УДК 336.763

Мубаракшин Самир Рафисович

магистрант направления подготовки «Бизнес-информатика в высокотехнологичных отраслях экономики»
Национальный исследовательский ядерный университет «МИФИ»
Россия, Москва
sammium1995@gmail.com

Samir R. Mubarakshin

Master's student in the direction of training "Business Informatics in High Technology Branches of Economy"
National Research Nuclear University MEPhI
Russia, Moscow

Королев Сергей Андреевич

магистрант направления подготовки «Бизнес-информатика в высокотехнологичных отраслях экономики»
Национальный исследовательский ядерный университет «МИФИ»
Россия, Москва

Sergey A. Korolev

Master's student in the direction of training "Business Informatics in High Technology Branches of Economy"
National Research Nuclear University MEPhI
Russia, Moscow

СОВРЕМЕННЫЕ ВОЗМОЖНОСТИ ПРОГНОЗИРОВАНИЯ НА ФОНДОВЫХ РЫНКАХ НА ОСНОВЕ СЕНТИМЕНТ-АНАЛИЗА

MODERN POSSIBILITIES OF FORECASTING IN STOCK MARKETS ON THE BASIS OF SENTIMENT ANALYSIS AFISOVICH

Аннотация

В статье исследуются современные методы прогнозирования цен акций на фондовом рынке на основе анализа тональности текста. Приводятся методы получения информации, а также достоинства и недостатки основных источников данных: СМИ и социальных сетей. Рассматриваются современные инструменты для получения необходимых данных. В конце приводится выводы об использовании анализа тональности текстов и его преимуществах перед стандартными инструментами фундаментального анализа.

Abstract

This article explores modern methods of forecasting stock prices on the stock market based on the analysis of the tone of the text. The methods of obtaining information, as well as the advantages and disadvantages of the main data sources are presented: Media and social networks. Modern tools for obtaining the necessary data are considered. At the end there are conclusions about the use of text tone analysis and its advantages over standard tools of fundamental analysis.

Ключевые слова:

анализ тональности текста, рынок ценных бумаг, социальные сети

Keywords:

text tone analysis, stock market, social networks

Введение

Прогнозирование фондового рынка очень важно для планирования деловой активности. Предсказание изменения цен акций привлекает внимание многих исследователей в различных дисциплинах, таких как IT, статистика, экономика, финансы. Недавние исследования показали, что огромное количество информации в Интернете, находящейся в открытом доступе, например, использование Википедии, новости из основных СМИ и обсуждения в социальных сетях, могут оказывать заметное влияние на мнение инвесторов о финансовых рынках. Надежность вычислительных

моделей для прогнозирования фондового рынка очень важна, поскольку он очень чувствителен к экономике и может напрямую привести к финансовым потерям.

В последние годы многие финансовые учреждения внедрили приложения машинного обучения для лучшего прогнозирования цен на акции. Вычислительные методы, такие как регрессия или классификация, дают финансовым аналитикам возможность предсказывать будущую цену акции, используя исторические данные. Совсем недавно анализ настроений использовался для сбора общественного мнения в надежде, что он даст значимую информацию, которая позволит создать более точные модели прогнозирования.

Сентимент-анализ, или анализ тональности текста, – одна из областей Data Science, заключающаяся в исследовании текстовых данных для определения их эмоциональной оценки, позитивной или негативной. В настоящее время благодаря Интернету аналитики могут получить большое количество публичных данных для своей работы. Инвесторы тоже могут получить выгоду от использования сентимент-анализа, изучая настроение в социальных сетях и СМИ, чтобы скорректировать свою стратегию и спрогнозировать движение цен на фондовом рынке.

Прогнозирование фондового рынка на основе исследования социальных сетей

В настоящее время социальные сети стали зеркалом, отражающим мысли и мнения людей по поводу любого конкретного события или новости. Позитивные или негативные настроения общественности, связанные с конкретной компанией, могут оказать воздействие на цены ее акций. Аналитики стремятся предсказать рыночные цены акций различных компаний путем проведения анализа настроений в социальных сетях, таких как Twitter, связанные с соответствующими компаниями. Поскольку реакция общественности на любое крупное событие доступна практически мгновенно в любой социальной сети, можно быстро уловить ее настроение и определить оценку волатильности цен на акции, обеспечив тем самым прогноз почти в реальном времени, подобный некоторым моделям прогнозирования погоды [2].

Для сбора постов Twitter предоставляет надежный API. Получаемый по запросу JSON-объект содержит посты и их метаданные. Он включает разнообразную информацию: имя пользователя, время, местоположение, ретвиты. Аналитики сосредотачиваются на времени и тексте твита для дальнейшего изучения. Основной

задачей при обработке данных из социальных сетей становится очищение постов из Twitter и фильтрация тех, которые имеют отношение к исследуемым компаниям.

Текст каждого сообщения содержит слишком много посторонних слов, которые не учитывают его настроение. Посты включают URL, теги для других и многие другие символы, которые не имеют никакого значения для настроения. Для точного определения настроения твита необходимо отфильтровать всё, что создает лишний «шум» для применения аналитических моделей в будущем. В основном, применяется следующий алгоритм:

1. Полученное сообщение разделяется по пробелам. Таким образом, из единой строки символов аналитик получает список отдельных слов.

2. Лемматизация: слова нормализуются, то есть приводятся в изначальную форму (имена существительные приводятся в именительном падеже, глаголы – в инфинитиве и т. д.).

3. Лишние слова, не несущие смысловой нагрузки, удаляются (например, предлоги или союзы, в английском языке убираются артикли).

В итоге получается список из некоторого количества слов, где каждое слово может быть в дальнейшем использовано как признак позитивности или негативности сообщения. Однако, при большой выборке данных, таких слов станет слишком много. В таких случаях отбираются наиболее значимые слова из выборки [1].

Таким образом, каждому полученному из социальной сети сообщению дается признак позитивности или негативности. В самом простом варианте каждое позитивное слово дает +1 балл, негативное слово дает -1 балл, получившийся знак суммы всех баллов определяет тональность этого поста. Затем проводится корелляционный анализ цен акций на фондовом рынке и настроения общественности в данный период времени. На основании полученных данных аналитики (или же, например, модель нейронной сети) могут делать прогноз движения цен акций компании на рынке.

Преимуществами данного метода прогнозирования цен на фондовом рынке будут:

1. Скорость реакции на события в мире. Каждое событие влечет за собой обсуждение его в обществе, а социальные сети сделали это доступнее и быстрее для всех, в том числе и для аналитиков.

2. Развитие технологий. Все действия из алгоритма выше на данный момент проводятся автоматизировано. Нейросетевые технологии на данный момент хорошо развиты и позволяют прогнозировать движение цен на основе огромных объемов данных.

3. Большая выборка. Согласно статистике от Твиттера в 2020 году, за день в соцсети появляется до полумиллиарда постов.

Однако есть и свои недостатки:

1. Мусорные сообщения. Некоторые сообщения могут не нести никакой смысловой нагрузки и состоять из случайных слов, эмодзи, специальных символов и т.п. Это может быть использовано в дальнейшем для создания большого количества таких псевдопозитивных или псевдонегативных сообщений в сети. С подобным социальные сети могут бороться так же, как с накруткой лайков или голосов в опросах, может быть введена автоматическая премодерация при обнаружении каких-то слов в сообщении.

2. Огромные объемы данных. Каждый пост будет храниться в виде списка слов, количество постов в зависимости от сложности задачи может доходить до миллиардов. Для хранения и обработки таких объемов данных потребуются дорогие вычислительные центры. Возможно использование облачных сервисов в дальнейшем [4].

3. Репрезентативность выборки. Не все позитивные или негативные действительно являются таковыми, люди могут использовать сарказм, иронию, что машинный алгоритм отличить не сумеет.

4. Языковая особенность социальной сети. Пользователи часто могут использовать жаргонизмы, сленговые выражения, выражать свои эмоции с помощью эмодзи или матерных слов или словосочетаний, или же могут использовать несколько языков в одном сообщении. Также стоит заметить, что Твиттер, например, имеет лимит по количеству символов в одном посте, что затрудняет написание сложносочиненных предложений и подталкивает пользователей к общению короткими фразами [7].

Прогнозирование фондового рынка на основе исследования новостных лент

Несмотря на большую роль социальных сетей в нашей жизни, люди также привыкли узнавать новости (особенно если они по определенной компании или отрасли) из специализированных СМИ. Доверие к ним выше, информации дается

больше. Тональность новостей может также стать психологическим фактором и повлиять на мнение читателя о теме [6].

Анализ тональности текста новостей технически похож на анализ постов в социальных сетях. Алгоритм действий такой же: берется новость, из нее выделяются наиболее важные слова, на основе этих слов дается оценка текста [5].

Однако есть некоторые особенности. В новостных лентах обычно используются одни и те же термины, не допускается излишней эмоциональной окраски при описании события. В новостной ленте не может появиться мусорных сообщений, как в ленте социальной сети. Таким образом, такие данные можно считать более достоверным источником информации. Большая надежность является ключевым фактором для аналитиков, так как высоки риски финансовых потерь.

Отличием новостных лент также будет их разрозненность. Если говорить о социальных сетях, то большинство пользователей будет использовать Facebook (в странах СНГ более популярен VK), Twitter, Instagram. Однако новостных интернет-изданий гораздо больше, не везде реализованы интерфейсы для получения данных для программной обработки, таким образом, аналитику придется разрабатывать собственное решение по сбору информации [3].

Таким образом, можно выделить следующие достоинства данного метода:

1. Большая надежность, нежели с при исследовании социальных сетей. В лентах серьезных новостных изданий не будет новостей, имеющих ироничный подтекст, не будет спама одной и той же новостью, чтобы повлиять на оценку.

2. Отсутствие лишнего шума: мусорных сообщений, излишней эмоциональной окраски, эмодзи, обсценной лексики. Количество сленговых и жаргонных выражений также будет сведено к минимуму.

3. Меньший объем данных, что удобно для аналитика. При сочетании с большей надежностью дает большую возможность для аналитика.

Данный метод будет проигрывать в следующих вещах:

1. Скорость реакции. Социальные сети отреагируют на событие быстрее, чем СМИ.

2. Ангажированность СМИ. Политика государства, редакции или компании-владельца СМИ могут серьезно повлиять на новость о событии. СМИ могут не сообщить о чем-то или вывернуть оценку в нужную им сторону. С подобным тяжело бороться

технически, обычно исследователи пытаются увеличить количество источников данных.

Современные инструменты для получения информации

Для аналитиков на данный момент существует много различных сервисов для сбора данных из СМИ и соцсетей. Например, в разделе «Инвестиции» на вебсайте РБК у каждой новости есть ссылка на цену акций компаний, прямо или косвенно упоминаемых в статье, на данный момент. Это позволяет трейдеру оценить влияние события на итоговую стоимость.

Однако, как уже отмечалось в сравнении выше, скорость новостной ленты гораздо меньше скорости ленты в Twitter, где аналитик может подписаться на интересующие его аккаунты. Это могут быть компании, с чьими акциями он работает, или же компании, составляющие рейтинги и индексы. Также это могут быть главные лица в такой компании. Ярким примером аккаунта в Twitter, влияющим на акции своей компании, является Илон Маск, генеральный директор Tesla и SpaceX [8]. Некоторые трейдеры предпочитают подписываться на других известных трейдеров и следовать их советам. Существуют различные приложения (например, Stocktwits), агрегирующие и систематизирующие посты в Twitter от известных авторов, в таких приложениях часто встроен анализ настроения данного поста.

Заключение

С учетом развития облачных и нейросетевых технологий возможности обработки большого потока текстовых данных растут с каждым годом, благодаря чему появляются новые пути для исследований настроений общества. Оценки, полученные на основе сентимент-анализа новостных статей и постов в социальных сетях, могут быть мощными индикаторами движения цен акций на фондовом рынке. Сентимент-анализ позволяет учесть эмоциональный фактор, который невозможно получить из финансовых отчетов компании.

Однако, данные методы нежелательно использовать в отрыве от классических методов фундаментального и технического анализа. Общей проблемой исследования настроений в СМИ и социальных сетях будет фактор самого исследователя: какие слова он будет считать главными признаками, насколько репрезентативна будет выборка постов и новостей.

Список использованных источников

1. Nuno Oliveira, Paulo Cortez and Nelson Areal, "The impact of microblogging data for stock market prediction: Using Twitter to predict returns volatility trading volume and survey sentiment indices", Expert Systems with Applications, vol. 73, pp. 125-144, 2017.

2. T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar and C. S. Lifna, "Stock Market Prediction based on Social Sentiments using Machine Learning," 2018 International Conference on Smart City and Emerging Technology (ICSCET), 2018, pp. 1-3, doi: 10.1109/ICSCET.2018.8537242.

3. D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

4. E. Seals and S. R. Price, "Preliminary Investigation in the use of Sentiment Analysis in Prediction of Stock Forecasting using Machine Learning," 2020 SoutheastCon, 2020, pp. 1-2, doi: 10.1109/SoutheastCon44009.2020.9368258.

5. Беляков Михаил Васильевич. Анализ новостных сообщений сайта МИД РФ методом сентимент-анализа (статья 2) // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. 2016. №4. URL: https://cyberleninka.ru/article/n/analiz-novostnyh-soobscheniy-sayta-mid-rf-metodom-sentiment-analiza-statya-2

6. Афанасьев Дмитрий Олегович, Федорова Елена Анатольевна, Рогов Олег Юрьевич. О влиянии тональности новостей в международных СМИ на рыночный курс российского рубля: текстовый анализ // Экономический журнал ВШЭ. 2019. №2. URL: https://cyberleninka.ru/article/n/o-vliyanii-tonalnosti-novostey-v-mezhdunarodnyh-smi-na-rynochnyy-kurs-rossiyskogo-rublya-tekstovyy-analiz

7. Бодрунова С.С. Кросс-культурный тональный анализ пользовательских текстов в Твиттере. URL: https://vestnik.journ.msu.ru/books/2018/6/kross-kulturnyy-tonalnyy-analiz-polzovatelskikh-tekstov-v-tvittere/

8. Что творится с рынком после взрывных твитов Илона Маска: 5 ярких примеров URL: https://quote.rbc.ru/news/article/5eb2c9719a7947889ba1dfa3