Роль методов интеллектуального анализа текста в автоматизации прогнозирования рынка ценных бумаг[1]

Е. Г. Андрианова, О. А. Новикова

Московский технологический университет (МИРЭА)
119454, Москва, пр-т Вернадского, 78

e-mail: *andrianova@mirea.ru, novikova@mirea.ru*

Аннотация. В статье выполнен анализ работ по прогнозированию изменений рынка ценных бумаг на основе методов интеллектуального анализа текста. Рассмотрена роль методов интеллектуального анализа текста в автоматизации исследований на основе традиционных методологий прогнозирования: техническом и фундаментальном анализе. Сделан вывод об актуальности и необходимости расширения традиционных методов средствами интеллектуального анализа текстовых и новостных данных, повышающего эффективность прогноза котировок рынка ценных бумаг. Обоснована необходимость учета данных, извлекаемых из новостных и текстовых сообщений, характеризующих эмоциональную окраску событий, влияющих на изменения рынка (поведенческая экономика, эмоциональный анализ). Сделан обзор применения методов анализа и прогнозирования изменений рынка ценных бумаг, основанных на таких моделях, как машинное обучение, нейронные сети, нечеткая логика, метод опорных векторов регрессии, генетического программирования сетевых приложений и др.
Ключевые слова: метод интеллектуального анализа текста, рынок ценных бумаг, поведенческая экономика, новостные сообщения.

1. Введение

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 времен-

[1] Работа выполнена за счет финансирования Министерством образования и науки РФ конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов. Номер проекта 28.2635.2017/ПЧ, наименование «Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

ных 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. 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, поэтому построение эффективного прогноза возможно только при совокупном использовании как технического, так и фундаментального анализа, и в первую очередь эмоционального анализа. 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.

Работа посвящена рассмотрению современного состояния исследований в области прогнозирования котировок рынка ценных бумаг на основе методов и инструментальных средств интеллектуального анализа текста и новостных сообщений.

2. Основные подходы прогнозирования котировок рынка ценных бумаг

2.1. Традиционные подходы

Рыночный прогноз непрост из-за сложной природы рынков и широкого спектра факторов влияния.

Главными инструментами прогнозирования котировок рынка ценных бумаг являются методы технического и фундаментального анализа. Применению этих методов посвящено большое число исследований.

Приведем работы [1–5], в которых обсуждаются сильные и слабые стороны применения технического и фундаментального анализов для рынка ценных бумаг. В данных работах отмечается, что технические аналитики полагаются на количественные исторические рыночные данные. Этот подход в значительной степени автоматизирован. Существуют компьютерные программы (роботы-предсказатели), автоматизирующие методы технического анализа для поддержки принятия решения при покупке или продаже ценных бумаг. Слабым местом данных средств является отсутствие возможности учета оперативной или долговременной информации из внешних источников и работа только на ограниченных коротких интервалах времени. Фундаментальные аналитики учитывают влияние внешних событий на поведение рынка, таких как, слияние/поглощение компаний, IPO и пр. Необходимость рассмотрения фундаментальных данных в расчетах прогнозирования неоднократно отмечалась такими успешными аналитиками, как Уоррен Баффет (Warren Buffet) [1]. Однако автоматизация подобного учета, по-прежнему остается сложной задачей.

Использование методик анализа рынков на основе извлечения правил можно условно считать промежуточной областью, соединяющей технический и фундаментальный анализ [2]. Данные методики ориентированы на работу с данными, извлеченными из финансовых новостей, которые, несмотря на текстовый формат, зашумлены меньше, чем обычные новости. Понятно, что сужение области исследования до рассмотрения только финансовых новостей упрощает автоматизацию извлечения и анализа данных, что является весьма существенным для валютного рынка.

В работе [3] исследовалось влияние частотных данных (макроновостей) на резкие изменения курсов валюты. Данные собирались из плановых структурированных макроэкономических новостей, отражающих такие индексы, как уровень безработицы в стране, уровень инфляции и т. д. Анализировались внутридневные данные (с частотой в 5 минут) для четырех валют за период 2005–2010 гг. Авторы доказали, что 9–15% резких изменений курсов валюты напрямую связаны с негативными новостями об экономике США. В исследовании применялась методика обнаружения прыжков для извлечения скачков валютных курсов и одновременных скачков по валютам. После установления точного времени внутрисуточных скачков валюты, авторы попробовали определить, можно ли предсказать курс валюты после его скачка, связанного с новостями. В качестве математического аппарата использовался регрессионный анализ.

Учитывая, что в ежедневной медиасфере перегрузка информации становится серьезной проблемой, авторы [5] предложили систему, основанную на модели информирования инвесторов о важных политических и экономических новостях в реальном времени. Модель базируется на взвешенных правилах ассоциации, использующихся для определения важности новости для инвесторов. Во время обучения на реальных данных алгоритм взвешенных ассоциативных правил обнаруживает множественные термины, часто одновременно появляющиеся в одном заголовке новостей. Интуитивное значение правила ассоциации заключается в том, что заголовки новостей в базе данных, которые содержат набор ключевых слов X, также содержат набор ключевых слов Y. Вес wks для отдельного ключевого слова p определяется как $wks = (\sum_j pf_{p,j}) \big/ n$, где $pf_{p,j}$ обозначает колебание цены закрытия акции на следующий торговый день и ключевое слово p появляется в заголовке (-ах) новостей в j-м дне, а n представляет собой общее количество дней, в которых ключевое слово p появляется в заголовках новостей. Таким образом, эти веса помогают решить, влияют ли ключевые слова в заголовках новостей на результат торговли. Модель была калибрована на реальных данных, вычисляя правила ассоциации весов и строя базу данных весов ключевых слов. Используя базу данных весов ключевых слов, модель обрабатывает новые заголовки новостей и вычисляет правила принятия решений о том, следует ли распространять новости среди инвесторов. Экспериментальные результаты показывают, что предлагаемая модель обеспечивает приемлемую производительность в анализе новостей в режиме реального времени [5].

В работе [6] приведены результаты изучения прогностических возможностей и прибыльности применения 60 технических правил торговли на пяти основных рынках Юго-Восточной Азии в Сингапуре, Малайзии, Таиланде, Индонезии и на Филиппинах в период с 1991 по 2008 г. Результаты этих исследований показали, что, например, правила переменной скользящей средней (VMA), правила фиксированной скользящей средней (FMA) и технические правила пробоя торгового диапазона (TRB) являются успешными для развивающихся рынков. Авторы в целом подтвердили успешность применения методов технического анализа для краткосрочных временных промежутков, хотя для различных рынков Юго-Восточной Азии успешность применения методов технического анализа имела незначительные расхождения [6].

2.2. Развитие традиционных подходов

Несовершенство существующих инструментов технического анализа подталкивает к поиску новых моделей и методов анализа и прогнозирования поведения рынков.

Например, основанных на различных видах или сочетаниях алгоритмов машинного обучения, таких как нейронные сети [1, 7–9], нечеткая логика [10], метод опорных векторов регрессии [11], набора правил на основе генетического программирования сетевых приложений [12].

В статье [1] описан новый подход к фундаментальному анализу данных для выявления связей между поведением рынка и внешней информацией, выдвинута гипотеза о возможной логической актуальности некоторой внешнеэкономической информации и ценовых движений валютной пары USD/GBP. Авторы [1] с использованием нейронных сетей провели ряд экспертиз с целью определения существования правдоподобных связей между внешними источниками информации и ценовыми движениями валютной пары. В работе [1] утверждается, что нейронные сети идеально подходят для моделирования движения цен. Они могут математически моделировать поведение цены, и у них есть возможность извлекать информацию из больших объемов данных, которая необходима для сложных сигналов, таких как движение финансовых цен.

Авторы [7] применяют как параметрические (нейронные сети с активными нейронами), так и непараметрические (аналоговые комплексообразующие) методы самоорганизующегося моделирования для ежедневного прогнозирования валютного рынка. В работе [7] предлагается комбинированный подход, при котором параметрические и непараметрические методы самоорганизации объединяются последовательно, используя преимущества отдельных методов с целью повышения их производительности. В исследовании сначала рассматривается параметрический подход для нейронных сетей с активными нейронами. Затем обсуждается непараметрический метод самоорганизующегося моделирования, известный как Analog Complexing (Аналоговое Комплексирование) и, наконец, описывается новый гибридный метод, вытекающий из последовательной комбинации двух предыдущих. Авторы [7] показали, что комбинированный метод дает хорошие результаты и превосходит отдельные при тестировании с двумя обменными курсами: американский доллар и немецкая марка по отношению к британскому фунту.

В работе [8] рассматривается методология проектирования и тестирования биржевых торговых систем (роботов), использующих технологии мягких вычислений (нечеткой логики), и искусственных нейронных сетей (ANN), разработанных в [9]. Представленная в работе [8] методология четко отделяет процесс обучения нейронных сетей и выбора параметров от процесса невыборочного бенчмаркинга. Для обучения и тестирования ANN данные должны быть логически (или физически) разделены как минимум на два набора: набор обучения и набор тестирования. По сути, основной принцип заключается в том, чтобы охватить как можно более разнообразную рыночную активность (с длительным обучением), сохраняя при

этом как можно более длинное окно тестирования (чтобы увеличить уверенность в модели).

Из-за нелинейности и изменчивости характеристик временных рядов цен на фондовом рынке традиционные методы моделирования, такие как интегрированная модель авторегрессии — скользящего среднего (ARIMA, модель Бокса — Дженкинса), не подходят для прогнозирования цен на фондовом рынке. Хуанг (Huang) и др. в [11] представили непараметрическую модель на основе теории хаоса для прогнозирования будущего поведения валютных курсов. Теория хаоса предполагает, что поведение финансовых рынков хаотично и обеспечивает основу для учета динамики нелинейных систем, которые могут выявлять скрытые шаблоны и тенденции в финансовых данных, которые не могут быть получены обычными статистическими методами.

Двухступенчатая модель, предложенная в [11], сочетает в себе фазовое пространство, реконструированное из одномерных хаотических временных рядов в комбинации с векторной регрессией. На первом этапе метод вложения координат задержки преобразует необработанные временные ряды в фазовое пространство, подходящее для прогнозирования. На втором этапе стандартная векторная регрессия поддержки (SVR) применяется к преобразованной серии для окончательного прогнозирования. Радиальная базисная функция используется как ядро SVR. Полученная модель, получившая название “Chaos-SVR”, превосходила обычную SVR, нейронную сеть обратного распространения (“BPNN”) и ее комбинацию с преобразованным фазовым пространством (“Chaos-BPNN”) на общих метриках для измерения производительности модели, таких как MSE (средняя квадратичная ошибка), RMSE (среднее значение квадрата ошибки) и MAE (средняя абсолютная ошибка). Были предложены расширения этой модели, такие как трехступенчатая модель, которая включает в себя оптимизацию гиперпараметров SVR с использованием «алгоритма светляка». Подводя итог, предложенная модель “Chaos-SVR” извлекает ключевые особенности динамики обменного курса и лучше подходит для финансового прогнозирования.

Фундаментальные данные сложнее использовать в качестве входных данных, особенно когда они не структурированы. Одним из подходов, который облегчает задачу извлечения структурированных данных из неструктурированных, является разработка специализированных поисковых систем, таких как семантическая поисковая система финансовых новостей [13]. Тем не менее по-прежнему остается проблемой извлечь из текста достоверный смысл, но, к сожалению, существующие поисковые системы, подобные [13], ограничиваются извлечением доступных числовых данных в соответствующих текстах.

2.3. Поведенческая экономика

Поведенческая экономика — это область экономики, которая изучает влияние социальных, когнитивных и эмоциональных факторов на принятие экономических решений отдельными лицами и учреждениями и последствия этого влияния на рыночные переменные (цены, прибыль, размещение ресурсов). Поведенческая экономика (behavioral economics) — имеет междисциплинарный характер и находится на стыке двух наук: психологии и экономики, что позволяет выявить психологическую основу и динамику принятия решений человеком и другими агентами.

Средства массовой информации не только сообщают о состоянии рынка, но и активно влияют на динамику рынка, основываясь на публикуемых новостях [14]. Робертсон (Robertson) и его коллеги исследовали влияние на рынок не самой новости, а время суток и процент новостей, когда и на которые рынок реагировал наиболее значимо [14]. Они определили гетерогенные временные ряды средней цены акций, объема, количества сделок, и собрали данные по каждой минуте в рамках бизнес-дня (от начала до конца торгового периода). Затем был определен временной ряд доходности, использованный для определения периодов высокой доходности, свидетельствующих о реакции рынка на новости. Кроме того, изменение временных рядов объема и изменение временных рядов количества торгов использовались для обнаружения внезапного увеличения объема и количества торгов в качестве индикаторов реакции рынка на новости. Временной ряд волатильности, вычисляющий годовую волатильность акций, использовался для определения временных моментов быстрого изменения цены акций, как потенциального следствия реакции рынка на новости. Для бинаризации событий были определены следующие ряды: ряд точечных событий (EPP), ряд событий с предшествующими новостями (ENPP), ряд событий без новостных событий (EWPP).

На основе данных рядов в [14] были разработаны тесты проверки влияния немакроэкономических новостей на рынок. В частности, тест «Пропорция событий, связанных с новостями (RERNE)» использовался для определения событий с высокими значениями RERNE, указывающими на предшествующие новости, ответственные за события. «Т-тест событий (ETT)» использовался для выполнения Student's t-test между распределением вероятности событий, происходящих с новостями, и распределением вероятности событий, происходящих без новостей.

Целью ETT было проверить нулевую гипотезу о том, что появление событий не зависит от появления новостей. «Т-тест новостей (NTT)» использовался для выполнения Student's t-test между распределением вероятности появления новостей перед событием и распределением вероятности появления новостей в течение каждого события. Цель NTT состояла в том, чтобы проверить нулевую гипотезу о том,

что появление новостей перед событиями совпадает с появлением новостей в обычном режиме.

Интерпретация людьми одной и той же информации сильно различается. На процесс интерпретации влияют когнитивные предубеждения, репрезентативная или информационная предвзятость, а также и другие предсказуемые человеческие ошибки при анализе и обработке информации [15]. Авторы [15] построили модель количественной оценки неправильной интерпретации информации с учетом аналогичных предыдущих ошибок лица, принимающего решения, показав влияние прошлых ошибок прогноза на результат текущего прогноза. Полученные результаты подтвердили теоретические результаты, предсказанные теорией когнитивного диссонанса, влияние которых на экономику дает ощутимый эффект.

Поведенческий подход присущ не только деятельности человека, но и деятельности фирм, рынков и регионов, также не всегда демонстрирующих рациональное поведение и принятие решений. Авторы [16] проанализировали текстовое содержимое ежедневных каналов Twitter с помощью инструментов отслеживания настроения, а именно OpinionFinder, который измеряет положительное и отрицательное настроения и Google-профиль состояний настроения (GPOMS), который измеряет настроение с точки зрения 6 измерений (спокойный, тревожный, уверенный, жизнерадостный, добрый и счастливый). Также были перекрестно проверены полученные временные ряды настроения, проведено сравнение их способности обнаруживать реакцию общественности на президентские выборы и День благодарения в 2008 г. Затем, используя анализ причинности Грейнджера и самоорганизующуюся нечеткую нейронную сеть, было проведено исследование гипотезы о том, что состояние общественного настроения, измеряемое временными рядами OpinionFinder и GPOMS, предсказывает изменения значений индекса закрытия DJIA. Результаты экспериментов показывают, что точность предсказаний DJIA может быть улучшена за счет включения конкретных измерений общественного настроения.

В настоящее время поведенческая экономика упоминается в литературе достаточно поверхностно, однако совместное использование методов поведенческой экономики с методами и инструментами интеллектуального анализа может привести к интересным результатам.

2.4. Эмоциональный анализ

Анализ эмоциональной окраски текстовых и новостных сообщений увеличивает свою актуальность параллельно с ростом подписчиков социальных сетей, профессиональных сообществ, с увеличением пользователей глобальной сети и их временем нахождений в ней. Эмоциональный анализ предоставляет широкие возможно-

сти для прогнозирования в различных предметных областях. Например, в [17] была изучена паника на голландском рынке ценных бумаг (2007–2009) и ее влияние на алгоритмическую торговлю.

В работе [18] предлагается использовать онтологические методы для анализа настроений в сообщениях Twitter, извлекая исходную информацию из популярных онлайн-сервисов глобальной сети Интернет (микроблогинг). При этом дается оценка настроений для каждого отдельного понятия в текстовом сообщении. Таким образом, авторы попытались перейти от эмоциональной окраски куска текста к сопоставлению его с неким продуктом.

В работе [21] предложена модель контекстной энтропии для расширения набора исходных слов из онлайн-статей о рынке ценных бумаг, обнаружения похожих эмоциональных слов и их интенсивностей. Это было достигнуто путем вычисления сходства между начальными словами и словами-кандидатами из контекстуальных распределений на основе меры энтропии. Когда начальные словари были расширены, то они также стали использоваться для классификации настроений новостных статей.

Эмоциональный анализ текста не подразумевает одномерности, т. е. положительности-негативности, а может быть многомерным [22]. В [23] предложена следующая шкала: результат оценки имеет две составляющие эмоции и чувства. Чувства имеют положительную или отрицательную полярность. Эмоции состоят из четырех аффективных измерений (концепция эмоций Мински (Minsky)): приятность, внимание, чувствительность и способность. Каждое измерение имеет шесть уровней градации. Каждый уровень представляет собой эмоциональное состояние души некоторой интенсивности.

В статье [24] описана модель, разграничивающая доходность фирмы от основного и оборотного капитала. Оборотный капитал, в отличие от основного капитала, подвержен рыночным настроениям, которые косвенно определяются «зашумленной» информацией, которая является сложной для адекватного реагирования.

Результаты исследования подтвердили, что в долгосрочно периоде основной капитал является основополагающим в прогнозировании. А факторы настроения рынка и «шума» можно использовать для прогнозирования доходности рыночного портфеля в краткосрочном периоде, с учетом закономерного возрастания уровня рисков [24].

3. Системы прогнозирования на основе текстовых данных

В работе [25] приведена сравнительная таблица входящих текстовых данных, которые использовались разными исследователями для прогнозирования рынков (рис. 1). В [3] в качестве входящих данных используются новости Dow Jones,

Bloomberg, а в [26] — Yahoo! Finance. Некоторые исследователи сосредоточены исключительно на Twitter и используют его для прогнозирования рынка и анализа общественных настроений [17, 18].

Еще одним классом текстовых источников для систем прогнозирования были ежегодные отчеты компании, пресс-релизы и корпоративная информация [3, 5]. В [5] и [27] входными данными являются не сами новости, а лишь их заголовки. Используемые тексты в виде отчетов могут иметь структурированный или полуструктурированный форматы, например, такие как макроэкономические объявления, поступающие от правительств или центральных банков по уровню безработицы или валовому внутреннему продукту. В [3] использовали такие же структурированные данные для прогнозирования скачков на валютном рынке (FOREX). Также для прогнозирования используются числовые значения в виде цен или рыночных индексов, например в [3] используется Промышленный Индекс Доу-Джонса (Dow Jones Industrial Average).

Доступность и качество экспериментальных наборов данных показывает, что одной из основных проблем в построении инструментальных средств прогнозирования на базе методов анализа текстов является отсутствие стандартизированных наборов данных, характеризующих рыночные тренды в течение определенного периода времени, которые исследователи могли бы использовать для объединения своих исследований.

Большинство исследователей попытались собрать свои собственные наборы данных или как-то сузить поле исследования. Это, естественно, привело к фрагментации методов и результатов. Попытки дать характеристику современным методам анализа текстов для прогнозирования котировок рынка были сделаны в [25, 28].

Авторы [28], проанализировав текущую ситуацию, представили сравнительную таблицу подходов к прогнозированию котировок рынка на основе методов текстового анализа. Сделанный ими вывод: несмотря на множество работ по прогнозированию котировок рынка ценных бумаг на основе текстовых данных, полученных из различных источников, до сих пор не сформулированы единые принципы и подходы, применимые к большинству практических ситуаций.

Авторы [25] рассмотрели работы по прогнозированию рынка на основе методов интеллектуального анализа текста. Выделены и проанализированы три категории: лингвистика (для понимания природы языка), машинное обучение (для создания возможностей вычислительного моделирования и распознавания образов), поведенческая экономика (для выделения экономического смысла). Авторы составили схему системы прогнозирования, подробно описав источники текстовых данных, алгоритм выделения и обработки текста, метод интеллектуального анализа.

На рис. 1 представлен компонент машинного обучения.

Рисунок 1. Схема компонента системы прогнозирования [25].

4. 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 данном обзоре рассмотрены исследования зарубежных ученых. До недавнего времени в нашей стране аналогичные исследования публиковались разрозненно и нерегулярно. Но осенью 2017 г. был опубликован доклад департамента исследований и прогнозирования Центробанка России [29], в котором представлен метод анализа экономической активности по публикациям в интернет-изданиях, основанный на обработке больших массивов данных (Big Data). В докладе отмечается, что «использование такой неструктурированной информации, как новости, является не менее важной составляющей при прогнозировании экономической активности, чем использование обычных статистических показателей». Кроме того, эти данные можно получить быстрее, чем статистику. В докладе отмечается, что разработанная

методика успешно справилась с задачей прогнозирования экономической динамики в системе Центробанка России. Результаты прогнозирования позволили сделать вывод о том, что новостные данные обладают достаточно хорошей прогнозной силой, а значит, возможно более грамотное и оперативное реагирование на текущую экономическую ситуацию и принятие решений ([29]). Попадание в массив новостной информации частично недостоверной информации вносит определенные погрешности, но в целом также отражает настроения, которые преобладают в экономике и в обществе. Именно эти ожидания, оценки предполагаемых изменений в будущем, оказывают заметное воздействие на принятие инвестиционных решений.

Литература

[1] *Khadjeh Nassirtoussi A., Ying Wah T., Ngo Chek Ling D.* A novel FOREX prediction methodology based on fundamental data // *African Journal of Business Management*. 2011. Vol. 5. P. 8322–8330.

[2] *Kaltwasser P. R.* Uncertainty about fundamentals and herding behavior in the FOREX market // *Physica A: Statistical Mechanics and its Applications*. 2010. Vol. 389, No. 8. P. 1215–1222.

[3] *Chatrath A., Miao H., Ramchander S., Villupuram S.* Currency jumps, cojumps and the role of macro news // *Journal of International Money and Finance*. 2014. Vol. 40. P. 42–62.

[4] *Fasanghari M., Montazer G. A.* Design and implementation of fuzzy expert system for Tehran Stock Exchange portfolio recommendation // *Expert Systems with Applications*. 2010. Vol. 37. P. 6138–6147.

[5] *Huang C.-J., Liao J.-J., Yang D.-X., Chang T.-Y., Luo Y.-C.* Realization of a news dissemination agent based on weighted association rules and text mining techniques // *Expert Systems with Applications*. 2010. Vol. 37. P. 6409–6413.

[6] *Yu H., Nartea G. V., Gan C., Yao L. J.* Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets // *International Review of Economics & Finance*. 2013. Vol. 25. P. 356–371.

[7] *Anastasakis L., Mort N.* Exchange rate forecasting using a combined parametric and nonparametric self-organising modelling approach // *Expert Systems with Applications*. 2009. Vol. 36. P. 12001–12011.

[8] *Vanstone B., Finnie G.* Enhancing stockmarket trading performance with ANNs // *Expert Systems with Applications*. 2010. Vol. 37. P. 6602–6610.

[9] *Vanstone B., Finnie G.* An empirical methodology for developing stockmarket trading systems using artificial neural networks // *Expert Systems with Applications*. 2009. Vol. 36. P. 6668–6680.

[10] *Sermpinis G., Laws J., Karathanasopoulos A., Dunis C. L.* Forecasting and trading the EUR/USD ex-change rate with gene expression and psi sigma neural networks // *Expert Systems with Applications*. 2012. Vol. 39, No. 10. P. 8865–8877.

[11] *Huang S.-C., Chuang P.-J., Wu C.-F., Lai H.-J.* Chaos-based support vector regressions for exchange rate forecasting // *Expert Systems with Applications*. 2010. Vol. 37. P. 8590–8598.

[12] *Premanode B., Toumazou C.* Improving prediction of exchange rates using differential EMD // *Expert Systems with Applications*. 2013. Vol. 40. P. 377–384.

[13] *Lupiani-Rui E., García-Manota I., Valencia-Garcí R., García-Sánchez F., Castellanos-Nieves D., Fernández-Breis J. T. et al.* Financial news semantic search engine // *Expert Systems with Applications*. 2011. Vol. 38. P. 15565–15572.

[14] *Robertson C., Geva S., Wolff R.* What types of events provide the strongest evidence that the stock market is affected by company specific news? // Proceedings of the fifth Australasian conference on Data mining and analystics. 2006. Vol. 61. P. 145–153.

[15] *Friesen G., Weller P. A.* Quantifying cognitive biases in analyst earnings forecasts // *Journal of Financial Markets*. 2006. Vol. 9. P. 333–365.

[16] *Bollen J., Huina M., Zeng Xiao-Jun* Twitter mood predicts the stock // *Journal of Computational Science*. 2010. Vol. 2. P. 1–8.

[17] *Kleinnijenhuis J., Schultz F., Oegema D. Atteveldt W.H. van.* Financial News and Market Panics in the age of high frequency trading algorithms // *Journalism*. 2013. Vol. 14, No. 2. P. 271-291

[18] *Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N.* Ontologybased sentiment analysis of twitter posts // *Expert Systems with Applications*. 2013. Vol. 40. P. 4065–4074.

[19] *Balahur A., Steinberger R., Goot E. v. d., Pouliquen B., Kabadjov M.* Opinion Mining on Newspaper Quotations // Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. 2009. Vol. 03. P. 523–526.

[20] *Cambria E., Schuller B., Yunqing X., Havasi C.* New Avenues in Opinion Mining and Sentiment Analysis // *IEEE Intelligent Systems.* 2013. Vol. 28. P. 15–21.

[21] *Yu L.-C., Wu J.-L., Chang P.-C., Chu H.-S.* Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news // *Knowledge-Based Systems*. 2013. Vol. 41. P. 89–97.

[22] *Ortigosa-Hernández J., Rodríguez J. D., Alzate L., Lucania M., Inza I., Lozano J. A.* Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers // *Neurocomputing*. 2012. Vol. 92. P. 98–115.

[23] *Loia V., Senatore S.* A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content // *Knowledge-Based Systems*. 2014. Vol. 58. P. 75–85.

[24] *Majumder D.* Towards an efficient stock market: Empirical evidence from the Indian market // *Journal of Policy Modeling*. 2013. Vol. 35. P. 572–587.

[25] *Khadjeh Nassirtoussi A., Aghabozorgi S., Ying Wah T., Ngo D. C. L.* Text mining for market prediction: A systematic review // *Expert Systems with Applications*. 2014. Vol. 41. P. 7653–7670.

[26] *Schumaker R. P., Zhang Y., Huang C.-N., Chen H.* Evaluating sentiment in financial news articles // *Decision Support Systems*. 2012. Vol. 53. P. 458–464.

[27] *Peramunetilleke D., Wong R. K.* Currency exchange rate forecasting from News Headlines // *Australian Computer Science Communications.* 2002. Vol. 24. P. 131–139.

[28] *Hagenau M., Liebmann M., Neumann D.* Automated news reading: Stock price prediction based on financial news using context-capturing features // *Decision Support Systems*. 2013. Vol. 55. No. 3. P. 685–697.

[29] *Яковлева К.* Оценка экономической активности на основе текстового анализа // Серия докладов об экономических исследованиях Центральный банк Российской Федерации. 2017. № 25. Октябрь.

Авторы:
Елена Гельевна Андрианова — кандидат технических наук, доцент, доцент кафедры корпоративных информационных систем Института информационных технологий, Московский технологический университет (МИРЭА)
Ольга Александровна Новикова — ассистент кафедры информационных технологий в государственном управлении Института инновационных технологий и государственного управления, Московский технологический университет (МИРЭА)

The role of text mining methods to automate securities markets forecasting

E. G. Andrianova, O. A. Novikova

Moscow Technological University (MIREA)
Prospekt Vernadskogo, 78, Moscow, Russia, 119454

e-mail: *andrianova@mirea.ru, novikova@mirea.ru*

Abstract. The article reviews advancements in forecasting dynamics of the securities market using text mining methods. The role of text mining methods in automating research based on traditional forecasting methodologies, such as technical and fundamental analysis, is considered. It was concluded that it is relevant and necessary to expand traditional methods with text data mining and news data, as they improve the efficiency of forecasts of securities market quotations. The need for taking into account data extracted from news and text messages that characterize emotional color of events affecting market changes (behavioral economics, emotional analysis) is presented. The application of methods of analysis and forecasting of the securities market based on models such as machine learning, neural networks, fuzzy logic, support vector machines, genetic programming of network applications, etc., is reviewed.
Key words: method of text mining, forecasting, securities market, fundamental analysis, technical analysis, emotional analysis, behavioral economics, news reports.

References

[1] *Khadjeh Nassirtoussi A.*, *Ying Wah T.*, *Ngo Chek Ling D.* (2011) *African Journal of Business Management*, 5:8322–8330.

[2] *Kaltwasser P. R.* (2010) *Physica A*, **389**(8):1215–1222.

[3] *Chatrath A.*, *Miao H.*, *Ramchander S.*, *Villupuram S.* (2014) *Journal of International Money and Finance*, 40:42–62.

[4] *Fasanghari M.*, *Montazer G. A.* (2010) *Expert Systems with Applications*, **37**:6138–6147.

[5] *Huang C.-J.*, *Liao J.-J. et al* (2010) *Expert Systems with Applications*, **37**:6409–6413.

[6] *Yu H.*, *Nartea G. V.*, *Gan C.*, *Yao L. J.* (2013) *International Review of Economics & Finance*, **25**:356–371.

[7] *Anastasakis L.*, *Mort N.* (2009) *Expert Systems with Applications*, **36**:12001–12011.

[8] *Vanstone B.*, *Finnie G.* (2010) *Expert Systems with Applications*, **37**:6602–6610.

[9] *Vanstone B.*, *Finnie G.* (2009) *Expert Systems with Applications*, **36**:6668–6680.

[10] *Sermpinis G., Laws J., Karathanasopoulos A., Dunis C. L.* (2012) *Expert Systems with Applications*, **39**(10):8865–8877.

[11] *Huang S.-C., Chuang P.-J., Wu C.-F., Lai H.-J.* (2010) *Expert Systems with Applications*, **37**:8590–8598.

[12] *Premanode B., Toumazou C.* (2013) *Expert Systems with Applications*, **40**:377–384.

[13] *Lupiani-Rui E., García-Manota I. et al.* (2011) *Expert Systems with Applications*, **38**:15565–15572.

[14] *Robertson C., Geva S., Wolff R.* (2006) What types of events provide the strongest evidence that the stock market is affected by company specific news? Proceedings of the fifth Australasian conference on Data mining and analystics, vol. 61, p. 145–153.

[15] *Friesen G., Weller P. A.* (2006) *Journal of Financial Markets*, **9**:333–365.

[16] *Bollen J., Huina M., Zeng Xiao-Jun* (2010) *Journal of Computational Science*, **2**:1–8.

[17] *Kleinnijenhuis J., Schultz F. et al.* (2013) *Journalism*, **14**(2):271-291

[18] *Kontopoulos E., Berberidis C., Dergiades T., Bassiliades N.* (2013) *Expert Systems with Applications*, **40**:4065–4074.

[19] *Balahur A., Steinberger R., Goot E. v. d., Pouliquen B., Kabadjov M.* (2009) Opinion Mining on Newspaper Quotations. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, p. 523–526.

[20] *Cambria E., Schuller B., Yunqing X., Havasi C.* (2013) *IEEE Intelligent Systems*, **28**:15–21.

[21] *Yu L.-C., Wu J.-L., Chang P.-C., Chu H.-S.* (2013) *Knowledge-Based Systems*, **41**:89–97.

[22] *Ortigosa-Hernández J., Rodríguez J. D. et al.* (2012) *Neurocomputing*, **92**:98–115.

[23] *Loia V., Senatore S.* (2014) *Knowledge-Based Systems*, **58**:75–85.

[24] *Majumder D.* (2013) *Journal of Policy Modeling*, **35**:572–587.

[25] *Khadjeh Nassirtoussi A. et al.* (2014) *Expert Systems with Applications*, **41**:7653–7670.

[26] *Schumaker R. P., Zhang Y., Huang C.-N., Chen H.* (2012) *Decision Support Systems*, **53**:458–464.

[27] *Peramunetilleke D., Wong R. K.* (2002) *Australian Computer Science Communications*, **24**:131–139.

[28] *Hagenau M., Liebmann M., Neumann D.* (2013) *Decision Support Systems*, **55**(3):685–697.

[29] *Yakovleva K.* (2017) Otsenka ekonomicheskoy aktivnosti na osnove tekstovogo analiza. Seriya dokladov ob ekonomicheskikh issledovaniyakh Tsentral'nyy bank Rossiyskoy Federatsii. No. 25. Oktyabr'. [In Rus]