# ПРОГНОЗИРОВАНИЕ СОБЫТИЙ С ПОМОЩЬЮ ЛЕНТЫ TWITTER

***Д.И. Самаль***
*Заведующий кафедрой электронных вычислительных машин БГУИР, кандидат технических наук, доцент*

***В.А. Прытков***
*Декан факультета компьютерных систем и сетей БГУИР, кандидат технических наук, доцент*

***А.И. Трубчик***

*Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь*
*E-mail: apri@tut.by, samal@bsuir.by, prytkov@bsuir.by*

*Absract.* The paper explores the possibility of forecasting based on the intelligent analysis of a feed of the Twitter social network. The object of this analysis is the Bitcoin crypto currency market. The theoretical basis for this forecasting approach is the efficient market hypothesis, which states that new information can be used for gaining economic advantage. The source of this information is the sentiment analysis of Twitter messages. The work proposes a new algorithm of determining the emotions of Twitter messages with three possible grade levels: negative, neutral and positive. Thanks to comprehensive correlation analysis and forecasting experiments, it was determined that there is a relation between the analyzed source data, and the prediction accuracy of Bitcoin market movements reached 63,27%.

Теоретическим обоснованием значительной части критики финансового прогнозирования такими методами как, например, технический анализ является гипотеза эффективного рынка (англ. – efficient market hypothesis) [1]. Согласно этой гипотезе, цены на финансовых рынках точно отражают всю имеющуюся на данный момент информацию и следовательно, стабильная генерация прибыли выше средней по рынку невозможна. Но, в то же время, важным выводом из гипотезы является предположение, что если удастся собрать некую информацию, не входящую в рынок, то она может быть использована для получения экономического преимущества.

Одним из способов получения новой информации является интеллектуальных анализ данных, которые лишь опосредованно связаны с анализируемыми активами. В рамках настоящей работы исходными данными для анализа выступают сообщения социальной сети Twitter и торговая история криптографической валюты Bitcoin. Аналогичные исследования, затрагивающие анализ тональности текстов Twitter, проводились применительно к фондовым биржам [2 - 4], прогнозированию популярности кинофильмов [4], а так же оценки общественного мнения [5]. Первым серьезным исследованием на эту тему была работа Боллена с коллегами [2], начиная с которой многие

исследователи обратили внимание на анализ данных Twitter, а именно эмоциональную окраску текста (тональность), и его успешное применение в прогнозировании различной направленности.

Целью настоящей работы являлось выявление корреляции между сообщениями в сети Twitter – «твитами» и изменением курса криптографической валюты Bitcoin. В рамках работы были решены задачи:

- загрузки, очистки и подготовки данных Twitter и Bitcoin;
- создания алгоритма оценки тональности сообщений Twitter;
- исследования взаимосвязи между данными из «твитов» и торговой историей Bitcoin;
- прогнозирование движения котировок Bitcoin с учётом данных, полученных из сообщений сети Twitter.

Первым этапом проведенного исследования являлось формирование базы сообщений сети Twitter. Из-за ограничений Twitter формирование базы заняло 180 дней - с 20 марта по 16 сентября. Всего в базу было загружено 10118 204 сообщений от 868 866 пользователей. В среднем у анализируемых пользователей насчитывалось по 2026 подписчиков, 31,4% сообщений являлись ретвитами, 1,4% – ответами на другие сообщения. Отбор сообщений Twitter для последующего анализа производился по ключевым словам: «bitcoin», «btc», «биткоин», «биткойн»

Особенностями сообщений сети Twitter являются их краткость, использование сленга, эмотиконов и хеш-тегов. Соответственно, для оценки тональности текста потребовалось производить их предварительную очистку и разработать собственный алгоритм анализа тональности текста. Анализатор тональности приписывает тексту положительный или отрицательный балл настроения (эмоции), сопоставляя его с лексиконом тональности. Используемый лексикон состоит из набора слов, отнесённых к двум классам - положительному или отрицательному. Он построен на основе базы субъективных маркеров для английского языка [6], разработанной Уилсоном с соавторами [7] для системы OpinionFinder. База субъективных маркеров содержит 8224 слов английского языка, отмеченных как положительные либо как отрицательные в двух дополнительных градациях, которым присвоена оценка 1 (слабо-положительное слово либо слабо-отрицательное) или 2 (сильно-положительное слово либо сильно-отрицательное).

Для увеличения точности анализа текста в лексикон были добавлены специальные теги, обозначающие эмотиконы, которым присвоен самый высокий уровень оценки равный 3. Алгоритм определения балла настроения исходной строки сообщения str состоял из следующих шагов:

1 Подготовка текста *str*:

- удаление стоп-слов («I», «the» и им подобных);
- удаление специальных символов («?», «!» и т.п.);
- удаление ссылок и идентификаторов, начинающихся с символа «@»;

– удаление символов хэш-тегов (пример «#apple»);
– замена эмотиконов («:)», «:-)», «:(» и других) на специальные теги.

2 Разделение строки str на отдельные слова (токенизация).

3 Выделение основы каждого слова (т.н. – стемминг).

4 Поиск оценок слов в лексиконе и сохранение в переменные pos и neg их максимальных значений (от 1 до 3).

5 На основе максимальной оценки одного из слов принятие решения:
– если *pos* и *neg* равны, то тональность твита считается нейтральной;
– если *pos* больше *neg,* то тональность твита – положительная;
– если *neg* больше *pos,* то тональность твита – отрицательная;

6 Переход к следующему тексту (сообщению) и повтор с п.1.

Для оценки точности анализатора были отобраны 505 твитов и вручную выставлены соответствующие категории. Зачастую встречались сообщения, которые сложно было отнести к какой-либо категории тональности. В итоге корректность алгоритма на проверочной выборке составила – 74,26% процентов. Для сравнения анализаторы на базе NaiveBayes и др. имеют точность в среднем около 80 процентов. Следует отметить, что хоть тестовые данные и отбирались и оценивались вручную, но из-за беспорядочного набора слов во многих твитах зачастую даже человеку было сложно определить к какой полярности он относится - положительной, негативной или нейтральной.

В качество источника данных торговых операций выступила биржа BTC-E и торговая пара BTC/USD - это Bitcoin торгуемый относительно доллара США.

История торгов от площадки загружается в виде хронологии последних исполненных заявок (ордеров). Соответственно эти данные и данные из Twitter необходимо было синхронизировать во времени. Для этой цели они были сгруппированы в торговые периоды с интервалом в 1 час. Попытка сделать интервал меньше приводила к появлению пропусков в данных, а при увеличении интервала – ряд сглаживался, что приводило к потере информативности.

Таким образом для торговых данных были получены новые переменные для каждого периода: цена открытия, цена закрытия, максимальная цена, минимальная цена, количество сделок, объем сделок. Для данных из сети Twitter – общее количество сообщений, количество положительных, отрицательных и нейтральных твитов, их процентное соотношение с общим количеством.

В ходе экспериментов полученные временные ряды подвергались дополнительным преобразованиям для увеличения корреляции между ними. Одним из вариантов подобных преобразований являлась нормализация данных по цене закрытия торгов. На рисунке 1 представлен пример нормированного временного ряда.

Для расчета степени корреляции между двумя наборами данных можно использовать разные меры, которые демонстрируют различные результаты на одних и тех же тестовых данных. В процессе исследований нами применялись коэффициент корреляции Пирсона и коэффициент корреляции ранга Кендалла.

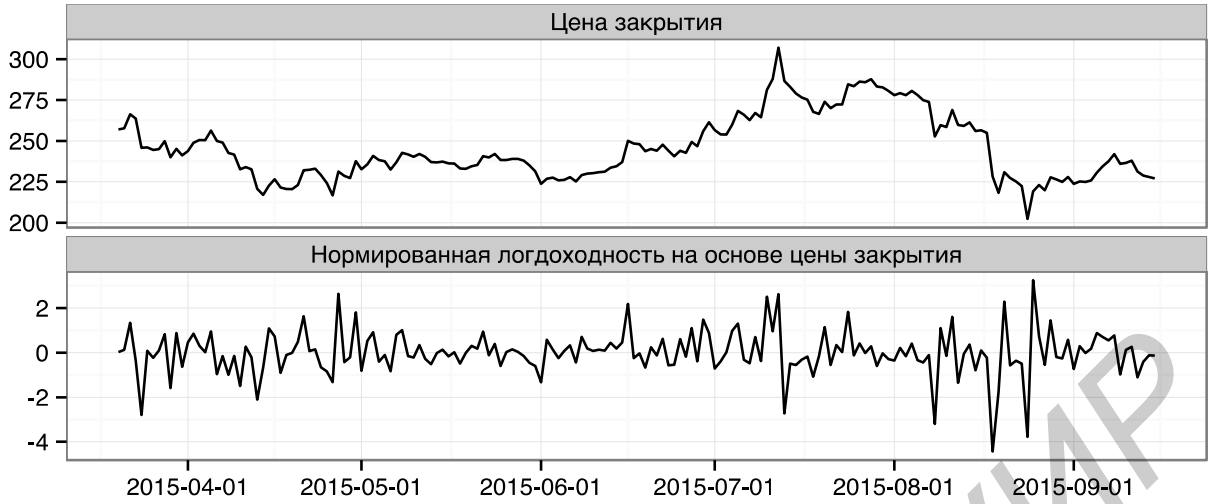


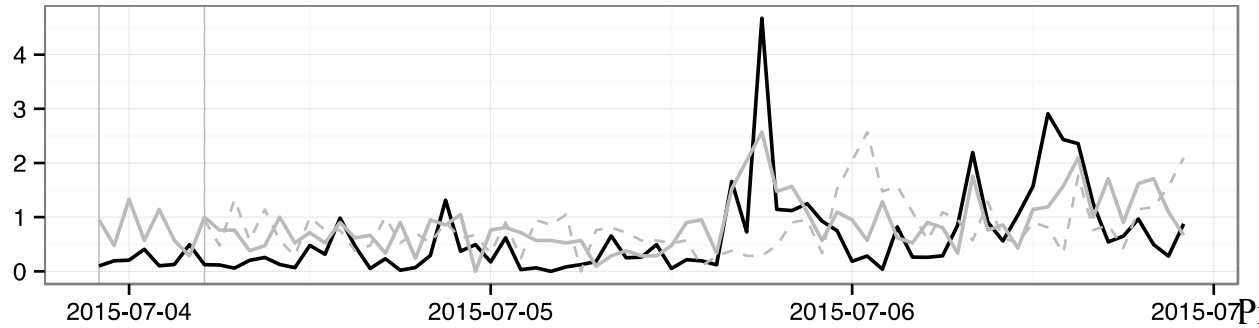

Рис. 1. Преобразование данных на примере цены закрытия

В ходе серии экспериментов данные, образующие временной ряд с целью более эффективного выявления временных закономерностей делились на сегменты (интервалы) и к ним локально по отдельности применялся корреляционный анализ. Таким образом производился массовый автоматический анализ тысяч сегментов различных временных рядов (без преобразований, с преобразованиями, с фильтрами и без).

В результате поиска интересующих корреляций на коротких промежутках в 24-48 часов были получены тысячи результатов с уровнем линейной корреляции выше 0,9. Ниже для наглядности приведен один пример на большем интервале в 72 часа – рисунок 2.

Для задачи прогнозирования, как и для предыдущего эксперимента, использовалась история торговой пары BTC/USD и сообщений Twitter. Самой простой разновидностью прогноза котировок является предсказание направления движения временного ряда. В данном случае это движение цены закрытия торгового периода и результатом прогноза будет выбор одного варианта из двух классов: «*вниз*» (0) и *«вверх»* (1). Это удобно с позиции реализации метода машинного обучения, так как есть возможность использовать бинарный классификатор, а в данном случае использовался метод опорных векторов с радиальной базисной функцией Гаусса в качестве ядра. Данное ядро похоже на нейронную сеть с радиальными базисными функциями и подходит для довольно широкого ряда задач, в том числе и прогнозирования.

Наилучшие результаты были получены с помощью группы признаков: разницы логарифмов («логдоходность») от цены закрытия торгового периода за 5 последних периодов. Другие возможные признаки, например на основе объема торгов, количества заявок и тому подобных, не показали какого-либо вклада в повышении качества прогноза, поэтому в статье не приводятся. Все признаки нормировались с помощью z-оценки.




Ри-

Рис. 2. График корреляции объема торгов и количества сообщений (цвет - серый), без фильтра в интервале времени 72 часа. Лаг между временными рядами 7 часов.

Для проверки того, насколько успешно предложенная модель способна работать на независимом наборе данных, использовалась кросс-валидация, которую иногда называют перекрестной проверкой. В нашем случае это кросс-валидация по К-блокам.

Качество модели оценивалось с помощью таблицы контингентности. В ней содержится информация сколько раз система приняла верное и сколько раз неверное решение по данным заданного класса. На основе её вычислялись характеристики: точность, каппа Коэна и F-мера.

Общее количество торговых периодов используемой выборки составило 4320, а данные были разделены на тренировочные и тестовые в пропорциях 80% и 20% соответственно. Результаты машинного обучения только на основе торговой истории приведены в таблице 1.

Таблица 1. Результаты только на основе торговых данных BTC-E

| | Действительный | | | *A* | *κ* | *F* |
|---|---|---|---|---|---|---|
| Прогноз | «вверх» | «вниз» | Всего | | | |
| «вверх» | 241 | 194 | 435 | 56,49% | 0,1301 | 0,5644 |
| «вниз» | 178 | 242 | 420 | | | |
| Всего | 419 | 436 | 855 | | | |

Доля предсказанных классов на основе данных BTC-E составила 56,49%. Это значение говорит об низкой точности прогнозирования. Можно утверждать, что такой прогноз лишь ненамного отличается от метода «подбрасывания монетки» в попытках угадать следующее движение котировки. Мера согласия каппа в данном случае имеет значение всего 0,1301 - это означает, что согласия почти нет.

В соответствии с гипотезой эффективного рынка, использование новой информации может улучшить точность прогноза котировок. Новой информацией в нашем случае являются сообщения Twitter и полученных из них оценки тональности. В результате экспериментов были выделены признаки, которые максимально повлияли на модель машинного обучения: доли положительных и отрицательных сообщений за период.

Результаты машинного обучения с использованием информации из сообщений Twitter приведены в таблице 2.

Таблица 2. Результаты на основе торговых данных BTC-E и информации из Twitter

| | Действительный | | | $A$ | $\kappa$ | $F$ |
|---|---|---|---|---|---|---|
| Прогноз | «вверх» | «вниз» | Всего | 63,27% | 0,266 | 0,629 |
| «вверх» | 266 | 161 | 427 | | | |
| «вниз» | 153 | 275 | 428 | | | |
| Всего | 419 | 436 | 855 | | | |

В данном случае по сравнению с предыдущим экспериментом доля предсказанных сообщений значительно увеличилась и по сравнению со значениями из таблицы 1 составила 63,27%. Мера согласия каппа достигла низкого уровня 0,266 и 0,233 соответственно, что говорит присутствии небольшого влияния признаков, созданных на основе сообщений Twitter.

Также были проведены эксперименты с различной величиной лага между торговыми данными и данными Twitter. Результаты показали резкое снижение доли угаданных классов.

Благодаря массовому корреляционному анализу и эксперименту по прогнозированию фактически было установлено наличие взаимосвязи между данными, которые можно извлечь из социальной сети Twitteri и торговой истории криптовалюты Bitcoin.

Используемые методы подготовки данных, в частности анализ тональности текста, позволил достичь точности прогноза в 63,27%. К сожалению, этого недостаточно, чтобы получить преимущество перед другими трейдерами на торговых площадках Bitcoin. Последние вычитают комиссию за каждый исполненный ордер в 0,1% (применяется дважды для обеих сторон сделки, то есть к каждому из ордеров на покупку и продажу), что полностью нивелирует полученное преимущество в виде прогнозирования движения котировок рынка.

В дальнейшем следует продолжить работу над подходами к очистке и подготовке данных. Например, в исходном наборе данных, используются только те сообщения, которые содержат слова «bitcoin», «btc» и «биткоин». Вполне возможно, что объема сообщений по этим ключевым словами недостаточно для точного анализа и прогнозирования рынка и одним из вариантов усовершенствования прогноза может быть расширение базы поступающих твитов с помощью новых ключевых слов. Безусловно, это увеличит размер базы данных в несколько раз, однако масштабируемость применяемых инструментов позволяет это сделать.

*Литература*

[1]. Sewell, M. History of the Efficient Market Hypothesis / M. Sewell // Research Note RN/11/04. — 2011.

[2]. Bollen, J. Twitter mood predicts the stock market / J. Bollen, H. Mao, X.J. Zeng // ArXiv e-prints. — 2010.

[3]. Porshnev, A. Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis / A. Porshnev, I. Redkin, A. Shevchenko // Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. — 2013. — Pp. 440–444.

[4]. Arias, Marta. Forecasting with Twitter Data / Marta Arias, ArgimiroArratia, Ramon Xuriguera // ACM Trans. Intell. Syst. Technol. — 2014. — 1. — Vol. 5, no. 1. — Pp. 8:1–8:24.

[5]. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. / B. O'Connor, R. Balasubramanyan, B. Routledge, N.A. Smith // ICWSM. — 2010. — Vol. 11, no. 122-129. — Pp. 1–2.

[6]. Subjectivity Lexicon [Электронный ресурс]. — Электронныедан- ные. — Режимдоступа: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/. — Датадоступа: 03.03.2016.

[7]. Wilson, T. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis / T. Wilson, J. Wiebe, P. Hoffmann // Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. — HLT '05. — Association for Computational Linguistics, 2005. — Pp. 347–354.