравнение с изменением цены дало результат -0.065837
3. Сравнение с волатильностью дало результат 0.00555

3.2.4. Шиба ину

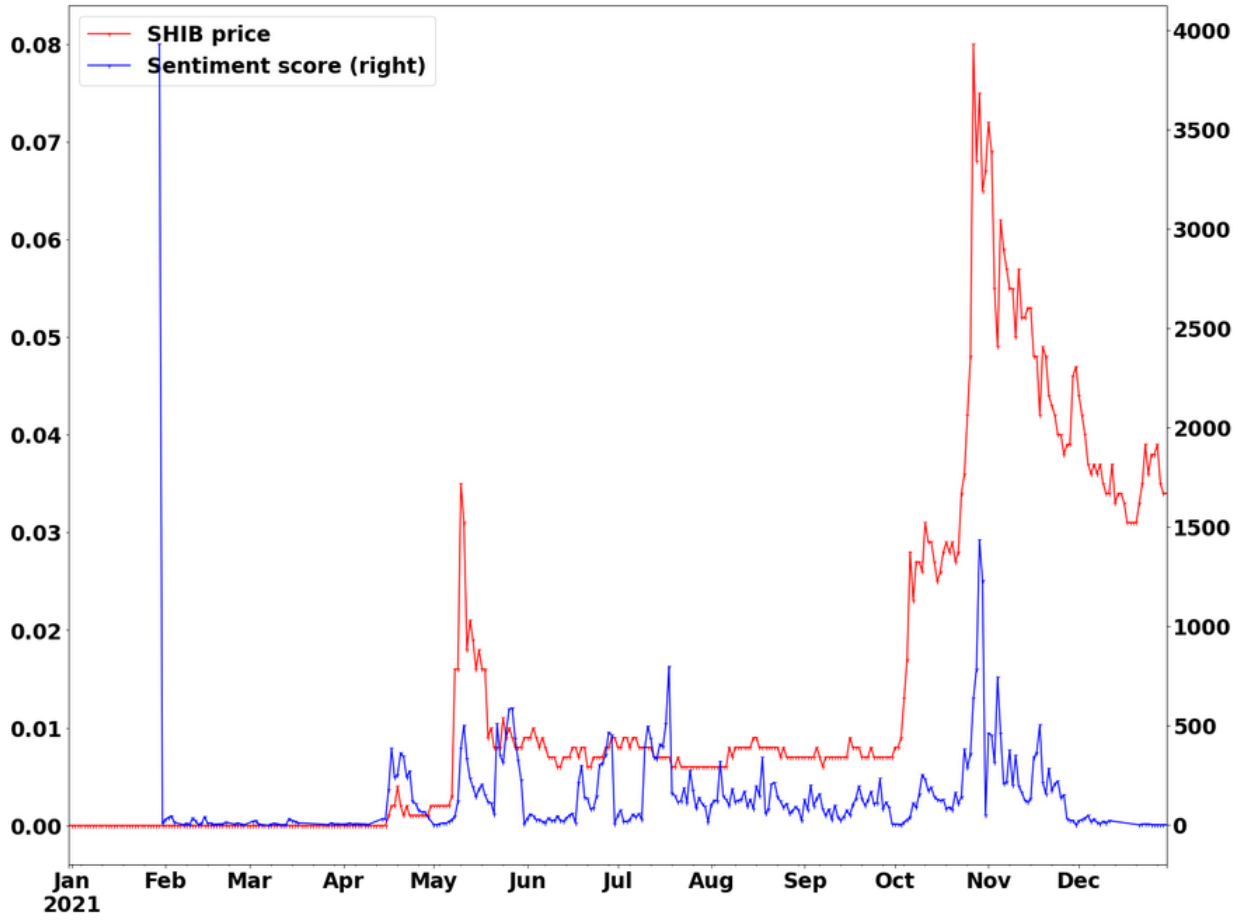

Рис. 12. Оценки тональности и цены шиба ину (12)

График тональности шиба ину (см. рис. 12) показывает лучшее соответствие изменению цены нежели график тональности доджкоина. Кривая тональности визуально соответствует крупным ростам цены. Значения корреляций имеет следующим вид:

1. Сравнение с изменением объёма дало результат 0.293025
2. Сравнение с изменением цены дало результат -0.028602
3. Сравнение с волатильностью дало результат 0.30833

Итак, наш анализ тональности показал, что существует некоторая корреляция между тональностью публикаций и цены. У всех криптовалют наблюдались высокие значения тональности в первом полугодии. Вопреки ожиданиям, корреляция между тональностью публикаций и цены не была обнаружена у доджкоина, как мемкоина, у которого в 2021 году ничего не имелось кроме поддержки пользователей, и наоборот была выявлена некоторая корреляция у эфириума, который имеет сильные перспективы войти в повседневную жизнь общества благодаря своим технологическим нововведениям. Другой мемкоин шиба ину показал достаточно хорошую корреляцию в сравнении с другими криптовалютами. Анализ тональности сообществ, посвящённых биткоину, ожидаемо не обнаружил корреляции, так как данная криптовалюта имеет интерес среди крупных инвестиционных фондов и даже государств, поэтому активность пользователей Reddit никак не влияет на цену биткоина. Лучшие показатели корреляции:

1) Эфириум - корреляция с изменением объёма равная 0.338299
2) Шиба ину - корреляция с изменением объёма равная 0.293025 и волатильностью равная 0.30833

Мы видим, что явной сильной корреляции между тональностью публикаций и движением цены не обнаруживается.

Выводы по главе 3

Итак, на основании исследования, проведённого в главе 3, можно сделать следующие выводы: тональность публикаций имеет разную корреляцию по отношению к изменению объёма, цены и волатильности в зависимости от криптовалюты. У эфириума выявлена корреляция слабой силы с изменением объёма, у шиба ину выявлена слабая корреляция с объёмом и волатильностью.

Глава 4. Тематическое моделирование англоязычного корпуса текстов социальной сети Reddit

4.1. Подготовка данных и инструментов для тематического моделирования англоязычного корпуса текстов социальной сети Reddit

Для тематического анализа мы будем применять алгоритм латентного размещения Дирихле (LDA) и языковую модель BERTopic. Также BERTopic будет использован для динамического тематического моделирования. Для LDA[9] нами будет использован модуль gensim, в котором уже реализована модель LDA, нам только потребуется подготовить тексты корпуса и обучить модель. Модель BERTopic [10]имеет реализацию в виде отдельного питоновского модуля.

Для упрощения работы на данном этапе исследования нам понадобится несколько изменить структуру нашего корпуса, а именно: объединить все тексты за конкретный день в один большой документ. После преобразования наш корпус состоит из ключа-даты и массива текстов публикаций за конкретную дату.

На данном этапе также следует сделать следующую оговорку. В работе [Blei 2012] внимания заслуживает нормализация данных. Из процентного изменения цены на тикер извлекалось процентное изменение цены индекса S&P, что исключало влияние общего настроения на рынке. Данное замечание является полезным в нашей работе, так как подобным индикатором настроения на рынке криптовалют является цена на биткоин, из чего можно сделать вывод, что при одномоментном росте биткоина и какого-либо

[9] https://radimrehurek.com/gensim/models/ldamulticore.html (дата обращения: 30.04.2022)
[10] https://pypi.org/project/bertopic/ (дата обращения: 15.05.2022)

альткоина, данный рост с большой долей вероятности вызван ростом цены биткоина.

Процедура тематического моделирования с помощью алгоритма LDA включает в себя следующую структуру.

1. Мы используем список стоп-слов для английского языка, который имеется в пакете NLTK[11]. Так как наши тексты уже априори имеют epre. тематическую направленность, то нам необходимо также убрать слова? связанные с очевидной тематикой торгов на рынке криптовалют. Для этого мы расширяем наш список стоп-слов словами, содержащимися в слове оценочной лексики NTUSD-fin, который мы использовали для анализа тональности. Таким образом, мы оставим больше лексики, которая потенциально может отражать внешнюю по отношению к криптовалютам информацию, но при этом присутствующую в наших текстах, что говорит об её актуальности.

2. Мы будем принимать во внимание только те дни, когда рост криптоактива превысил 10% за день для всех 4 криптовалют. Для альткоинов мы дополнительно вычитаем дни, когда такой сильный рост показывал биткоин. Также дополнительно мы сохраняем процентное изменение цены за день − это значение поможет нам при анализе.

3. Ещё одним полезным в анализе показателем будут оценка тональности за конкретный день и средняя оценка тональности у отдельного криптоактива. Среднее значение нам необходимо, чтобы сравнивать содержание тем с текущим значением тональности.

4. На данном шаге мы производим предобработку текста. Стоп-слова удаляются из текста, текст токенизируется.

[11] https://www.nltk.org/book/ch02.html (дата обращения: 02.05.2022)

5. Затем из токенизированного объекта создаётся корпус и словарь для обучения модели LDA.

6. Для обучения модели передаётся созданный на предыдущем шаге словарь и корпус. После настройки параметров тематического моделирования и в результате подбора оптимальных значений метрик когерентности c_v и U_mass мы остановились на 5 темах и 15 итерациях обработки корпуса в момент обучения. На выходе мы получаем по 10 слов на каждую тему.

7. Результат работы для каждого дня сохраняется вместе с оценкой тональности, средней оценкой тональности и изменением цены.

Для алгоритма BERTopic последовательность действий выглядит несколько иначе − мы экспериментально установили, что модель успешно группирует слова по темам даже без удаления стоп-слов. Благодаря реализации алгоритма в виде библиотеки для языка Python не требуется никаких дополнительных манипуляций с данными, поэтому шаги 4-6 из схемы для LDA не используются.

Наш алгоритм тематического моделирования с помощью BERTopic состоит из двух шагов.

1. В качестве отдельного документа рассматриваются текст публикации и комментарии к ней. Данные корпуса уже готовы для работы с выбранной нами моделью. Наши эксперименты показали, что BERTopic не нуждается в удалении стоп-слов или удалении пунктуации. На данном этапе мы определяем темы для всего корпуса конкретной группы сообществ Reddit за год (стандартная тематическая модель).

2. Далее, мы высчитываем изменение регулярности регистрации каждой темы на протяжении всего 2021 года (динамическая тематическая модель).

4.2. Анализ результатов тематического моделирования с помощью алгоритма LDA

В тематические модели, описывающие ситуацию на рынке криптовалют, вошли публикации за 6 дней для BTC, 11 дней для ETH, 14 дней для DOGE и 44 дня для SHIB.

4.2.1. Биткоин

Особое внимание мы уделили темам со словами, описывающими внешние факторы, которые могут влиять на ситуацию на рынке криптовалют. Рассмотрим некоторые примеры тем, связанных с сообществом BTC.

Первый пример содержит упоминания Илона Маска 8 февраля 2021 года. Действительно, в этот день Илон Маск публиковал сообщения о доджкоине, что нашло отражение и в сообществе Reddit с другой тематикой. Темы выглядят следующим образом:


```
        [
                0,
                "0.027*\"bch\" + 0.021*\"per\" +
0.017*\"cash\" + 0.015*\"nodes\" + 0.013*\"ln\" +
0.013*\"big\" + 0.012*\"nano\" + 0.012*\"seed\" +
0.012*\"away\" + 0.009*\"node\""
            ],
            [
                1,
                "0.019*\"subreddit\" + 0.015*\"bch\"
+ 0.013*\"badger\" + 0.013*\"egon\" + 0.010*\"find\" +
0.010*\"post\" + 0.010*\"blockstream\" + 0.010*\"mod\"
+ 0.008*\"long\" + 0.008*\"ver\""
            ],
            [
                2,
```

```
                "0.035*\"bch\" + 0.016*\"ntp\" +
0.012*\"low\" + 0.010*\"huge\" + 0.007*\"elon\" +
0.007*\"bit\" + 0.007*\"bro\" + 0.007*\"registrar\" +
0.007*\"domain\" + 0.007*\"stay\""
            ],
            [
                3,
                "0.117*\"bch\" + 0.029*\"cash\" +
0.018*\"nano\" + 0.017*\"gold\" + 0.015*\"long\" +
0.012*\"many\" + 0.011*\"pay\" + 0.008*\"elon\" +
0.008*\"low\" + 0.007*\"hold\""
            ],
            [
                4,
                "0.022*\"restore\" + 0.015*\"app\" +
0.015*\"post\" + 0.012*\"tx\" + 0.012*\"seed\" +
0.012*\"two\" + 0.012*\"icloud\" + 0.012*\"encrypted\"
+ 0.008*\"sign\" + 0.008*\"cash\""
            ]
```

В данный день оценка тональности имела очень низкий показатель 95 в сравнении со средним 524. Цена закрылась на 18% дороже цены открытия.

В другом примере на 7 сентября 2021 года мы получили следующий список тем:

```
        [
                0,
                "0.014*\"statistical\" +
0.007*\"layer\" + 0.006*\"long\" + 0.005*\"per\" +
0.005*\"dca\" + 0.004*\"pay\" + 0.004*\"cost\" +
0.004*\"tx\" + 0.004*\"hold\" + 0.003*\"run\""
            ],
            [
                1,
```

```
                "0.038*\"el\" + 0.035*\"salvador\" +
0.012*\"long\" + 0.011*\"big\" + 0.010*\"usd\" +
0.010*\"many\" + 0.008*\"news\" + 0.007*\"hold\" +
0.005*\"pay\" + 0.005*\"cash\""
            ],
            [
                2,
                "0.012*\"el\" + 0.008*\"salvador\" +
0.008*\"many\" + 0.007*\"node\" + 0.007*\"usd\" +
0.007*\"long\" + 0.006*\"custodial\" + 0.006*\"bit\" +
0.005*\"nodes\" + 0.005*\"big\""
            ],
            [
                3,
                "0.017*\"bch\" + 0.006*\"pay\" +
0.005*\"hold\" + 0.004*\"cash\" + 0.004*\"app\" +
0.004*\"post\" + 0.003*\"ever\" + 0.003*\"byte\" +
0.003*\"per\" + 0.003*\"pool\""
            ],
            [
                4,
                "0.037*\"el\" + 0.032*\"salvador\" +
0.011*\"long\" + 0.011*\"news\" + 0.009*\"many\" +
0.008*\"big\" + 0.007*\"usd\" + 0.006*\"tender\" +
0.006*\"ago\" + 0.006*\"pay\""
            ]
```

В темах 1, 2 и 4 мы видим упоминание государства Сальвадор. Именно 7 сентября Сальвадор объявил биткоин легальным способом оплаты на своей территории. Оценка тональности равна 813 при среднем уровне в 524. Однако, это не имело никакого влияние на цену и в этот день цена биткоина закрылась на отметке на 11 % ниже цены открытия. В данном случае падение цены объясняется сезонностью – в сентябре традиционно на рынке преобладают «медведи».

4.2.2. Эфириум

Многие темы в группе ETH также не отражают состояния на рынке криптовалют, но имеются и положительные примеры. Так, 2 февраля 2021 г. мы находим упоминание другой криптовалюты − кардано, тикер ADA. Список тем выглядит следующим образом:

```
     [
                    0,
                    "0.041*\"gas\" + 0.024*\"tx\" +
0.012*\"game\" + 0.009*\"apy\" + 0.009*\"workaround\" +
0.009*\"team\" + 0.009*\"claymore\" + 0.006*\"hi\" +
0.006*\"run\" + 0.006*\"defi\""
                ],
                [
                    1,
                    "0.043*\"cardano\" + 0.030*\"gas\" +
0.016*\"ada\" + 0.013*\"tokens\" + 0.013*\"layer\" +
0.012*\"defi\" + 0.010*\"rollups\" + 0.009*\"native\" +
0.009*\"many\" + 0.009*\"uniswap\""
                ],
                [
                    2,
                    "0.015*\"cache\" + 0.012*\"node\" +
0.012*\"nano\" + 0.012*\"mew\" + 0.012*\"staking\" +
0.009*\"long\" + 0.009*\"app\" + 0.006*\"bit\" +
0.006*\"celsius\" + 0.006*\"nodes\""
                ],
                [
                    3,
                    "0.071*\"gas\" + 0.017*\"murall\" +
0.016*\"cardano\" + 0.016*\"gwei\" + 0.016*\"txn\" +
0.014*\"tx\" + 0.012*\"pay\" + 0.010*\"mind\" + 0.008*\"defi\" +
0.008*\"long\""
                ],
                [
                    4,
```

```
                    "0.092*\"gas\" + 0.021*\"tx\" +
0.019*\"staking\" + 0.019*\"pay\" + 0.017*\"uniswap\" +
0.015*\"many\" + 0.011*\"long\" + 0.011*\"cost\" +
0.011*\"gwei\" + 0.009*\"defi\""
                ]
```

В темах 1 и 3 мы видим упоминание cardano и ada. Ко 2 февраля 2021 г. данная криптовалюта за месяц увеличилась в цене 3 раза и начиная со 2 февраля росла до 15 мая прибавив в цене 600%. В данный день эфириум подорожал на 10%. Значение тональности равно 200 при среднем показателе 108.

Следующий пример приходится на 23 февраля 2021 г. На этот раз упоминается проект Chainlink, чей токен линк, тикер LINK, обрушился на 17%, и далее рос в цене вплоть до 10 мая, подорожав в 2 раза. Темы выглядят следующим образом:

```
          [
                    0,
                    "0.039*\"gas\" + 0.023*\"layer\" +
0.023*\"app\" + 0.016*\"addresses\" + 0.013*\"many\" +
0.013*\"tezos\" + 0.011*\"erc\" + 0.011*\"mine\" +
0.011*\"seed\" + 0.008*\"loopring\""
                ],
                [
                    1,
                    "0.046*\"gas\" + 0.024*\"pay\" +
0.020*\"long\" + 0.012*\"dapps\" + 0.012*\"loopring\" +
0.012*\"cz\" + 0.011*\"cost\" + 0.011*\"defi\" + 0.010*\"many\"
+ 0.010*\"centralized\""
                ],
                [
                    2,
                    "0.046*\"gas\" + 0.017*\"staking\" +
0.017*\"chainlink\" + 0.012*\"key\" + 0.012*\"rediculous\" +
0.010*\"nft\" + 0.010*\"centralized\" + 0.010*\"pay\" +
0.010*\"dai\" + 0.008*\"low\""
```

```
                ],
                [
                    3,
                    "0.024*\"subreddit\" + 0.018*\"moderate\" +
0.018*\"mod\" + 0.015*\"mist\" + 0.015*\"modmail\" +
0.012*\"long\" + 0.012*\"argent\" + 0.012*\"sub\" +
0.009*\"phrases\" + 0.009*\"mnemonic\""
                ],
                [
                    4,
                    "0.020*\"gwei\" + 0.017*\"pay\" +
0.017*\"staking\" + 0.013*\"usd\" + 0.013*\"form\" +
0.013*\"node\" + 0.010*\"big\" + 0.010*\"gas\" + 0.010*\"find\"
+ 0.010*\"true\""
                ]
```

Информация о проекте Chainlink содержится в теме 2. Цена на сам эфириум упала на 11% и имела оценку тональности 119 при среднем значении 108.

4.2.3. Доджкоин

В сообществах, посвящённых доджкоину, очень часто упоминается Илон Маск как самый видный сторонник данной криптовалюты. Так, мы видим упоминание Илона Маска 24 февраля 2021 г.:

```
                [
                    0,
                    "0.029*\"elon\" + 0.012*\"post\" +
0.011*\"rh\" + 0.007*\"fun\" + 0.006*\"bit\" + 0.006*\"exodus\"
+ 0.006*\"big\" + 0.006*\"mine\" + 0.006*\"shelter\" +
0.005*\"meme\""
                ],
                [
```

```
                        1,
                        "0.038*\"rh\" + 0.016*\"long\" +
0.008*\"im\" + 0.008*\"big\" + 0.007*\"robhinhood\" +
0.007*\"two\" + 0.006*\"hold\" + 0.006*\"cash\" + 0.006*\"ever\"
+ 0.006*\"give\""
                    ],
                    [
                        2,
                        "0.030*\"elon\" + 0.021*\"hold\" +
0.020*\"long\" + 0.017*\"game\" + 0.013*\"many\" +
0.009*\"huge\" + 0.007*\"ago\" + 0.007*\"big\" + 0.007*\"give\"
+ 0.007*\"fun\""
                    ],
                    [
                        3,
                        "0.017*\"mine\" + 0.012*\"per\" +
0.011*\"post\" + 0.010*\"referral\" + 0.010*\"tax\" +
0.010*\"pool\" + 0.009*\"many\" + 0.009*\"multidoge\" +
0.008*\"link\" + 0.008*\"elon\""
                    ],
                    [
                        4,
                        "0.016*\"find\" + 0.013*\"long\" +
0.013*\"hold\" + 0.010*\"elon\" + 0.010*\"app\" +
0.010*\"dogecoins\" + 0.008*\"cost\" + 0.007*\"keys\" +
0.007*\"many\" + 0.007*\"sodogetip\""
                    ]
```

В темах 0 и 2 мы видим имя Илона. В этот день Илон Маск опубликовал в своём твиттере картинку, изображающую собаку породы шиба ину в скафандре на луне. Цена на мемкоин взлетела на 19%. Показатель тональности равен 753 при среднем 265.

28 апреля 2021 года мы опять находим упоминание имени Илона Маска, но уже в контексте американского шоу Saturday Night Life (SNL), на котором Илон Маск должен был быть ведущим 7 мая 2021 г.

```
     [
                    0,
                    "0.015*\"rh\" + 0.010*\"hold\" +
0.010*\"long\" + 0.008*\"many\" + 0.007*\"pay\" + 0.007*\"give\"
+ 0.006*\"big\" + 0.006*\"man\" + 0.006*\"elon\" + 0.006*\"im\""
                ],
                [
                    1,
                    "0.031*\"rh\" + 0.011*\"find\" +
0.009*\"many\" + 0.009*\"give\" + 0.008*\"compose\" +
0.008*\"sodogetip\" + 0.008*\"sidebar\" + 0.008*\"wiki\" +
0.008*\"app\" + 0.007*\"cash\""
                ],
                [
                    2,
                    "0.021*\"elon\" + 0.017*\"hold\" +
0.014*\"long\" + 0.011*\"tax\" + 0.010*\"rh\" + 0.008*\"big\" +
0.008*\"pay\" + 0.008*\"biden\" + 0.008*\"snl\" +
0.006*\"many\""
                ],
                [
                    3,
                    "0.017*\"big\" + 0.011*\"post\" +
0.010*\"hold\" + 0.009*\"long\" + 0.008*\"many\" +
0.006*\"elon\" + 0.005*\"give\" + 0.005*\"im\" + 0.005*\"rh\" +
0.005*\"nice\""
                ],
                [
                    4,
                    "0.031*\"elon\" + 0.018*\"snl\" +
0.011*\"long\" + 0.008*\"hold\" + 0.008*\"many\" +
0.008*\"post\" + 0.007*\"per\" + 0.006*\"pay\" + 0.006*\"low\" +
0.006*\"ellen\""
                ]
```

Имя Маска упоминается в темах 2 и 4. В четвертой теме вслед за именем идёт сокращённое название шоу. В этот день мемкоин подорожал на 18% со значением тональности 3106 при среднем 265.

В целом, в текстах сообществ DOGE много упоминаний Илона Маска, когда он публикует мем или текст о доджкоине, что приводило к росту криптоактива в 2021 году.

4.2.4. Шиба ину

Шиба ину имел наибольшее количество дней с ростом цены, превышающей 10% по сравнению с другими рассматриваемыми нами криптовалютами. Однако, в большинстве дней не находится сколько-нибудь значимых тем, но есть и положительные примеры. Например, 13 мая 2021 г. пользователи обсуждали создателя блокчейна эфир Виталика Бутерина, который пожертвовал токены шиба ину общей стоимостью миллиард долларов Индии на помощь с борьбой против коронавируса:


```
        [
                0,
                "0.043*\"mil\" + 0.014*\"hold\" +
0.013*\"usdt\" + 0.013*\"long\" + 0.012*\"elon\" +
0.011*\"bitmart\" + 0.010*\"inu\" + 0.009*\"name\" +
0.008*\"dog\" + 0.007*\"im\""
            ],
            [
                1,
                "0.017*\"news\" + 0.011*\"vitalik\" +
0.010*\"wazirx\" + 0.010*\"usd\" + 0.010*\"vb\" +
0.009*\"long\" + 0.009*\"covid\" + 0.008*\"link\" +
0.008*\"big\" + 0.008*\"quadrillion\""
            ],
            [
                2,
```

```
                "0.017*\"gas\" + 0.014*\"vitalik\" +
0.012*\"hold\" + 0.010*\"inu\" + 0.009*\"man\" +
0.008*\"donation\" + 0.008*\"vb\" + 0.008*\"true\" +
0.008*\"many\" + 0.008*\"meme\""
            ],
            [
                3,
                "0.028*\"bro\" + 0.016*\"vb\" +
0.016*\"app\" + 0.015*\"hold\" + 0.014*\"give\" +
0.012*\"long\" + 0.012*\"man\" + 0.010*\"meme\" +
0.010*\"mom\" + 0.009*\"many\""
            ],
            [
                4,
                "0.015*\"huge\" + 0.013*\"life\" +
0.010*\"hold\" + 0.010*\"many\" + 0.009*\"covid\" +
0.008*\"find\" + 0.008*\"donation\" + 0.008*\"pay\" +
0.008*\"ever\" + 0.008*\"job\""
            ]
```


Имя Виталика содержится в темах 1 и 2. Слово *donation* (*пожертвование*) также упоминается в темах 2 и 4. В этот день мемкоин подорожал на 16% с тональной оценкой 235 при среднем 152

4.3. Анализ результатов тематического моделирования с помощью алгоритма BERTopic (стандартная модель)

Для построения тематических моделей с помощью библиотеки BERTopic мы берем аналогичные дни, что и для алгоритма LDA: 6 дней для BTC, 11 дней для ETH, 14 дней для DOGE и 44 дня для SHIB. Первое, что

можно отметить, это то, что аналогично с предыдущей моделью у нас есть дни без явно выраженных тем.

4.3.1. Биткоин

По сравнению с предыдущей моделью мы получили гораздо больше тем. Однако, большинство тем описывают лишь общие явления, связанные с криптовалютами: комиссии за транзакции, программное обеспечение и кошельки. Но и наблюдаются примеры таких тем, которые могут иметь отношение к рынку криптовалют.

Так, 26 апреля из 5 тем лишь одна говорит о конкретном событии:

"tesla bitcoin sold btc sell elon company liquidity earnings holdings"

Речь идёт о том, что компания Илона Маска Tesla продала часть своих биткоин общей стоимость 272 миллиона долларов. В данный день оценка тональности имела очень высокий показатель 1165 в сравнении со средним 524. Цена закрылась на 10% дороже цены открытия. Описанное событие можно расценивать как сигнал к продаже криптоактива, так как вслед за одной крупной компанией другие также будут фиксировать прибыль, что приведёт к снижению ликвидности рынка. Действительно, через 2 недели цена биткоина обрушилась примерно в два раза.

Только в один день за выбранные нами даты была обнаружена информация, полезная для анализа рынка.

4.3.2. Эфириум

У группы сообществ ETH также множество дней не имеют ярко выраженных тем. Значимая тема обнаруживается 3 мая:

“algorand ethereum blockchain eth smart contract people like many would"

Пользователи активно обсуждали проект Algorand про криптовалютный блокчейн протокол. На следующий день цена токена алго, который реализует этот протокол, упала на 10%, но затем в течение 3 дней показала рост в 20%. В данный день оценка тональности имела высокий показатель 438 в сравнении со средним 108. Цена эфириума закрылась на 16% дороже цены открытия.

Только одна ярко выраженная тема была обнаружена и она не имела отношения к самому эфириуму.

4.3.3. Доджкоин

У группы сообществ DOGE обнаружена гораздо большее разнообразие тем и доля дней с определёнными темами. Возьмём для примера субботу 30 января, когда цена обрушилась на 39%. Одна тема выглядит весьма интересно, приведем ключевые слова темы:

```
[('pump', 0.0302890068786388 74),
 ('monday', 0.0281113682692569 48),
 ('saturday', 0.0250157860801515 47),
 ('around', 0.0216663591906754 6),
 ('increments', 0.0208761082581151 55),
 ('sell', 0.0196837195655570 13),
 ('doge', 0.0195317356088910 95),
 ('people', 0.0191096860774707 35),
 ('january', 0.0187618395601136 6),
```

```
('buy', 0.017679378823058953)]
```

Пользователи обсуждали взлёт цены (с англ. pump) и упоминали понедельник, когда цена вернулась примерно к тому значению, с которого обрушилось в субботу.

Однако, мы не находим в выбранных датах тем, упоминающих о публикациях в Twitter Илона Маска, за исключением 24 октября:

"foundation doge elon board dogecoin people one im know guy"

Речь идёт о Dogecoin Foundation, организации, которая занимается развитием данной криптовалюты. Она была основана в 2014 г., но потом упразднена. Организация объявила о возобновлении своей работы в 2021 г. В данный день люди обсуждали публикации в Twitter от Илона Маска об этой организации. В данный день оценка тональности имела высокий показатель 315 в сравнении со средним 265. Цена доджкоина закрылась на 10% дороже цены открытия.

4.3.4. Шиба ину

Шиба ину имел большее количество дней в 2021 г., когда цена закрывалась дороже 10%. Однако, в большинстве дней темы не выражены. Наблюдается частое упоминание биржи Coinbase ещё до официального размещения криптовалюты на этой бирже, но мы не находим отдельной темы, упоминающей её, в день официального листинга.

4.4. Анализ результатов тематического моделирования с помощью алгоритма BERTopic (динамическая модель)

4.4.1. Биткоин

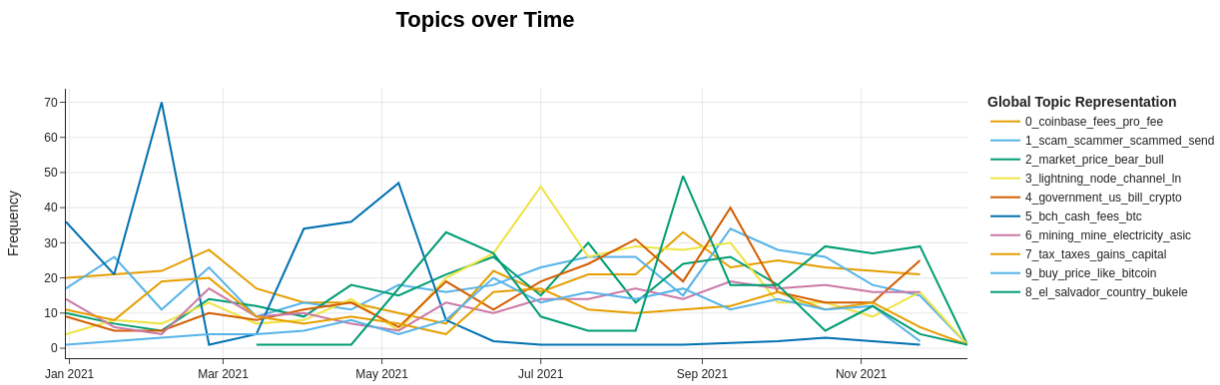

Рисунок 13. Динамическая модель BERTopic для биткоина

На графике 13 мы видим, что в группе сообществ BTC на протяжении года не было одной преобладающей темы. Каждая преобладающая тема сменялась другой. Мы видим темы, не только связанные с торгами и биржами, но и проблему потребления электричества при майнинге биткоина − тема 6, *скамы* (с англ. *мошеннические проекты*) − тема 1, налоги − тема 7. Также популярной темой была легализация биткоина в Сальвадоре. Заметно, что тема была популярна в период, когда криптовалюта переживала сильный спад в цене, следующий пик обсуждения соотносится с ростом цены.

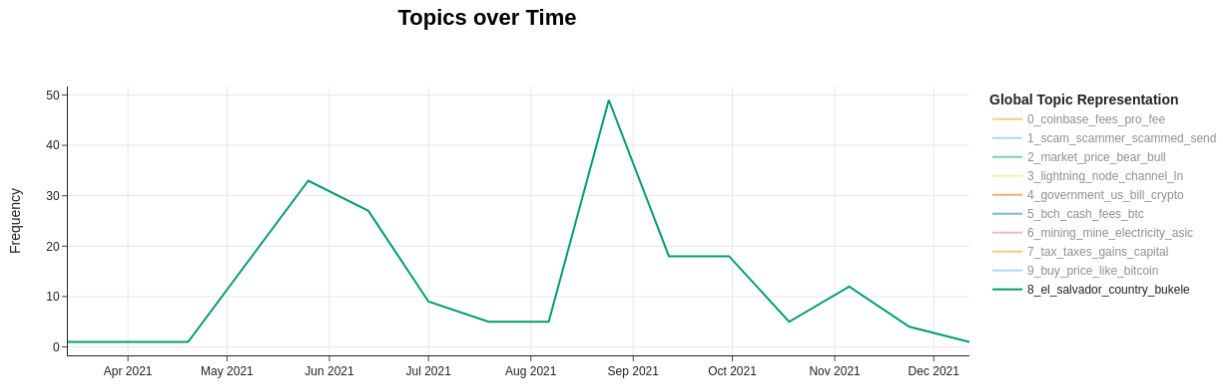

Рисунок 14. Изменение популярности темы легализации биткоина в Сальвадоре

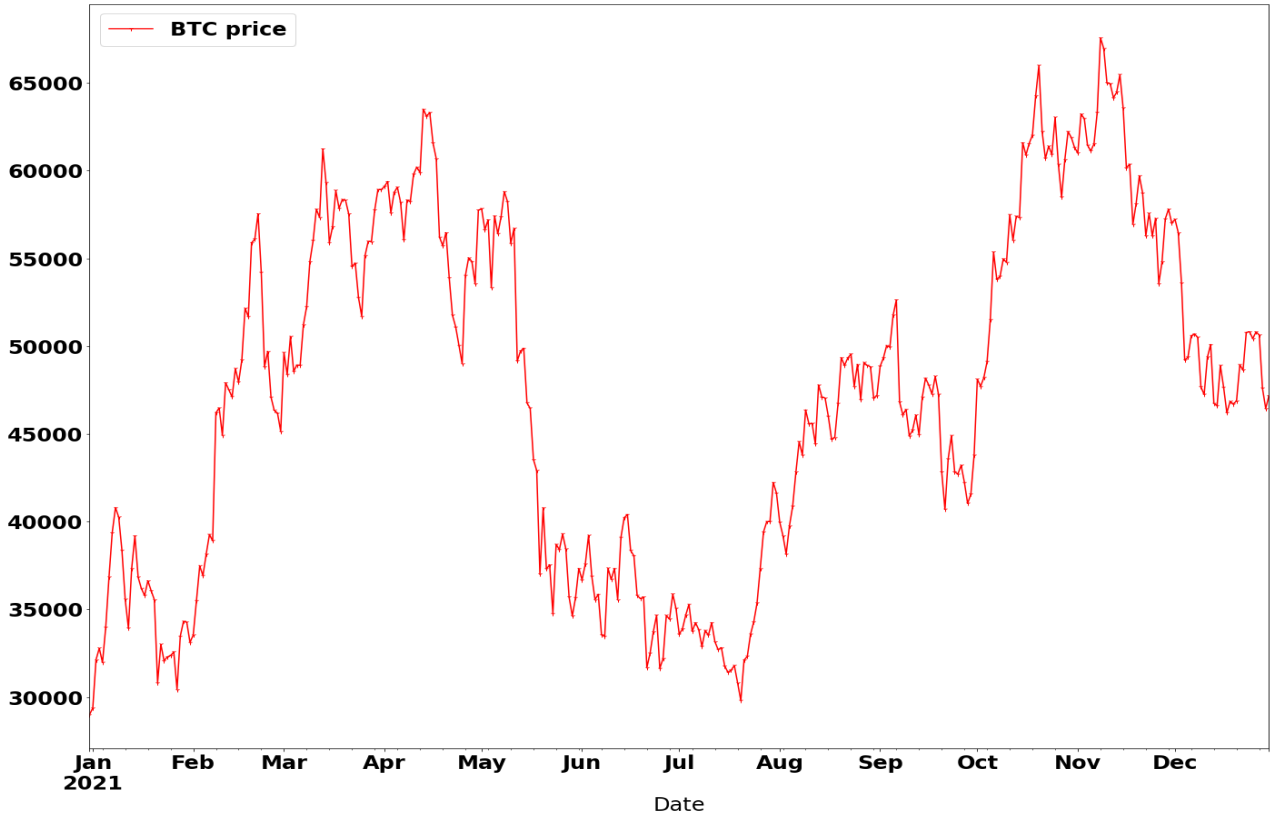

Рисунок 15. Цена биткоина за 2021 год

Также обращает на себя внимание (ср. рис. 16 и 17) популярность темы о BCH − биткоин кэш, вариант биткоина отделившийся от основной версии в 2017 г. и с тех пор существующий как отдельная криптовалюта. Хотя мы и не рассматриваем биткоин кэш в настоящем исследовании, преобладающая

популярность данной темы в первом полугодии свидетельствует о каком-то ярком событии. Если мы сравним график изменения популярности данной темы с ценой биткоин кэш, то мы увидим, что обсуждение сопутствовало росту цены. Однако, также можно заметить, что пиковая популярность не соответствует интенсивности роста цены. Когда цена с февраля по март показывала рост в 2 раза, в публикациях данная криптовалюта активно обсуждалась, но далее популярность уже не находит такого уровня, когда цена возрастает больше, чем в 2 раза.

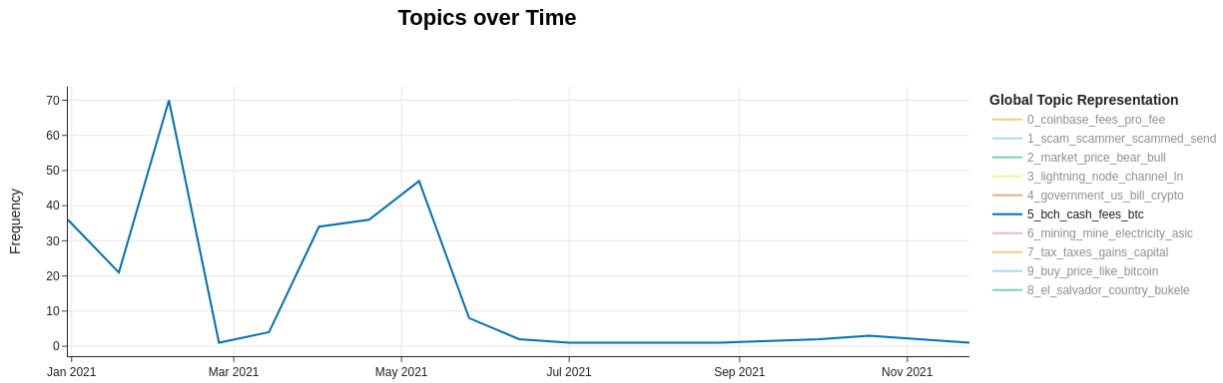

Рисунок 16. Изменение популярности биткоинкэша

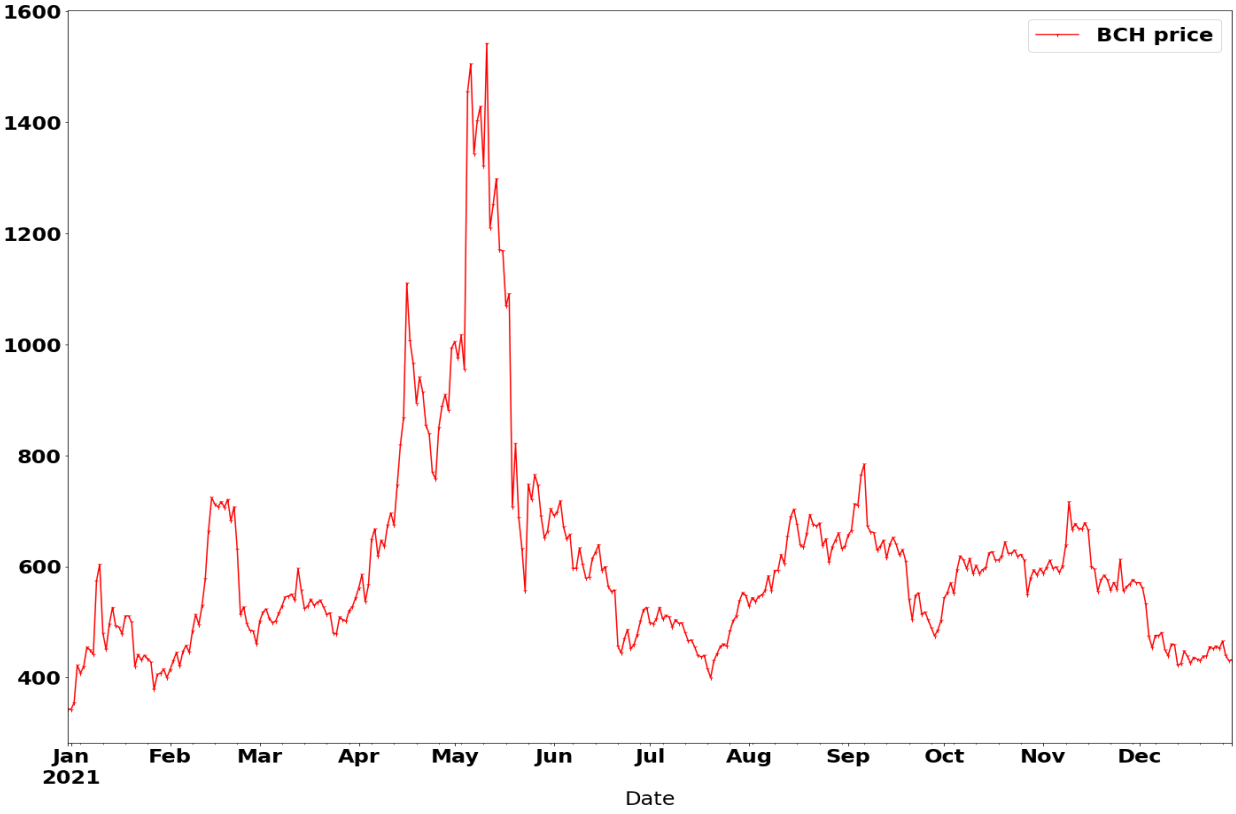

Рисунок 17. Цена биткоинкэша за 2021 год

4.4.2. Эфириум

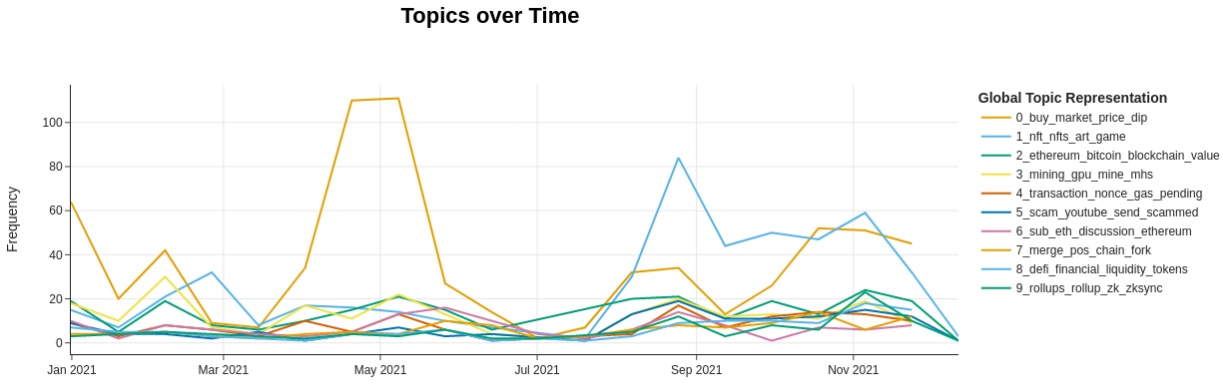

Рисунок 18. Динамическая модель BERTopic для эфириума

Как видно на графике в группе сообществ ETH, главной темой с конца марта по конец мая были торги криптовалютой, тема 0. В этот период цена эфириума поднялась с 1400 долларов за единицу до рекордных 4000 долларов. Мы можем видеть, как тема торгов становится вновь актуальным с моментами сильного роста цены. Однако, не наблюдается роста частотности темы пропорционально росту цены эфириума.

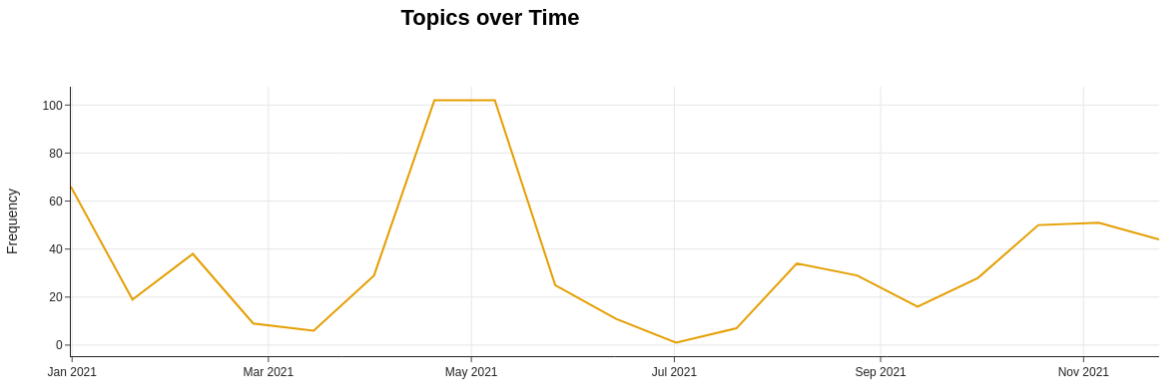

Рисунок 19. Тема покупки эфириума

Рисунок 20. Изменение цены эфириума за 2021 год

Также обращает на себя внимание тема 1, посвящённая *нфт* (с англ. *non-fungible token*) − объектам цифрового искусства записанным в блокчейн. В 2021 г. данная тема была популярна в связи с многими объекты цифрового искусства с сотен и тысяч долларов дорожали до нескольких миллионов. График показывает явные скачки в интересе пользователей к данным проектам.

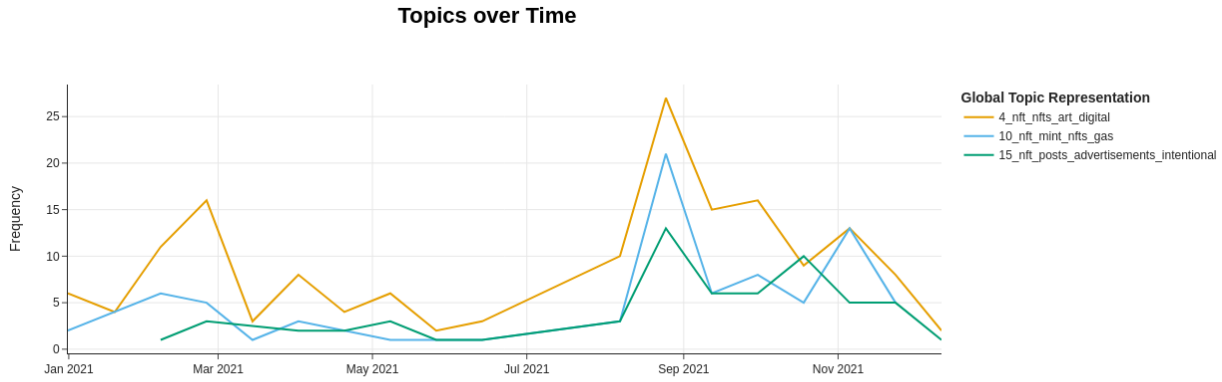

Рисунок 21. Изменение популярности нфт в группах сообществ ETH

4.4.3. Доджкоин

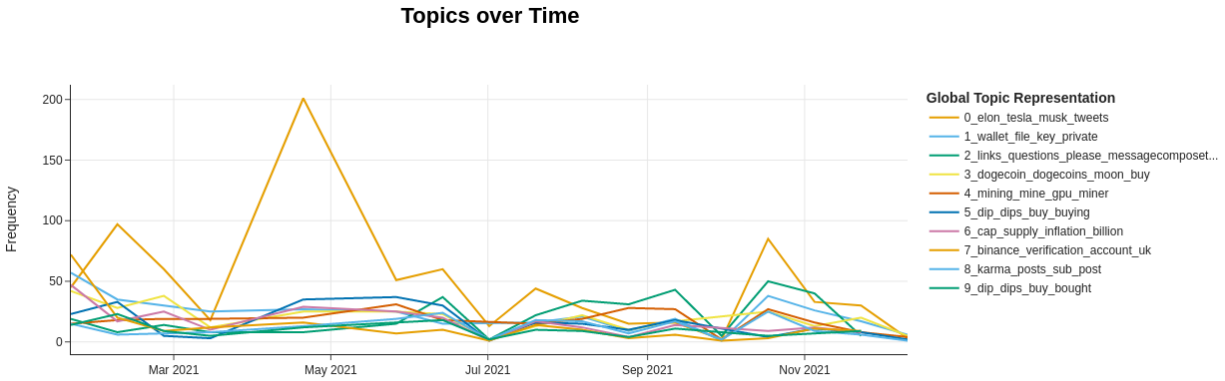

Рисунок 22. Динамическая модель BERTopic для доджкоина

График (см. рис. 22) показывает сильную популярность Илона Маска в группе сообществ DOGE. BERTopic смог успешно уловить не только Маска и его поддержку криптовалюты в твиттере, но и ассоциировал компанию Tesla с её создателем. Остальные представленные на графике темы отражают темы общие для криптосообществ - такие как криптокошельки, криптобиржи, майнинг криптовалюты, но и видно, что очень популярна тема неограниченной эмиссии доджкоина. Всего модель обнаружила 149 тем.

Также участники криптосообществ обсуждали, как видно на графике (см. рис. 23) и политические темы - налоги в США, Дональда Трампа, а также запрет на криптовалюты в КНР:

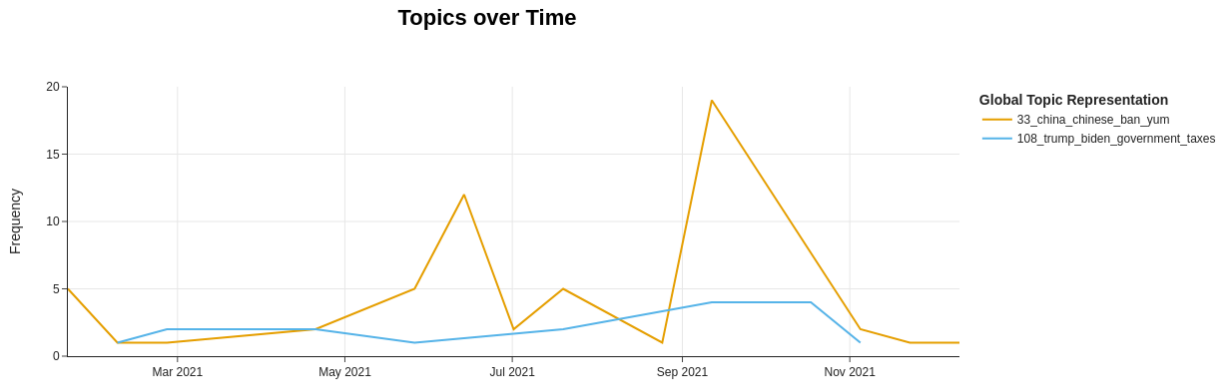

Рисунок 23. Политические темы в группах сообществ DOGE

Сравним график частотности тем с графиком цены доджкоин (ср. рис. 23 и 24). Как мы видим Илон Маск обсуждался очень интенсивно в период с апреля по май, когда цена криптовалюты росла и обновила исторический максимум. В этот период Илон Маск очень активно публиковал в Twitter своё мнение о криптовалюте, таким образом популяризируя её.

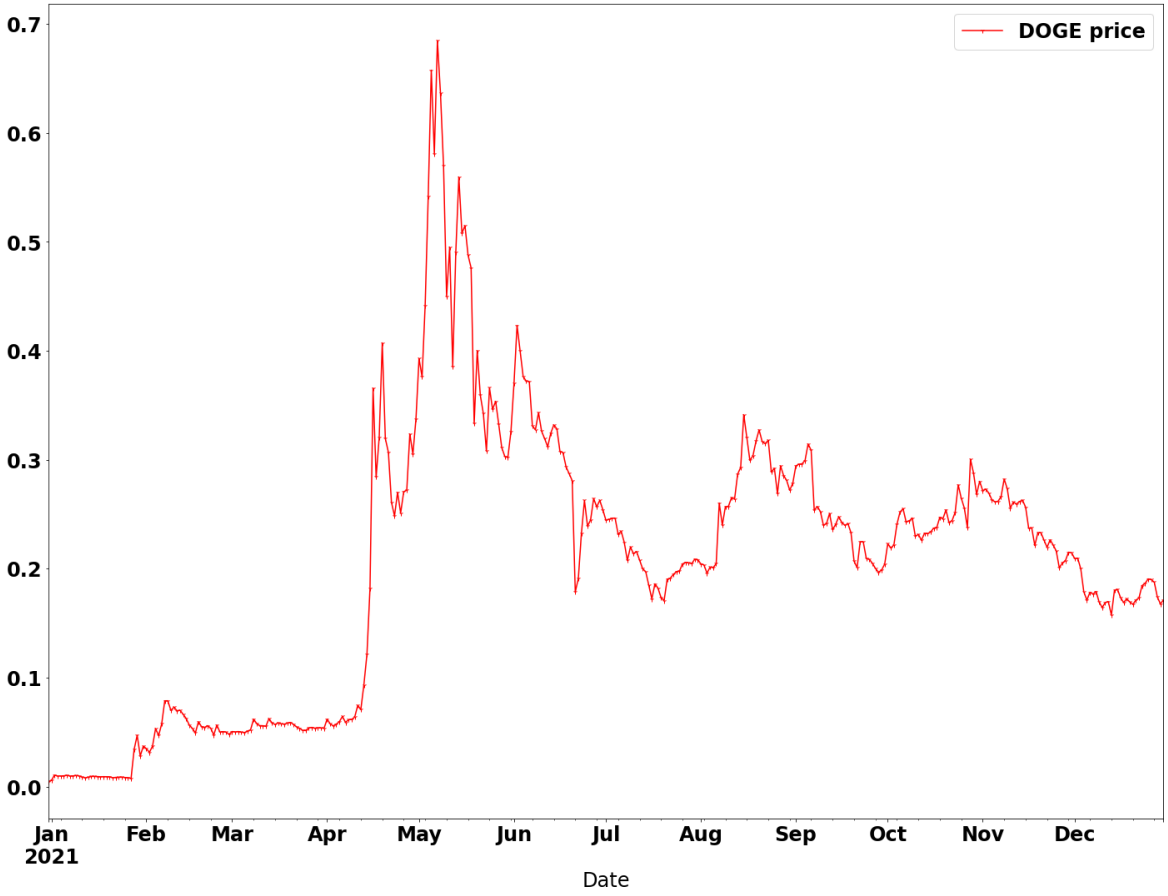

Рисунок 24. Цена доджкоин за 2021 год

4.4.4. Шиба ину

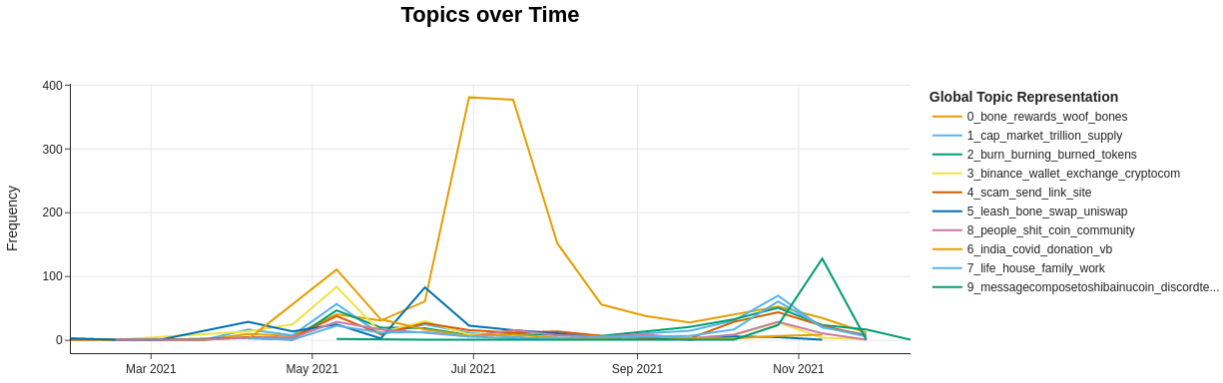

Рисунок 25. Динамическая модель BERTopic для шиба ину

Преобладающей темой 0 в группе сообществ SHIB оказались так называемые мемы. Также видно, что пользователей на протяжении года волновала тема 1 об огромной эмиссии криптовалюты и темы 2 о сжигании объёма предложения. При этом достаточно популярными была тема 6 о коронавирусе и пожертвовании шиба ину на миллиард долларов Виталика Бутерина Индии.

Рассмотрим некоторые важные для цены темы в отдельности. Как видно на графике BERTopic смог выявить очевидно важные темы для цены шиба ину - листинг на биржах Binance в мае, тема 3, и Coinbase в сентябре, тема 19. Так же, как и у доджкоин, Twitter Илона Маска имел определённую популярность среди пользователей. Интересно отметить тему 43, в которой пользователи обсуждали организацию сбора подписей для финансовой площадки Robinhood, чтобы на ней разрешили торговлю шиба иной. Данная тема появляется в середине мая, а Robinhood удовлетворил просьбы сторонников шиба ину только в апреле 2022 г., но это уже выходит за временные рамки нашего исследования.

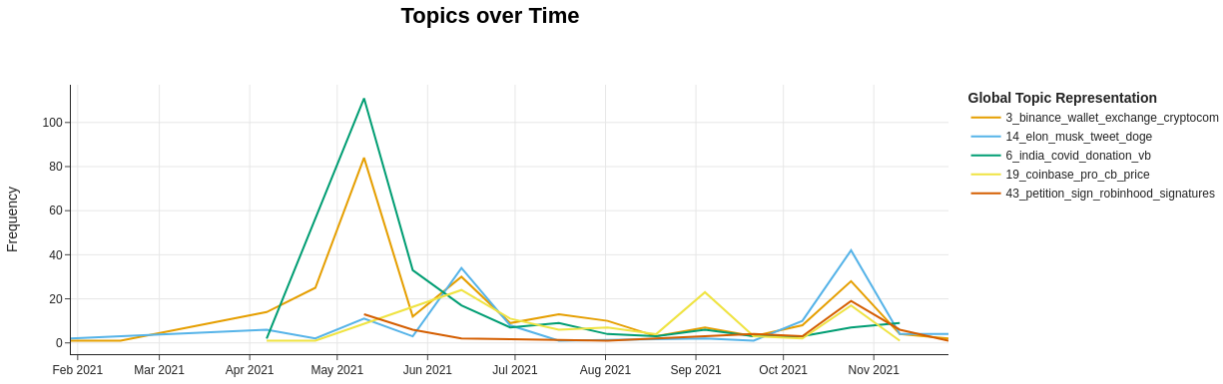

Рисунок 26. Наиболее значимые темы для шиба ину

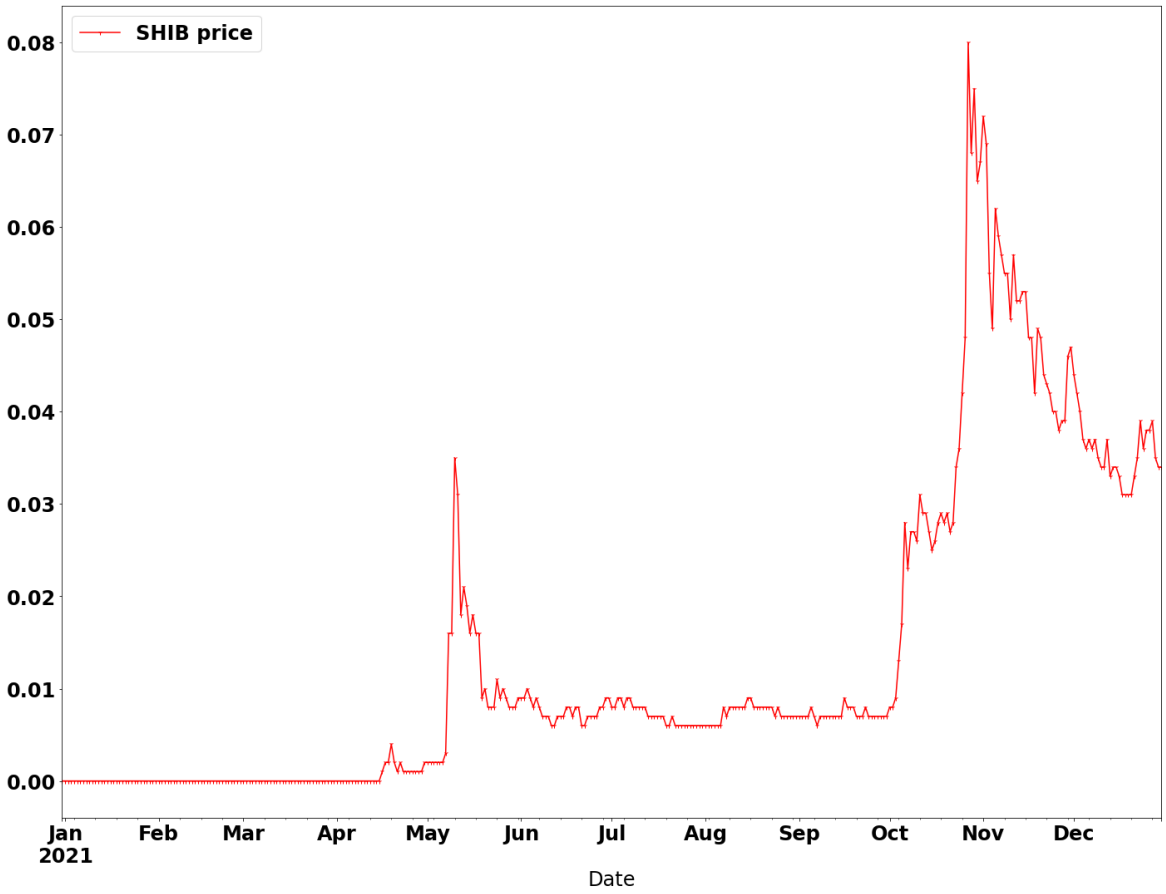

Рисунок 27. Цена шиба ину за 2021 год

Как мы можем видеть, скачок в цене в мае соотносится сразу с двумя темами − пожерствованием Виталика Бутерина и листингом на Binance. Новость о миллирдном пожертвовании в криптовалюте послужила популяризации криптоактива, а листинг на бирже дал доступ большему количеству инвесторов. Однако, вопреки ожиданиям, листинг на бирже Coinbase в сентябре никак не отразился на цене шиба ину.

Выводы по главе 4

Итак, проведенный нами тематический анализ корпуса публикаций показал разные результаты для двух моделей − LDA и BERTopic. С одной стороны, явным преимуществом BERTopic является автоматическое определение оптимального количества тем, но модель LDA смогла показать результаты, отличные от результатов BERTopic. Так, например, модель LDA смогла выявить в группе сообществ DOGE обсуждение другого проекта Cardano, который только начинал свой путь в мире блокчейн, а его токен ADA имел относительно низкую цену.

Стоит отметить, что обе модели не всегда справлялись с поставленной задачей и не могли соотнести слова за конкретный день ни к одной теме. Это можно объяснить, во-первых, небольшим количеством текстов в отдельные дни и, во-вторых, коммуникация в Reddit является куда более сложным явлением, чем общение в Twitter: нами не брались во внимание картинки, на которых могли быть изображены мемы, очень часто пользователи общаются посредством эмоджи, делятся ссылками и т.д.

Динамическое тематическое моделирование смогло показать, как сменялись тренды в Reddit, а также отобразить ключевые события, влияющие на цены. Многие темы имели цикличную частотность и соответствовали росту цены хотя и не пропорционально. Были обнаружены обсуждения и сторонних проектов и криптовалют. Так, например, в группе сообществ ETH активно обсуждались нфт, что никак не связано с ценой эфириума, но может свидетельствовать об определённых событиях на рынке цифрового искусства.

Результаты между выбранными нами криптовалютами неоднородны. Сравнивая результаты всех четырёх криптовалют, мы не находим особенной

выразительности данных, объясняющих движение цены, у мемкоинов DOGE и SHIB. Из наших данных видно, что одно событие, как, например, листинг на Binance у шиба ину, и другое аналогичное ему событие, как листинг той же криптовалюты на Coinbase, имеют разные последствия для цены. Определенно, у нас есть эпизодические примеры соответствия реакции сообщества и цены на публикации Илона Маска в Twitter, но такие примеры не определяют и не объясняют изменение цены в длительной перспективе. Таким образом, мемкоины не обнаруживают лучшего соответствия данных исследования изменению цены, чем у таких крупных криптовалют как биткоин и эфириум.

Заключение

Наше исследование показало, что количество публикаций в тематических сообществах социальной сети Reddit не имеет никакой корреляции с ценой на соответствующую криптовалюту. Однако, мы обнаружили, что пользовательская поддержка криптовалюты может быть сильной несмотря на то, что данный криптоактив ещё не торгуется на криптобиржах и не имеет реальной стоимости в фиатной валюте.

Тональность публикаций имеет разную корреляцию по отношению к изменению объёма, цены и волатильности в зависимости от криптовалюты. У эфириума выявлена корреляция слабой силы с изменением объёма и у шиба ину выявлена слабая корреляция с объёмом и волатильностью.

Мы доказали, что тексты Reddit содержат разнообразную информацию как о внешних событиях, так и о самих криптовалютах, которая может оказывать влияние на цену. Популярность темы не всегда пропорциональна изменению цены. Также тексты в сообществах, посвящённых одним криптовалютам, могут содержать информацию о других проектах. Тексты сообществ Reddit, посвящённых мемкоинам не показывают лучшее соответствие изменение цены, чем у крупных криптовалют.

Получив результаты обоих анализов, можно заключить, что в отдельности ни один из подходов не может дать достаточной информации для определения тренда на рынке криптовалют. Тематический анализ показал присутствие разнообразия тем в

большинстве дней, что может оказывать влияние на общую оценку тональности отдельного дня, поэтому представляется перспективным подход использования анализа тональности на текстах, разделённых тематическими моделями.

Результаты данной работы могут быть применены в создании системы для отслеживания популярных тем в сотнях сообществах Reddit. В данном имплементации важен не только динамический анализ тем, показывающий изменение трендов, но и тематический анализ на ежедневных текстах. Такой анализ может не только показать факторы потенциально важные для рынка, но и отслеживать блокчейн проекты на ранних стадиях становления или выявлять новые проекты цифрового искусства в виде нфт.

Список использованной литературы

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных - 2017 - М.: НИУ ВШЭ - С. 269.
2. Разрабатываем простую модель глубокого обучения для прогнозирования цен акций с помощью TensorFlow, 2018 URL: https://habr.com/ru/company/iticapital/blog/354732/ (дата обращения 01.04.2022)
3. Islam MR, Nguyen N. Comparison of Financial Models for Stock Price Prediction. Journal of Risk and Financial Management. 2020
4. Burba Davide An overview of time series forecasting models, 2019 URL: https://towardsdatascience.com/an-overview-of-time-series-forecasting-models-a2fa7a358fcb (дата обращения: 01.04.2022)
5. Mick Smith A Comparison of Time Series Model Forecasting Methods on Patent Groups, 2015 URL: http://ceur-ws.org/Vol-1353/paper_13.pdf (дата обращения: 01.04.2022)
6. Philip J. Stone,Robert F. Bales,J. Zvi Namenwirth,Daniel M. Ogilvie The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of
7. J. Kim, J. Seo, M. Lee and J. Seok, "Stock Price Prediction Through the Sentimental Analysis of News Articles," 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 2019, pp. 700-702
8. Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia and David C. Anastasiu Stock Price Prediction Using News Sentiment Analysis

// IEEE Fifth International Conference on Big Data Computing Service and Applications, 2019

9. László Nemes & Attila Kiss (2021) Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication, 5:3, 375-394
10. Arora, Arjun. “Using news titles and financial features to predict intraday movements of the DJIA.” (2019).
11. Heeyoung Lee, Mihai Surdeanu, Bill MacCartney and Dan Jurafsky. On the Importance of Text Analysis for Stock Price Prediction. Language Resources and Evaluation Conference (LREC). 2014
12. Kalyani Joshi , Prof. Bharathi H. N. , Prof. Jyothi Rao STOCK TREND PREDICTION
    USING NEWS SENTIMENT ANALYSIS, 2016 URL: https://arxiv.org/pdf/1607.01958.pdf (дата обращения: 05.03.2022)
13. Kari Lee and Ryan Timmons Predicting the Stock Market with News Articles, Stanford University, 2016 URL: https://nlp.stanford.edu/courses/cs224n/2007/fp/timmonsr-kylee84.pdf (дата обращения: 04.04.2022)
14. Y. Shynkevich, T. M. McGinnity, S. Coleman and A. Belatreche, "Predicting Stock Price Movements Based on Different Categories of News Articles," 2015 IEEE Symposium Series on Computational Intelligence, 2015, pp. 703-710, doi: 10.1109/SSCI.2015.107.
15. Chahat Tandon, Sanjana Revankar, Hemant Palivela, Sidharth Singh Parihar, How can we predict the impact of the social media messages on the value of cryptocurrency? Insights from big data analytics, International Journal of Information Management Data Insights, Volume 1, Issue 2,2021

16. Muxi Xu NLP for Stock Market Prediction with Reddit Data”NLP for Stock Market Prediction with Reddit Data // Stanford University, 2022
17. Wooley, Stephen, Andrew Edmonds, Arunkumar Bagavathi and Siddhartha Krishnan. “Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment.” 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (2019): 500-505.
18. Ramon Hinojosa Alejandro. “Twitter and Reddit posts analysis on the subject of Cryptocurrencies”, 2021.
19. Chuluunsaikhan, Tserenpurev, Ga-Ae Ryu, Kwan-Hee Yoo, HyungChul Rah, and Aziz Nasridinov. 2020. "Incorporating Deep Learning and News Topic Modeling for Forecasting Pork Prices: The Case of South Korea" Agriculture 10, no. 11: 513.
20. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (by C.J. Hutto and Eric Gilbert) Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
21. Chen, Chung-Chi, Hen-Hsen Huang and Hsin-Hsi Chen. “NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications.” (2018).
22. Самигулин Тимур Русланович, Джурабаев Анвар Эркин Угли АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ // Научный результат. Информационные технологии. 2021. №1. URL: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-metodami-mashinnogo-obucheniya (дата обращения: 23.05.2022).
23. Bengio Y. Learning deep architectures for AI // Foundations and Trends in Machine Learning, 2009.

24. Ahmad M. et al. Machine learning techniques for sentiment analysis: A review //Int. J. Multidiscip. Sci. Eng. – 2017. – T. 8. – No. 3. – P. 27.
25. Пескишева Т.А. Методы анализа тональности текстов на естественном языке // Общество. Наука. Инновации (НПК-2017). – 2017. – С. 1730-1742.
26. Schofield, Alexandra, Måns Magnusson, Laure Thompson and David Mimno. “Pre-Processing for Latent Dirichlet Allocation.” (2017).
27. file:///home/kartashow/Downloads/vkr_Zamiraylova.pdf
28. Blei D. M. Introduction to Probabilistic Topic Models // Communications of the ACM. Vol. 55. Issue 4. 2012, C. 77–84.
29. Maarten Grootendorst BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022 URL: https://arxiv.org/abs/2203.05794 (дата обращения: 05.05.2022)
30. Senyuk Lyubomyr 5 Natural Language Processing (NLP) Applications In Finance, 2021
URL: https://www.avenga.com/magazine/nlp-finance-applications/ (дата обращения: 30.05.2022)
31. Tsarouva Maria The unprecedented revolution of NLP in finance, 2020 URL:
https://www.itechart.com/blog/natural-language-processing-in-finance/ (дата обращение: 30.05.2022)
32. Чернова Анастасия Бычий и медвежий рынок: кто такие быки и медведи на бирже, 2022 URL:
https://www.nalogia.ru/articles/551-bychiy-i-medvezhiy-rynok-kto-takie-byki-i-medvedi-na-birzhe.php (дата обращения: 30.05.2022)

33. Хобсон Лейн, Ханнес Хапке, Коул Ховард Обработка естественного языка в действии - 2020 - СПб: "Издательский дом ""Питер""", 2020 - С. 576
34. Российский семинар по Оценке Методов Информационного Поиска URL: http://romip.ru/en/ (дата обращения: 10.04.2022)
35. SemEval-2022 URL: https://competitions.codalab.org/competitions/33556 (дата обращения: 10.04.2022)
36. TextBlob: Simplified Text Processing URL: https://textblob.readthedocs.io/en/dev/ (дата обращения: 30.04.2022)
37. NLTK Documentation. Sample usage for sentiment URL: https://www.nltk.org/howto/sentiment.html#sample-usage-for-sentiment (дата обращения: 10.04.2022)
38. Анализ тональности в русскоязычных текстах, часть 1: введение URL: https://habr.com/ru/company/vk/blog/516214/ (дата обращения: 20.05.2022)
39. Smetanin S., "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," in IEEE Access, vol. 8, pp. 110693-110719, 2020
40. Blei D.M, Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3.