УДК 621.3.078.4+519.876.5

Анализ новостного фона криптовалют

А. В. Копица, Е. О. Савкова
ГОУ ВПО «Донецкий национальный технический университет», г. Донецк
helen-savkova@rambler.ru

Аннотация
Данная статья посвящена анализу новостного фона для прогнозирования цен криптовалют с учетом эмоциональной окраски, связанных с ними новостей. Для этого был определен источник новостей, разработан алгоритм их получения и формирования базы для хранения извлеченной информации. Предложена методика обработки новостей, в результате чего появляется возможность классифицировать новости на категории с различной оценкой тональности текста. В результате разработано программное приложение для анализа новостного фона. С помощью данного приложения производится сбор новостей в автоматическом режиме из новостного источника, связанного с криптовалютными биржами, а затем с помощью выбранного классификатора выполняется автоматическая оценка новостей. Результаты исследований в дальнейшем могут быть использованы для разработки системы более точного прогнозирования цен криптовалют.

Введение

В наше время множество людей покупают и продают акции, валюту, фьючерсы, опционы и т.д. Одно из основных допущений теории принятия решений состоит в том, что человек делает рациональный выбор. Рациональный выбор означает, что человек принимает решение в результате упорядоченного процесса мышления.

В 70-е годы появились работы, в которых систематически демонстрировалось отклонение поведения людей от рационального.

Авторами наиболее известных работ были психологи: А.Тверский, П.Словик, Б.Фишхоф, Д.Канеман и др.

Многочисленные эксперименты продемонстрировали отклонение поведения людей от рационального, определили эвристики, которые используются при принятии решений.

Известный физик Ньютон был большим любителем играть на бирже, но делал он это видимо не из расчета рационального поведения людей. Играл он не очень удачно. Даже совсем неудачно. Один раз, после крупного проигрыша, он произнес знаменитую фразу: – Я могу прогнозировать движение космических тел, но не могу прогнозировать безумство людской толпы (движение цен и поведение спекулянтов).

Каковы же тогда причины в этой нерациональности человеческого поведения? Причинами нерациональности человеческого поведения являются:

1) недостаток информации у ЛПР (лицо, принимающее решение) в процессе выбора;

2) недостаточный опыт ЛПР: он находится в процессе обучения и поэтому меняет свои предпочтения;

3) ЛПР стремится найти решение, оптимальное с точки зрения совокупности критериев (целей), строго упорядоченных по важности, но не может его найти;

4) различие между объективно требуемым временем для реализации планов и субъективным горизонтом планирования ЛПР.

Именно поэтому и необходимо разрабатывать инструменты для определения возможных действий людей на основании их поведения, а благодаря веку информационных технологий, прогнозирование поведения людей становится не такой сложно задачей, поскольку люди подвергаются влиянию поступающим к ним новостей, а в свою очередь за счет этого принимают решения. Так же если научится связывать новости с ценами на биржах, можно привести торговлю к большей осознанности. Поэтому создание инструмента позволяющего прогнозировать цены на бирже с учетом новостей, является необходимым.

Влияние новостей на биржу

В основе каждого рынка лежит равновесие спроса и предложения. Участники рынка обеспечивают спрос или предложение, основываясь на собственном эмоциональном восприятии происходящих событий, освещение которых они получают из новостной и/или текстовой информации. Информационные (и в первую очередь новостные) потоки оказывают существенное влияние на котировки рынка, а степень влияния зависит от характера и типа

информационного события.

Традиционные методы технического анализа, эффективно работающие на коротких временных интервалах, не принимают во внимание влияния эмоционального восприятия участниками рынка информационного потока. Фундаментальный анализ (к которому относят и анализ информационного потока — эмоциональный анализ) рассматривает в основном регулярно публикуемые финансовые показатели компаний, т. е. статистическую информацию, имеющую тенденцию к запаздыванию. Таким образом, методы традиционных методологий (технический и фундаментальный анализ) прогнозирования котировок рынка ценных бумаг ориентированы на количественные данные. Тем не менее сегодня в практике анализа рынка ценных бумаг существует много работ, использующих качественные данные, такие как текстовая и новостная информация, которая может быть извлечена из новостных статей, блогов, аналитических отчетов, социальных опросов. Обычно доступ к внутрикорпоративной информации отсутствует, поэтому для анализа особенностей деятельности той или иной компании рассматривается информация из регулярно публикуемых финансовых отчетов, аналитических прогнозов, новостных статей, отражающая некоторые особенности деятельности компаний [1].

Анализ новостей способен выявить особенности компании, которые традиционный финансовый анализ выявить не может. Причина этого в том, что лица, принимающие решения эмоциональны, и, следовательно, субъективны. Причем их реакция на одни и те же события может быть совершенно различной (поведенческая экономика). Изучение текстовой и новостной информации позволяет сформировать более полное суждение о финансовой состоятельности участника рынка. Ручное отслеживание новостного потока и его последующий анализ является весьма затруднительным для аналитика, поэтому применение методов интеллектуального анализа и создание инструментальных средств их поддержки — востребованная задача.

Сбор данных

В начале надо решить задачу выбора источника новостей для анализа. Поскольку работа проводится с криптовалютными биржами, то был выбран самый популярный новостной агрегатор с web-приложением доступным по адресу https://ru.tradingview.com.

Извлечение данных из сайтов называется веб-скрапингом. Это полезная практика, когда доступная вам информация доступна через веб-приложение, которое не предоставляет соответствующий API (как в нашем случае). Для извлечения данных из современных веб-приложений требуется некоторая нетривиальная работа, но зрелые и хорошо продуманные инструменты, такие как Selenium и BeautifulSoup решают подобные задачи [2].

Так как веб-приложение не имеет собственного API извлекать новости приходится методом обхода всех страниц. Использование BeautifulSoup не считается возможным, поскольку в веб-приложении предусмотрена защита от ботов для обхода страниц вне зависимости от настроек метода. Был выбран более требовательный, но эффективный инструмент, предназначенный для тестирования, но подходящий также для нашей задачи - Selenium. Данный инструмент является средством, запускающим браузер и осуществляющий переходы путем имитации действий человека, что, в итоге, не позволяет определить это как работу бота.

Определившись с инструментом, мы можем получить все новости по такому алгоритму:

1) Запускается браузер с помощью selenium через специальный драйвер.

2) Происходит обход всех страниц со списком кратко описанных новостей до последней страницы с новостями для получения точного диапазона страниц.

3) Происходит обход с последней страницы и до первой включительно по блокам html страницы из которых извлекаются новости в базу данных в формате: - дата новости; - заголовок; -ссылка на полную новость.

4) По списку ссылок на полные новости, которые мы храним в БД, определяем в какие записи не был внесен полный текст и обходим их как описано в пункте 3, только извлекается полная новость и если есть дата и время, то обновляется запись в БД с полученными данными.

Собранная база новостей позволяет проводить анализ текста, предварительно проведя обработку и токенизацию текста. За счет этого мы также сможем определять привязку криптовалют к новостям.

Предварительная обработка и токенизация текста

Одной из главных проблем анализа текстов является большое количество слов в документе. Если каждое из этих слов подвергать анализу, то время поиска новых знаний резко возрастет и вряд ли будет удовлетворять требованиям пользователей. В то же время очевидно, что не все слова в тексте несут полезную информацию. Кроме того, в силу гибкости естественных языков формально различные слова (синонимы и т. п.) на самом

деле означают одинаковые понятия. Таким образом, удаление неинформативных слов, а также приведение близких по смыслу слов к единой форме значительно сокращают время анализа текстов. Устранение описанных проблем выполняется на этапе предварительной обработки текста [1, 3].

На этапе предварительной обработки текста действия выполняются в 4 этапа.

Этап 1 - текст приводится к нижнему регистру.

Этап 2 - удаляются знаки препинания и спецсимволы, т.е. такие элементы, как запятые, точки, дефисы, знаки восклицания, знаки вопроса, различные скобки, кавычки и математические знаки. При необходимости список можно расширять;

Этап 3 - удаляются стоп-слова. К стоп-словам (или шумовым словам), как правило, относят предлоги, союзы, междометия, частицы и другие части речи, которые часто встречаются в тексте, являются служебными и не несут смысловой нагрузки – являются избыточными. Типовыми удаляемыми стоп-словами являются: -служебные части речи (предлоги, союзы, частицы); -самостоятельные части речи (местоимения). Следует отметить, что стоп-слова являются контекстно зависимыми – для текстов различной тематики стоп-слова могут отличаться. Как и в случае со спецсимволами, необходимо проанализировать исходный текст и выявить стоп-слова, которые не вошли в типовой набор.;

Этап 4 - проводится лемматизация. У всех слов происходит объединение времени, объединение единственного и множественного числа.

Для последующей обработки очищенный текст разбивается на составные части – токены. Процесс разбиения называется токенизацией. При анализе текста на естественном языке применяется разбиение на символы, слова и предложения. Разбивая обработанный текст на слова можно провести частотный анализ текста, который так же может позволить трейдеру понять какие слова сейчас на слуху (в топе). Частотный анализ текста, связанного с новостями про криптовалюты представлен на рис. 1.

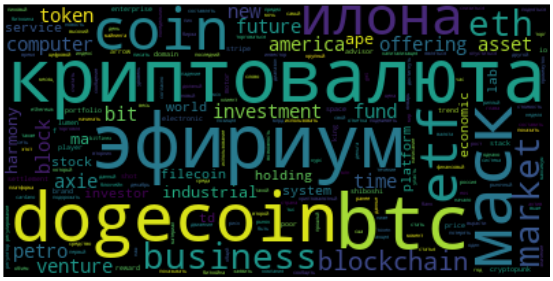

Рисунок 1 – Частотный анализ текста в виде облака слов

На рисунке слова, выделенные разным цветом, разделяются друг от друга. Размером слова в свою очередь указывается частота слова в тексте (чем чаще слово встречается, тем больше его размер). Расположение слов не имеет никакого значения.

Обучение модели, определяющей оценку новости

Классификация текстов (документов) — задача компьютерной лингвистики, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Анализ тональности текста — задача компьютерной лингвистики, заключающаяся в определении эмоциональной окраски (тональности) текста и, в частности, в выявлении эмоциональной оценки авторов по отношению к объектам, описываемым в тексте.

Для того чтобы мы могли качественно определять тональность текста нам необходимо обучить модель классификации для выполнения необходимых задач. Определим для начала этапы, которые необходимо выполнить для обучения модели классификации текстов:

1) Получение набора новостей;

2) Предварительная обработка новостей;

3) Определение возможных типов новостей по фону(тональности);

4) Ручная оценка 20-30% новостей;

5) Разбиение выборки на обучающую и тестовую, из новостей, которые имели оценку.

6) Выбор классификатора или конвейера классификаторов для обучения модели.

7) Обучение классификатора на предварительно обработанных новостях со связанными оценками.

8) Выбор главного классификатора (на основании метрик) для следующей работы с ним.

Выполнение первых двух пунктов детально описаны ранее.

Этап 3. Обозначим минимальный набор типов новостей по фону и определим количественную шкалу для оценки. Шкала представлена в табл. 1.

Таблица 1 – Шкала оценок новостей

№	Численное значение	Качественное значение
1	2	очень позитивная
2	1	позитивная
3	0	нейтральная
4	-1	негативная
5	-2	очень негативная

После этого для обучения необходимо оценить 20% (порядка 600) новостей в соответствии с типами, определенными ранее. В качестве эксперта в данном случае выступает автор, при этом вручную просматриваются (вычитываются) все новости выбранной части набора данных и выставляется его субъективная оценка.

После того как выборка новостей полностью оценена, весь набор данных разбивается на тренировочную и тестовую выборки в соотношении 80 к 20. Тестовый набор данных будет применяться для тестирования правильности работы классификатора.

Этап 4. Выберем классификаторы и дополнительные инструменты для обучения модели классификации. Опишем используемые средства для моделей.

Используемые классификаторы:

- SGDClassifier;
- KNeighborsClassifier;
- LinearSVC.

Дополнительные используемые средства:

- TfidfVectorizer;
- CountVectorizer;
- SelectKBest;

Классификатор стохастический градиентный спуск (SGDClassifier) - классификатор стохастического градиентного спуска в основном реализует простую процедуру обучения SGD, поддерживающую различные функции потерь и штрафы за классификацию.

Стохастический градиентный спуск (SGD) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь, такие как (линейные) машины опорных векторов и логистическая регрессия. Несмотря на то, что SGD существует в сообществе машинного обучения уже давно, совсем недавно он привлек значительное внимание в контексте крупномасштабного обучения.

SGD успешно применяется для крупномасштабных и разреженных проблем машинного обучения, часто встречающихся при классификации текста и обработке естественного языка. Учитывая разреженность данных, классификаторы в этом модуле легко масштабируются на проблемы с более чем 10^5 учебными примерами и более чем 10^5 функциями [4].

Классификация на основе соседей (NeighborsClassifier) — это тип обучения на основе экземпляров или необобщающего обучения: он не пытается построить общую внутреннюю модель, а просто сохраняет экземпляры обучающих данных. Классификация вычисляется простым большинством голосов ближайших соседей каждой точки: точке запроса назначается класс данных, который имеет наибольшее количество представителей среди ближайших соседей точки.

Существует два разных классификатора ближайших соседей: KNeighborsClassifier реализует обучение на основе k ближайшие соседи каждой точки запроса, где k - целочисленное значение, указанное пользователем. RadiusNeighborsClassifier реализует обучение на основе количества соседей в фиксированном радиусе r каждой тренировочной точки, где r — это значение с плавающей запятой, указанное пользователем [5].

Метод классификации опорных векторов (LinearSVC) — это мощные, но гибкие контролируемые методы машинного обучения, используемые для классификации, регрессии и обнаружения выбросов. LinearSVR обладает большей гибкостью в выборе функций штрафов и потерь. Он также лучше масштабируется для большого количества выборок [6].

CountVectorizer – данный метод используется для преобразования заданного текста в вектор на основе частоты каждого слова, которое встречается во всем тексте. Это полезно, когда имеется несколько текстов и мы хотим преобразовать каждое слово в каждом тексте в векторы (для использования в дальнейшем анализе текста) [7].

TfidfVectorizer – данным метод основан на принципе Term Frequency – Inverse Document Frequency(Частота термина – обратная частота документа). Это один из наиболее важных методов, используемых для поиска информации, чтобы показать, насколько важно конкретное слово или фраза для данного документа. Давайте возьмем пример, у нас есть строка и мы должны извлечь из нее информацию, тогда мы сможем использовать этот подход. Значение tf-idf увеличивается пропорционально количеству раз, когда слово встречается в документе, но часто компенсируется частотой слова в корпусе, что помогает скорректировать тот факт, что некоторые слова появляются чаще в целом.

В TF-IDF используются два статистических метода: первый - частота терминов, а другой - обратная частота документов. Частота термина относится к общему количеству раз, когда данный термин t появляется в документе doc из общего количества всех слов в документе и обратного показателя частоты документа о том, сколько информации предоставляет слово. Он измеряет вес данного слова во всем документе. IDF показывает, насколько распространенным или редким является данное слово во всех

документах. TF-IDF может быть вычислен как tf * idf [8].

Выбор К лучших (SelectKBest) - это метод, при котором выбираются те функции в данных, которые вносят наибольший вклад в целевую переменную. Другими словами, выбираются лучшие прогнозы для целевой переменной [9];

Для того чтобы выбрать лучший классификатор для определения новостного фона нам необходимо результаты работы классификации оценить метриками [10, 11].

Для понимания принципов работы метрик оценки классификации можно обратить внимание на рис. 2. .

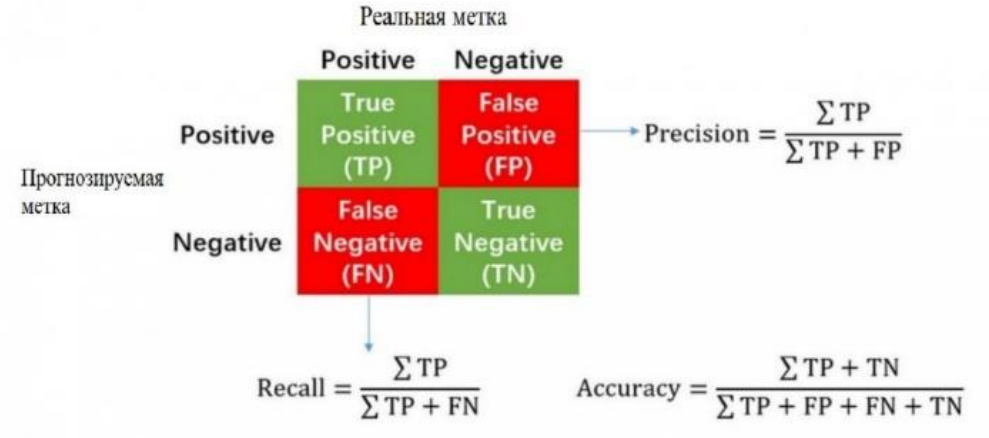

Рисунок 2 – Диаграмма, описывающая метрики

Positive и Negative - это предсказания нашей модели, а True и False- это оценка того, правильно ли определила модель наш класс.

Метрики оценки классификатора:

- Precision (точность);
- Recall (полнота);
- F1-score.

Precision - отношение TP к TP + FP. Это доля объектов, названными классификатором положительными и при этом действительно являющимися положительными.

Recall - отношение TP к TP + FN. Это то, какую долю объектов положительного класса из всех объектов положительного класса нашёл алгоритм.

Можно также подчеркнуть суть этих метрик, с ростом recall уменьшается precision, и наоборот.

Модели очень удобно сравнивать, когда их качество выражено одним числом. В случае пары Precision-Recall существует популярный способ скомпоновать их в одну метрику - взять их среднее гармоническое. Данный показатель эффективности исторически носит название F1-меры (F1-score). Оценка F1 достигает своего лучшего значения при 1 (идеальная точность и отзывчивость) и худшего при 0.

Результаты обучения разных моделей классификации и метрики приведены в таб. 2.

Выбор модели классификации проводится на основании взвешенного среднего используемых метрик, а также равномерности отношения к разным оценкам. Поскольку в 3-ей модели взвешенное среднее для 3-х метрик имеет лучший результат, а также довольно равномерно распределены оценки (нет особого перевеса к какой-то оценки), то мы можем использовать данную модель классификации в дальнейшей работе. В результате работы классификатора для новости присваивается оценка, к типу которой новость относится (табл. 2).

Тестирование модели

Для проверки работы классификатора, выполнено тестирование на реальных данных. На основании обученного классификатора на предыдущем этапе, был проведен эксперимент. В базе данных были удалены все оценки у новостей, а затем с помощью модели были определены и занесены в БД автоматические оценки 3 тысяч новостей. Экспертом были проверены и исправлены неправильно классифицированные новости. Неправильно было классифицировано около 5% новостей, что считается допустимым и можно сказать, что модель работает достаточно хорошо.

Под неправильной классификацией подразумевается то, что эксперт, пересмотрев все новости 5% новостей посчитал неверно классифицированными и вручную обозначил их принадлежность. Было решено, что на большем количестве новостей нет необходимости переобучать классификатор, поскольку, это не даст скорее всего значительного прироста при классификации, однако значительно замедлит работу классификатора.

Для выполнения анализа было разработано приложение, примеры результатов работы которого представлены на рис. 3 и рис. 4.

Таблица 2 – Классификаторы с метриками

№	Описание	Метрики				
		тип оценки новости	Оценка классификатора			support (кол-во обьектов каждого класса)
			precision	recall	f1-score	
1	TfidfVectorizer, SGDClassifier (random_state=42)	-2	0.40	1.00	0.57	4
		-1	0.76	0.70	0.73	204
		0	0.40	0.57	0.47	35
		1	0.83	0.81	0.82	298
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.77	0.75	0.76	541
2	TfidfVectorizer (), KNeighborsClassifier (n_neighbors=10)	-2	0.20	0.50	0.29	4
		-1	0.69	0.68	0.68	191
		0	0.54	0.40	0.46	67
		1	0.75	0.78	0.76	279
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.70	0.69	0.69	541
3	TfidfVectorizer (ngram_range=(1, 2)) SGDClassifier (penalty='elasticnet', class_weight='balanced', random_state=42)	-2	0.50	0.56	0.53	9
		-1	0.72	0.77	0.74	176
		0	0.62	0.52	0.56	60
		1	0.84	0.82	0.83	296
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.77	0.77	0.77	541
4	TfidfVectorizer () SelectKBest (chi2, k=1000) LinearSVC ()	-2	0.20	1.00	0.33	2
		-1	0.66	0.72	0.69	172
		0	0.42	0.57	0.48	37
		1	0.86	0.75	0.80	332
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.76	0.73	0.74	541
5	CountVectorizer () TfidfTransformer () RandomForestClassifier (n_estimators=100, random_state=0)	-2	0.00	0.00	0.00	0
		-1	0.61	0.80	0.70	143
		0	0.26	0.65	0.37	20
		1	0.93	0.72	0.81	378
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.83	0.74	0.76	541
6	CountVectorizer () TfidfTransformer () SGDClassifier ()	-2	0.40	0.00	0.00	0
		-1	0.73	0.70	0.72	196
		0	0.38	0.51	0.44	37
		1	0.84	0.79	0.81	303
		2	0.00	0.00	0.00	0
		Взвешенное среднее:	0.77	0.74	0.75	541

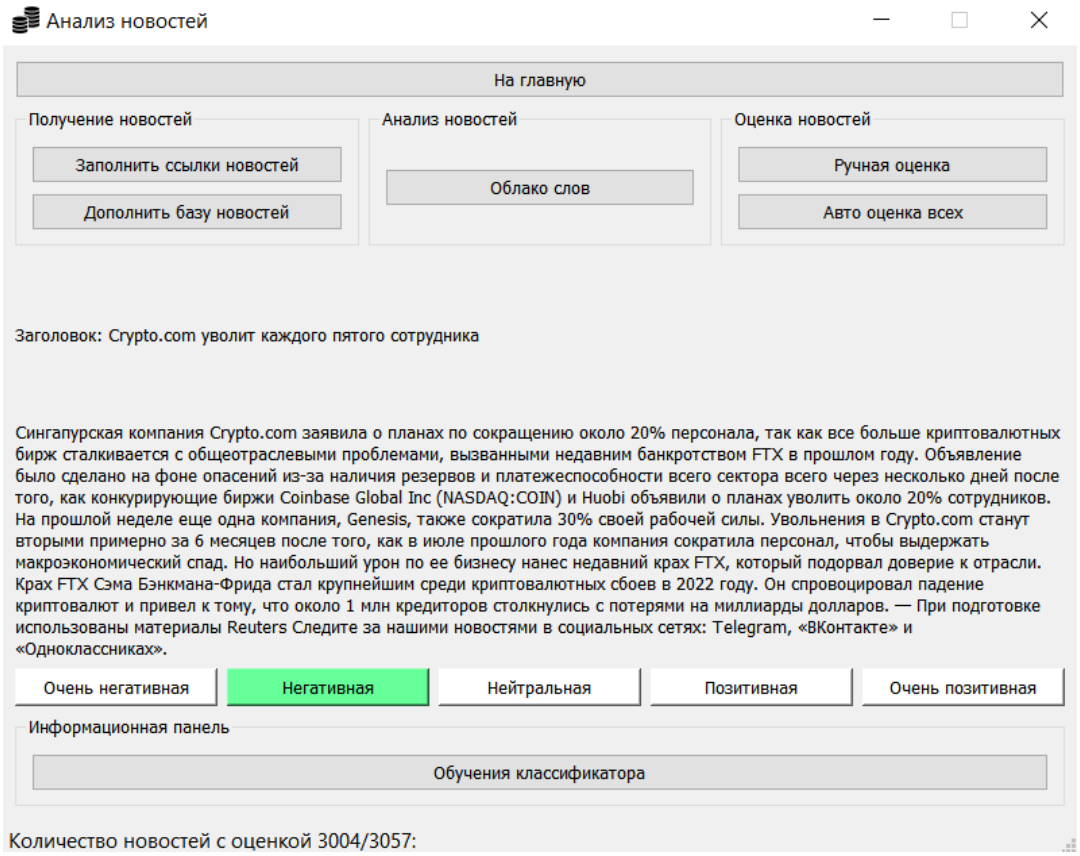

Рисунок 3 – Определение негативной оценки моделью

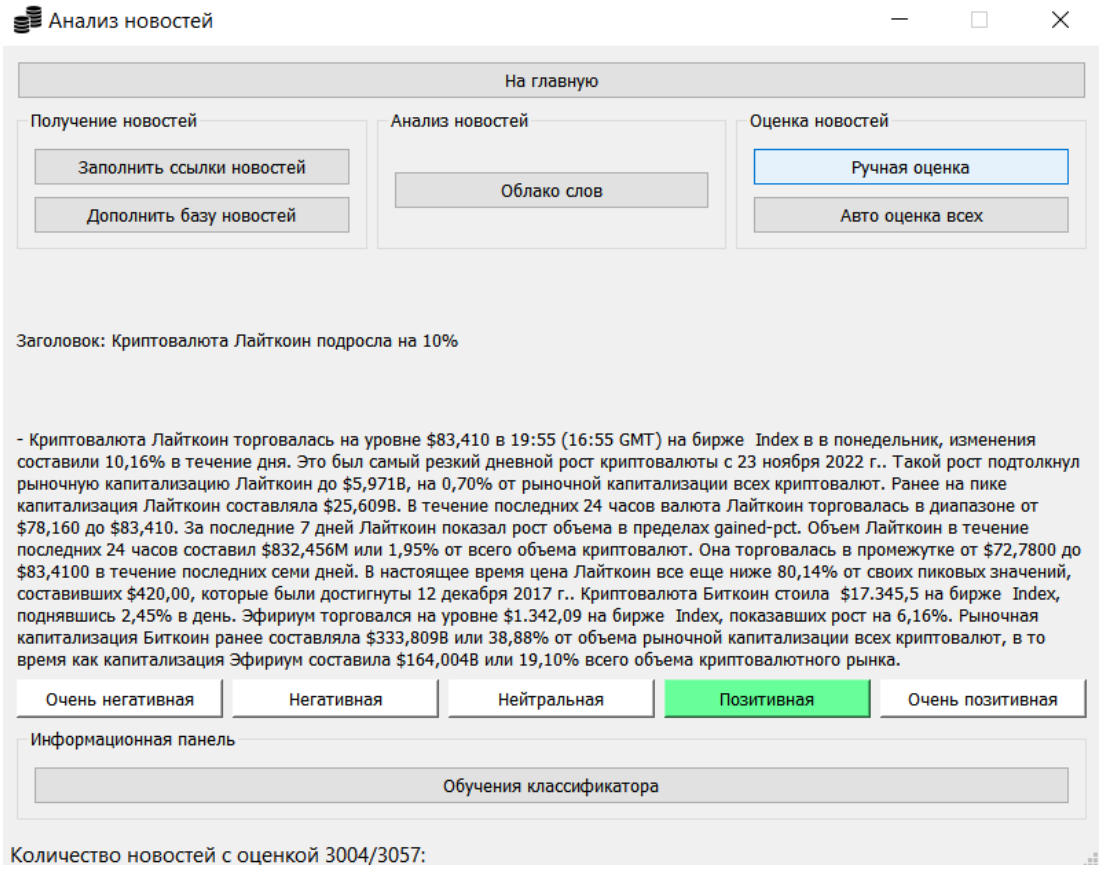

Рисунок 4 – Определение позитивной оценки моделью

Блок «Получение новостей» предназначен для сбора ссылок на новости в БД, а также сбор самих новостей по этим ссылкам.

Блок «Анализ новостей» предназначен для проведения частотного анализа слов, представленного в виде облака слов.

После нажатия на кнопку «Обучение классификатора» осуществлялся процесс обучения классификатора на основании новостей с оценкой из БД, а также в консоли приложения выводились метрики оценки работы всех классификаторов, которые ранее были сведены в таблицу 2.

Блок «Оценка новостей» предназначен для оценки новостей как в ручном, так и в автоматическом режиме. После нажатия на кнопку «Ручная оценка» из БД выбирается новость без оценки и с помощью классификатора определяется возможная принадлежность этой новости и кнопка с подписью вероятной оценки подсвечивается зеленым светом и в этот же момент сама новость выводится выше, что также можно увидеть на рис. 3, 4.

С помощью классификатора эксперту проще проводить оценку новостей в ручном режиме, а из-за того, что результаты достаточно точны, классификацию можно использовать полностью в автоматическом режиме.

Заключение

В условиях глобализации экономики необходимо глубокое понимание изменения котировок рынка ценных бумаг как финансового инструмента, влияющего на различные стороны жизни миллионов людей.

Негативность отсутствия подобных инструментов наглядно продемонстрировал финансовый кризис 2008 г.

Отмечая определенные успехи, достигнутые в прогнозировании котировок рынка ценных бумаг, нельзя не обратить внимания, на то, что учет социальной реакции на различные события увеличил бы эффективность прогнозирования.

Одним из перспективных направлений в данной области является интеллектуальный анализ текста как основа будущего жизнеспособного инструментального решения прогнозирования изменений рынка. Тем более, что исходные текстовые данные широко доступны в Интернете в режиме реального времени.

В результате проведения анализа и оценки новостей нам открывается возможный горизонт дальнейших действий.

Поскольку в рассматриваемом случае были оценены новости, мы их можем связать с ценами в день новости и использовать в дальнейшем в задаче прогнозирования многомерного временного ряда.

Хотелось бы подчеркнуть, что расширение формата анализа новостей может позволить прогнозировать даже более глобальные события, даже не относящиеся к финансовому сектору, например войны, пандемии, кризисы и т.д.

Разработанное приложение полноценно демонстрирует проведение анализа с удовлетворительной погрешностью (около 5% новостей неправильно классифицированы), что позволяет сделать вывод о возможности использовать это приложение в практических целях.

Список используемых источников

1. Андрианова, Е. Г., Новикова, О. А. Роль методов интеллектуального анализа текста в автоматизации прогнозирования рынка ценных бумаг [Электронный ресурс]. – Режим доступа: http://scyberleninka.ru/article/n/rol-metodov-intellektualnogo-analiza-teksta-v-avtomatizatsii-prognozirovaniya-rynka-tsennyh-bumag (дата обращения: 18.11.2022).

2. Gigi Sayfan. Современный веб-скрапинг с BeautifulSoup и Selenium. [Электронный ресурс]. - Режим доступа: https://code.tutsplus.com/ru/tutorials/modern-web-scraping-with-beautifulsoup-and-selenium--cms-30486 (дата обращения: 18.11.2022).

3. Частотный анализ русского текста и облако слов на Python. [Электронный ресурс].- Режимдоступа: habr.com/ru/post/517410/ (дата обращения: 18.11.2022).

4. Метод Стохастического градиентного спуска [Электронный ресурс]. – Режим доступа: https://scikit-learn.ru/1-5-stochastic-gradient-descent/ (дата обращения: 18.01.2023).

5. Метод ближайших соседей в задачах классификации. [Электронный ресурс]. – Режим доступа: https://scikit-learn.ru/1-6-nearest-neighbors/ (дата обращения: 18.01.2023).

6. Метод опорных векторов [Электронный ресурс]. – Режим доступа: https://ww.tutorialspoint.com/scikit_learn/scikit_learn_support_vector_machines.htm (дата обращения: 18.01.2023).

7. Использование CountVectorizer для извлечения объектов из текста. [Электронный ресурс]. – Режим доступа: https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/ (дата обращения: 18.01.2023).

8. Извлечение функций с помощью TF-IDF. [Электронный ресурс]. – Режим доступа: https://www.geeksforgeeks.org/sklearn-feature-extraction-with-tf-idf/ (дата обращения: 18.01.2023).

9. Выбор функций для контролируемых моделей с использованием SelectKBest [Электронный ресурс]. – Режим доступа: https://runebook.dev/ru/docs/scikit_learn/modules/ge

nerated/sklearn.feature_selection.selectkbest (дата обращения: 18.01.2023).

10. Лабинцев, Е. Метрики в задачах машинного обучения [Электронный ресурс]. - Режим доступа: https://habr.com/ru/company/ods/blog/328372/ (дата обращения: 18.01.2023)

11. Осовский, Н. Precision и recall. Как они соотносятся с порогом принятия решений? [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/661119/ (дата обращения: 18.01.2023).

Копица А. В., Савкова Е. О. Анализ новостного фона криптовалют. *Данная статья посвящена анализу новостного фона для прогнозирования цен криптовалют с учетом эмоциональной окраски, связанных с ними новостей. Для этого был определен источник новостей, разработан алгоритм их получения и формирования базы для хранения извлеченной информации. Предложена методика обработки новостей, в результате чего появляется возможность классифицировать новости на категории с различной оценкой тональности текста. В результате разработано программное приложение для анализа новостного фона. С помощью данного приложения производится сбор новостей в автоматическом режиме из новостного источника, связанного с криптовалютными биржами, а затем с помощью выбранного классификатора выполняется автоматическая оценка новостей. Результаты исследований в дальнейшем могут быть использованы для разработки системы более точного прогнозирования цен криптовалют.*

Ключевые слова: *криптовалюта, новость, классификация, прогнозирования, анализ, метрика, новостной фон.*

Kopitsa A., Savkova E. Analysis of the news background of cryptocurrencies*. This article is devoted to the analysis of the news background for predicting the prices of cryptocurrencies, taking into account the emotional coloring of the news related to them. To do this, a news source was identified, an algorithm for obtaining them and forming a database for storing the extracted information was developed. A method of news processing is proposed, as a result of which it becomes possible to classify news into categories with a different assessment of the tonality of the text. As a result, a software application has been developed to analyze the news background. With the help of this application, news is collected automatically from a news source associated with cryptocurrency exchanges, and then an automatic news assessment is performed using the selected classifier. The research results can be used in the future to develop a system for more accurate prediction of cryptocurrency prices.*

Keywords: *cryptocurrency, news, classification, forecasting, analysis, metric, news background.*

Статья поступила в редакцию 15.12.2022
Рекомендована к публикации профессором Аноприенко А.Я.