

АВТОМАТИЗИРОВАННЫЙ АЛГОРИТМ КВАНТИФИКАЦИИ ИНФОРМАЦИОННОЙ СРЕДЫ ФИНАНСОВОГО РЫНКА

© 2022 **Родионов Дмитрий Григорьевич**

доктор экономических наук, профессор Высшей инженерно-экономической школы
Санкт-Петербургский политехнический университет Петра Великого (СПбПУ),
Россия, Санкт-Петербург
E-mail: dmitry.rodionov@spbstu.ru

© 2022 **Пашина Полина Александровна**

студент Высшей инженерно-экономической школы
Санкт-Петербургский политехнический университет Петра Великого (СПбПУ),
Россия, Санкт-Петербург
E-mail: pashinapolina@yandex.ru

© 2022 **Конников Евгений Александрович**

кандидат экономических наук, доцент
Высшей инженерно-экономической школы
Санкт-Петербургский политехнический университет Петра Великого (СПбПУ),
Россия, Санкт-Петербург
E-mail: konnikov.evgeniy@gmail.com

Изучение информационной среды финансового рынка предполагает агрегацию новостного массива, сформированного вокруг компаний, представленных на финансовом рынке, их ценных бумаг и иных новостей об их деятельности и изменениях в ней. Ранее в статье «Квантификаторы информационной среды финансового рынка» была разработана система квантификаторов информационной среды финансового рынка, предполагающая расчёт показателей, характеризующих некоторые свойства или проявления этой среды. В рамках данной статьи будет рассмотрен автоматизированный алгоритм квантификации информационной среды финансового рынка, для чего автором статьи предварительно были отобраны компании, цены акций и волатильность цены которых будет исследоваться в данной работе, а также определён период, в рамках которого предполагается рассмотрение выбранных параметров финансового рынка.

Ключевые слова: информационная среда, квантификаторы, квантификация, финансовый рынок, токены, токенизация, тональность, автоматизированный алгоритм, парсинг.

По мнению автора статьи, для выявления базовых выводов в квантификации информационной среды финансового рынка необходимо исследовать крупные компании, представленные на мировом рынке, упоминания о которых часто встречаются в новостях, т.е. наиболее ярко представлены в информационной среде. Во-первых, это компании, акции которых «на слуху» для большинства людей и обладают особой привлекательностью для неквалифицированных инвесторов-новичков. Во-вторых, это компании, регулярно вносящие изменения в свою деятельность, в свою продукцию и анонсирующие это в сети Интернет. В-третьих, это компании — онлайн-маркетплейсы. Так, для анализа влияния предложенных показателей информационной среды на основные параметры финан-

сового рынка в качестве примера были выбраны крупнейшие международные компании, присутствие в информационной среде мирового финансового рынка оценивается как высокое, — Tesla Inc., Apple Inc. и Amazon.com Inc. Причина выбора для рассмотрения вышеуказанных компаний заключается в значительной представленности данных компаний в информационной среде, что обеспечит отсутствие большого количества нулевых значений выбранных автором исследования показателей информационной среды вследствие их значительного присутствия в новостных потоках. Кроме того, важно рассмотреть, как информационная среда влияет на динамично развивающуюся компанию Tesla Inc., мнения инвесторов о которой довольно противоречиво, на устойчивую компанию Apple Inc., явля-

ящуюся IT-гигантом с огромной капитализацией, а также вызывающую существенный интерес для инвестирования у «новичков» на финансовом рынке, на крупнейший маркетплейс и мировой гигант электронной торговли — компанию Amazon.com Inc., которая обладает особой привлекательностью для инвесторов, имела феноменальный рост котировок акций в 2020 году и перспективы дальнейшего развития которой обсуждаются сотнями экспертов на финансовом рынке.

По данным компаниям с американской биржи Nasdaq были собраны цены акций (в долларах) и рассчитана их волатильность за период 01.04.2020–31.08.2020 гг. по формуле (1) [27].

$$V = \frac{P_{max} - P_{min}}{P_{min}} * 100\% \quad (1)$$

где P_{max} — максимальная цены акции за 1 день, долл.

P_{min} — минимальная цены акции за 1 день, долл.

Причина выбора периода 01.04.2020–31.08.2020 гг. обуславливается особым влиянием информационных потоков в финансовой среде, приходящейся на пандемийный период, когда фондовый рынок переполнился частными неквалифицированными инвесторами, инфор-

мационная среда характеризовалась созданием огромного количества чатов инвесторов и люди решили найти новые способы заработка. Так, в статье «Конец эпохи низких ставок: вернутся ли инвесторы в депозиты?» одними из основных тезисов являются следующие факты:

1. «В развитых странах физические лица активно участвуют в деятельности на фондовом рынке. Так, в США ценными бумагами владеет более половины населения, в Японии — немного менее половины, в Великобритании — около трети населения. В развивающихся странах доля участия розничных инвесторов невысока, но наблюдается стремительный рост числа физлиц на фондовом рынке».

2. «Скорее всего, к прежнему низкому уровню присутствия населения на фондовом рынке мы уже не вернемся: тренд на увеличение числа индивидуальных инвесторов продолжится, как и в других развивающихся экономиках» [28].

Полноценный анализ состояния информационной среды, его позитивных, негативных и содержательных компонент позволяет сформировать универсальный автоматизированный алгоритм его калькуляции. Предложенный автором алгоритм включает в себя 2 этапа.

Предварительно импортируем необходимые инструментальные библиотеки для всего автоматизированного алгоритма (рис. 1).

- 1) Установка необходимых инструментальных библиотек
 - Transformers — библиотека для обучения state-of-the-art моделей в обработке естественного языка.
 - Numpy — библиотека, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых математических функций для операций с этими массивами.
 - pandas — библиотека для обработки и анализа структурированных данных.
 - Pygooglenews — библиотека инструментов парсинга новостного агрегатора Google.ru
 - datetime — библиотека определения и генерации временной информации.
 - tqdm.notebook — библиотека, обеспечивающая простой индикатор выполнения для операций.
 - Bs4 — библиотека для извлечения данных из файлов HTML и XML.
 - Selenium webdriver — программная библиотека для управления браузером.
 - Time — библиотека для выставления задержки между действиями.
 - Collections — это встроенный модуль, реализующий специализированный контейнер типов данных.
 - Nltk — наиболее известная и эффективная библиотека инструментов для обработки естественного языка:
 - Nltk.tokenize,
 - Nltk.stem,
 - Nltk.corpus.
 - Re — модуль для работы с регулярными выражениями.
 - Fuzzywuzzy — библиотека, предназначенная для нечётного сравнения строк.

```
from transformers import BertTokenizer, BertForSequenceClassification
import numpy as np
from selenium import webdriver
import pandas as pd
from pygooglenews import GoogleNews
from datetime import datetime, timedelta
from tqdm.notebook import tqdm
from bs4 import BeautifulSoup
from time import sleep
from collections import Counter
from nltk.tokenize import word_tokenize
from nltk import pos_tag
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
import re
from fuzzywuzzy import fuzz
```

Рис. 1. Импорт необходимых инструментальных библиотек

Далее рассмотрим последовательно каждый из этапов:

1. *Формирование взвешенного массива токенов по тематике «финансовые рынки».* На данном этапе автором исследования был определен первоначальный список токенов, так или иначе характеризующих тематику финансового рынка. Присутствие этих слов в новостной единице позволило бы отнести данный информационный поток (новостную единицу) к новостям, описывающих какие-либо события на финансовых рынках. Для этого автор исследования рассматривал основные финансовые инструменты, присутствие в новостной единице которых позволило бы рассчитать коэффициент присутствия тематики финансового рынка. Это были следующие токены: «акция», «облигация», «валюта», «фьючерсы». Так как данного списка слов недостаточно для поиска новостей с тематикой финансового рынка, автором было принято решения в использовании ассоциативного алгоритма, который обеспечил расширение списка токенов путём поиска ассоциаций к каждому первоначально определённому токеноу. Для целей автоматизированного поиска ассоциаций использован язык программирования Python 3 и часть фреймворка Selenium — Selenium WebDriver. По назначению Selenium WebDriver представляет собой драйвер браузера, то есть программную библиотеку, которая позволяет разрабатывать программы, управляющие поведением браузера [3]. Поиск ассоциаций производился с помощью сайта wordassociations.net/ru. В результате был рассчитан удельный вес каждого токена в общей совокупности ассоциативных токенов. Всего было собрано 31 токенов, из которых 2 (exchange и mint) повторялись 2 раза, 4 (stocks, bonds, currency, futures) — 3 раза, поэтому их доля в общей совокупности ассоциативных токенов больше (всего $31 + 2 \cdot 1 + 4 \cdot 2 = 41$). Так, автором данного исследования было решено осуществить сбор за каждый день. Для целей повышения специализации данного исследования, в качестве базового языка парсинга был выбран английский, так как анализируются мировые события и компании.

На рисунке 2 (с учётом рисунка 1) предлагается рассмотрение подробного алгоритма автоматизированной реализации данного этапа с применением языка программирования Python 3. Результатом реализации данного этапа является формирование взвешенного массива токенов по тематике «финансовые рынки» (таблица 1).

2. *Извлечение информации относительно новостного фона компаний.* На данном этапе в пер-

вую очередь был определён и зафиксирован список компаний, новостные потоки вокруг которых важны для исследования. Кроме того, был определён временной период, за который будет проводиться сбор новостного массива. Оптимальным источником первичной информации в данном случае выступает платформа агрегации новостей от Google. Данный выбор обусловлен следующей особенностью: он представляет собой непрерывный поток ссылок на статьи, организованные из тысяч издателей и журналов. Такая особенность в рамках данного исследования позволяет оценить все опубликованные мировые новости, т.е. благодаря Googlenews имеется возможность получить информацию о всех событиях, происходящих в мире и опубликованных в сети Интернет на информационных ресурсах. Для целей автоматизированного парсинга данной информации может быть использован язык программирования Python 3 и библиотека GoogleNews. Инструменты данной библиотеки позволяют определить наиболее широко распространённые новостные заголовки в рамках заданного периода времени и на заданном языке (английский). Собранный новостной массив обрабатывался: происходили удаление «стоп-слов», не несущих в себе важной содержательной информации (например, предлоги) и лемматизация (процесс приведения словоформы к лемме — её нормальной (словарной) форме) с целью сокращения потерь важной информации. Далее новостной массив токенизировался и для каждой новостной единицы рассчитывалась тональность с помощью библиотеки Transformers (BERT). Bidirectional Encoder Representations from Transformers (BERT) — это метод машинного обучения на основе трансформера для предварительной подготовки к обработке естественного языка, разработанный Google [4]. В рамках данной работы будет использоваться именно эта библиотека, так она позволяет анализировать тональность финансовых текстов. Определение тональности — это метод анализа речи в *компьютерной лингвистике*. Текст токенизируется на смысловые единицы, далее позитивные и негативные компоненты рассчитываются как сумма слов с позитивным и негативным подтекстом делённое на число упоминаний каждой компоненты. Также на данном этапе происходили оценка коэффициента тематики финансового рынка и расчёт выбранных автором квантификаторов информационной среды, ранее рассматриваемых в статье «Квантификаторы информационной среды финансового рынка».

Кроме ранее указанной библиотеки на данном этапе предусмотрено и применение инструментов библиотеки Pandas, которую можно использовать для чтения и записи файлов *Excel*, что позволит

создать необходимый датафрейм на базе MS Excel.

На рисунке 3 представлен автоматизированный алгоритм извлечения информации относительно новостного фона компаний.



Рис. 2. Автоматизированный алгоритм формирования взвешенного массива токенов по тематике «финансовые рынки»

Таблица 1. Массив токенов по тематике финансового рынка

Токен	Удельный вес
stock	0,024390244
debt	0,024390244
trade	0,024390244
cash	0,024390244
acid	0,024390244
commodity	0,024390244
trading	0,024390244
investor	0,024390244
treasury	0,024390244
swap	0,024390244
transaction	0,024390244

solvent	0,024390244
buy	0,024390244
carbon	0,024390244
reserve	0,024390244
penny	0,024390244
stabilize	0,024390244
valuation	0,024390244
asset	0,024390244
bond	0,024390244
breeding	0,024390244
worthless	0,024390244
complementary	0,024390244
price	0,024390244
privatization	0,024390244
purchasing	0,024390244
exchange	0,036585366
mint	0,036585366
bonds	0,073170732
currency	0,073170732
stocks	0,073170732
futures	0,073170732

Настройка инструментов извлечения.

- gn – свойства извлекаемой новостной информации

```
full_data = {}
gn = GoogleNews(lang = 'en')
finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone', num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')
```

Определение объектов (компаний)

```
company_list = ['Tesla', 'Apple', 'Amazon']
```

Определение временного диапазона поиска новостных единиц

```
start = '2020-04-01'
finish = '2020-08-31'

date_list = [datetime.strftime(datetime.fromisoformat(start) + timedelta(days = x), format = '%Y-%m-%d') for x in range(0, (datetime.fromisoformat(finish) - datetime.fromisoformat(start)).days + 2)]
```

Извлечение новостной информации

```
stop_words = stopwords.words('english')
lemmatizer = WordNetLemmatizer()

for company in tqdm(company_list):
    company_data = {}
    for d in tqdm(date_list[:-1]):
        k_list = []
        k_list_2 = []
        info = gn.search(company, from_ = d, to_ = date_list[date_list.index(d) + 1])
        news = [x['title'] for x in info['entries']]
```

Токенизация единиц новостной информации

```
if len(news) > 0:
    inputs = tokenizer(news, return_tensors="pt", padding=True)
    outputs = finbert(*inputs)[0]
    df = pd.DataFrame(np.array(outputs.detach()), columns = ['neutral', 'positive', 'negative'])
    text_tokens_list = []
    for text in tqdm(news):
        text_1 = re.sub('[^\w\s]', '', text.lower())
        text_2 = re.sub('un', '', text_1)
        if text_2.strip() != "":
            text_3 = word_tokenize(text_2.strip())
            text_4 = []
            for token in text_3:
                lemmetized_word = lemmatizer.lemmatize(token)
                text_4.append(lemmetized_word)
            text_5 = []
            for token in text_4:
                if token not in stop_words:
                    text_5.append(token)
            if len(text_5) > 0:
                text_tokens_list.extend(text_5)
```

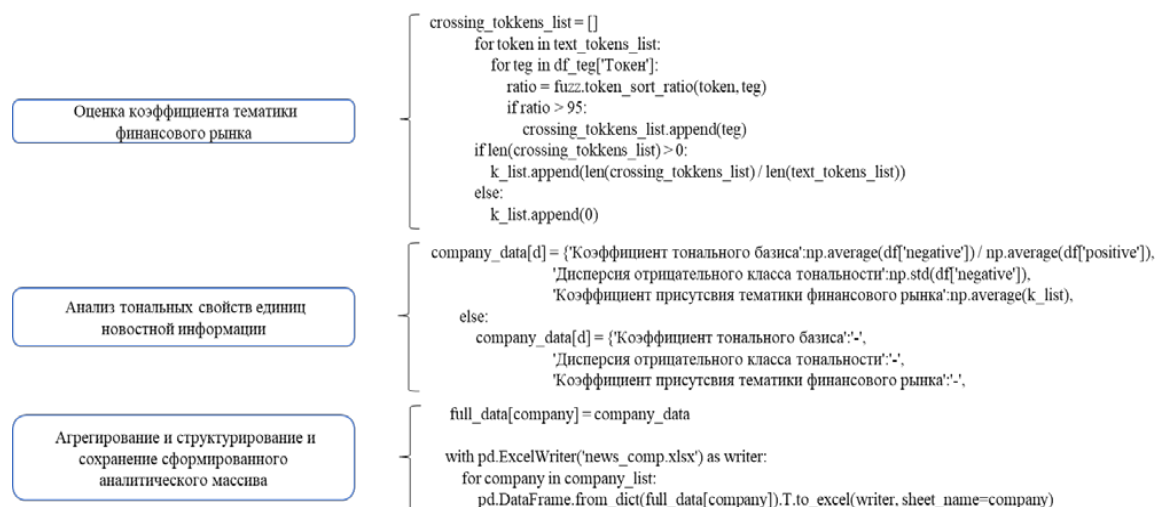


Рис. 3. Извлечение информации относительно новостного фона компаний

Результатом применения данного алгоритма является собранный датафрейм по трём компаниям за заданный промежуток времени и представляет собой таблицы с 4 столбцами: дата,

коэффициент тонального базиса, дисперсия отрицательного класса тональности, коэффициент присутствия тематичности финансового рынка.

Библиографический список

1. Nasdaq. [Электронный ресурс]. — Режим доступа: <https://www.nasdaq.com/> (дата обращения: 15.03.2022).
2. Конец эпохи низких ставок: вернутся ли инвесторы в депозиты?. [Электронный ресурс]. — Режим доступа: https://raexpert.ru/researches/individual_investors_2021/ (дата обращения: 15.03.2022).
3. Что такое Selenium WebDriver?. — Текст: электронный // habr.com: [сайт]. — URL: <https://habr.com/ru/post/152971/> (дата обращения: 16.03.2022).
4. BERT (языковая модель). — Текст: электронный // neerc.ifmo.ru: [сайт]. — URL: [https://neerc.ifmo.ru/wiki/index.php?title=BERT_\(языковая_модель\)](https://neerc.ifmo.ru/wiki/index.php?title=BERT_(языковая_модель)) (дата обращения: 16.03.2022).
5. Родионов, Д. Г. Квантификация информационной среды как инструмент инвестиционного анализа / Д. Г. Родионов, П. А. Карпенко, П. А. Пашина // Экономические науки. — 2021. — № 204. — С. 144–153. — DOI 10.14451/1.204.144. — EDN FOZMSH.
6. Родионов, Д. Г. Методология системного анализа информационной среды / Д. Г. Родионов, Е. А. Конников, О. А. Конникова // Экономические науки. — 2021. — № 196. — С. 160–174. — DOI 10.14451/1.196.160. — EDN EAZTRR.