Отчёт

**По теме: “Анализ рынка акций в зависимости от новостного контента”**

Выполнили: Березовский Александр Дмитриевич, Холуев Владислав Евгеньевич, Муравьев Антон Сергеевич.

Оглавление

[1. Введение. 3](#_Toc90814225)

[1.1. Цель работы. 3](#_Toc90814226)

[1.2. Задание. 3](#_Toc90814227)

[2. Разработка ПО. 4](#_Toc90814228)

[2.1. Получение dataset новостных постов. 4](#_Toc90814229)

[2.2. Обработка dataset и создание зависимой информации. 4](#_Toc90814230)

[2.3. Анализ и подготовка к обучению для прогнозирования результатов. 4](#_Toc90814231)

[2.4. Обучение логистической регрессии. 4](#_Toc90814232)

[3. Анализ и тестирование программы. 5](#_Toc90814233)

[3.1. Анализ. 5](#_Toc90814234)

[3.2. Тестирование. 5](#_Toc90814235)

[4. Вывод 6](#_Toc90814236)

[5. Код программы. 7](#_Toc90814237)

# Введение.

Прогнозирование инвестиций – это сложный, многоэтапный процесс, связанный с изучением вероятностных аспектов вложения денежных средств в те или иные сферы экономики.

Инвестиционные прогнозы важны всегда. Однако их значение многократно возрастает в условиях экономического кризиса и международной политической нестабильности. В подобной ситуации умение инвестора правильно спрогнозировать основные тенденции развития финансовых рынков ценится особенно высоко.

Такие прогностические способности остаются залогом успешного осуществления инвестиционной деятельности. В связи с этим каждый современный инвестор должен уметь на основании аналитических методов разбирать текущее состояние рынков и делать точные прогнозы на среднесрочную и долгосрочную перспективу.

## Цель работы.

Целью данной работы является демонстрация автоматизации информационных процессов, в функционал которого входит связь поступающих данных, а также их аналих, с целью выявления прогноза (зависимости).

## Задание.

Найти новостной сайт (ы) относительно определённых корпораций. Пропарсить его, связать время постинга совокупности информации и время падения (подъема) акций. Узнать статус тех или иных постов, их влияние на изменение акций.

# Разработка ПО.

## Получение dataset новостных постов.

Программа обращается к новостным и биржемым порталам для получения необходимой информации. К сожалению, многие порталы закрывают доступ, а посты недоступны (примерно 200 из 5000, что составляет примерно 4%). Для получения новостных каналов использовался Telegram.

**Telegram** – это мессенджер, позволяющий пользователям обмениваться мгновенными сообщениями и файлами различных форматов, совершать звонки, создавать ботов, вести собственные каналы и тд.

Для того, чтобы считать данные с биржи, используется Tinkoff Invest API.

**Tinkoff Invest API** – это api, необходимое чтобы вести собственную статистику прибыльности.

## Обработка dataset и создание зависимой информации.

При обработки dataset берутся некоторые интересующие нас данные (время, рост/падение акций, тест поста и тд), удаляется ненужная информация и формируется конечный набор данных.

Идёт обработка напрямую связанная с тикерами.  
**Тикер** — это сокращенное наименование определенного актива на бирже, состоящего из латинских букв и цифр, применяемого на конкретной бирже. Обычно его длина составляет от одного до шести символов. Тикером обозначают акции, облигации, фьючерсы, фонды и евробонды.

Каждый тикер обрабатывается отдельно от остальных и связан с определенной компанией, что даёт возможность легко анализировать тот или иной пост.

## Анализ и подготовка к обучению для прогнозирования результатов.

Производится некоторый анализ отфильтрованной информации, берутся частоиспользуемые слова, результаты падений/роста акций, количество успешных постов и записываются в некоторую БД для дальнейшего использования.

Для анализа берётся дата публикации определённого поста его тикеры и результат роста/падения акций с биржи.

## Обучение логистической регрессии.

Обучение происходит при помощи полученной конечной информации из БД. При получение нового информационного поста происходит сравнение с характеристиками из прошлых постингов и выявление прогнозирования для той или иноф компании.

**Логистическая регрессия** или логит-модель — статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой. Эта регреcсия выдаёт ответ в виде вероятности бинарного события (1 или 0).

Основная идея логистической регрессии заключается в том, что пространство исходных значений может быть разделено линейной границей (т.е. прямой) на две соответствующих классам области. В случае двух измерений — это просто прямая линия без изгибов. В случае трех — плоскость, и так далее. Эта граница задается в зависимости от имеющихся исходных данных и обучающего алгоритма. Чтобы все работало, точки исходных данных должны разделяться линейной границей на две вышеупомянутых области. Если точки исходных данных удовлетворяют этому требованию, то их можно назвать линейно разделяемыми.

# Анализ и тестирование программы.

Проект состоит из 3 основных частей

1. Сбор датасета из телеграм каналов

Данные берутся из постов телеграм каналов, которые публикуют новости о компаниях, которые торгуются на фондовом рынке.

Телеграмм каналы:

<https://t.me/AK47pfl>,

<https://t.me/proinvesty>,

<https://t.me/cozytrade>,

<https://t.me/financemap>

Из телеграм каналов были собраны и обработаны все посты в текстовом виде.

Для обработки собранных данных был написан следующий пайплайн:

def format\_one\_tiker(self):

# pipeline

self.delete\_tikers()

self.delete\_start\_dot()

self.delete\_double\_stars()

self.delete\_spec\_symbols()

self.delete\_links()

self.delete\_round\_brackets()

self.delete\_footer()

self.delete\_new\_lines()

self.delete\_colon()

self.delete\_square\_brackets()

return self.text

def format\_many\_tikers(self):

# pipeline

self.delete\_start\_dot()

self.delete\_double\_stars()

self.delete\_spec\_symbols()

self.delete\_links()

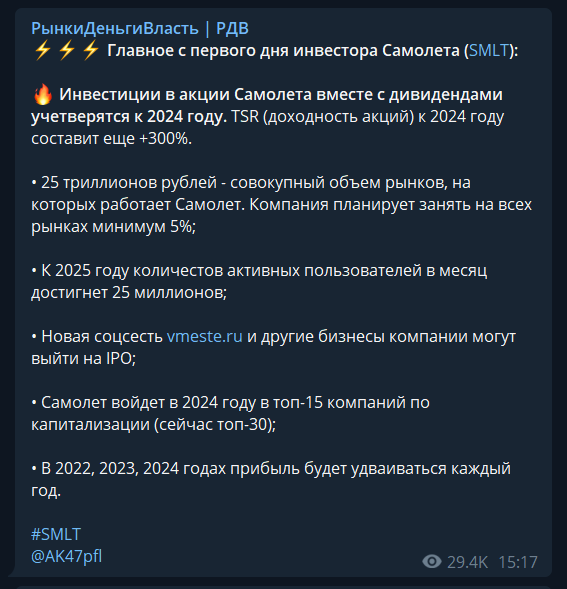
self.delete\_round\_brackets()

self.delete\_footer()

self.delete\_colon()

return self.text

**Пример поста до обработки**:



В каждом посте ищутся тикеры компаний. После обработки поста удаляется мусорная информация и далее, записывается в файл csv.

**После обработки:**

2021-11-29 12:17:31+00:00," Главное с первого дня инвестора Самолета Инвестиции в акции Самолета вместе с дивидендами учетверятся к 2024 году. TSR доходность акций к 2024 году составит еще +300%. 25 триллионов рублей - совокупный объем рынков, на которых работает Самолет. Компания планирует занять на всех рынках минимум 5%; К 2025 году количестов активных пользователей в месяц достигнет 25 миллионов; Новая соцсесть vmeste.ru и другие бизнесы компании могут выйти на IPO; Самолет войдет в 2024 году в топ-15 компаний по капитализации сейчас топ-30; В 2022, 2023, 2024 годах прибыль будет удваиваться каждый год. ",SMLT

1. Сбор биржевых данных

Для сбора данных фондового рынка была использована функциональность приложения Тинькофф инвестиции через Тинькофф API

Для работы с тинькофф API потребовался брокерский счет. Брокерский счет позволил создать подключение к фондовому рынку и собрать данные о компаниях, которые были опубликованы в новостях телеграм каналов.

Код для получения данных о компании, о которой был опубликован пост.

Сначала по тикеру получаем figi (Financial Instrument Global Identifier (Финансовый Глобальный Идентификатор инструмента)), далее по figi получаем данные за 4 недели о компании, по которой был опубликован пост.

В датасет записываются данные за первые 4 недели после публикации поста.

def get\_company\_info(client, tiker, date):

    '''

    В данной функции 2 запроса к API

    'c' - цена закрытия

    'o' - цена открытия

    'h' - максимальная цена торгов

    'l' - минимальная цена торгов

    '''

    # -! ЗАПРОС

    res = client.market.market\_search\_by\_ticker\_get(tiker)

    print(res)

    company\_info = res.to\_dict()

    figi = company\_info['payload']['instruments'][0]['figi']

    print(figi)

    start\_date = str(date).replace(' ', 'T')

    print(start\_date)

    end\_date = str(date + timedelta(days=31)).replace(' ', 'T')

    print(end\_date)

    # -! ЗАПРОС

    market\_info = client.market.market\_candles\_get(

        figi, start\_date, end\_date, 'week').to\_dict()

    if market\_info['status'] == 'Ok':

        return market\_info['payload']['candles']

    else:

        print('Error')

        exit(1)

При сборе данных возникла проблема, связанная с ограничением количества запросов через приложение тинькофф инвестиции.

Лимит запросов в минуту составляет – 100 запросов в минуту

Результатом программы является датасет, содержащий пост, тикер компании, а также показатели на 4 недели вперед, относительно даты и времени публикации поста.

1. Анализ

В данной части ведется основная работа по работе с информацией.

Создается итоговый датасет, который является конечным для работы с моделью.

Для этого берется прошлый датасет, далее из него берутся данные по 4 неделям, и с помощью следующей функции, производится создание класса для классификации, 1- если цена акции выросла, 0 – не выросла

def create\_up\_column(x):

    start\_value = x['week1\_open']

    week1\_max = x['week1\_high']

    week2\_max = x['week2\_high']

    week3\_max = x['week3\_high']

    week4\_max = x['week4\_high']

    res = []

    for st, wm1, wm2, wm3, wm4 in zip(start\_value, week1\_max, week2\_max, week3\_max, week4\_max):

        val = max(wm1, wm2, wm3, wm4)

        if (val - st) / st \* 100 > 8:

            res.append(1)

        else:

            res.append(0)

    return pd.Series(res)

Далее полученный вектор значений добавляется в существующий датасет.

new\_df = df.assign(up\_class=create\_up\_column)

new\_df['up\_class'].value\_counts()

Получаем конечный датасет.

Заключительным этапом является обучение.

Применяется TfidfVectorizer для представления постов в виде математической модели.

from sklearn.feature\_extraction.text import TfidfVectorizer

Данные разбиваются на тренировочную и тестовую выборки.

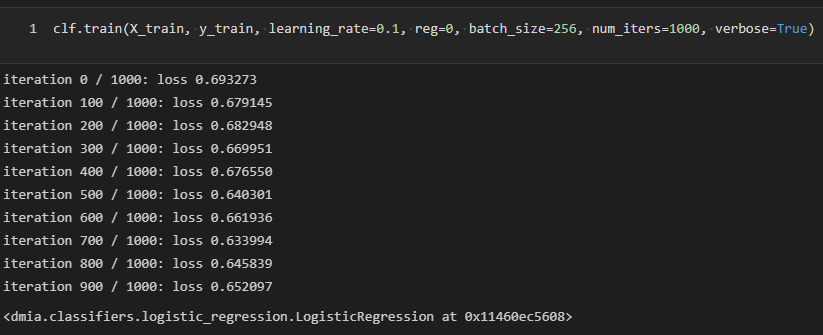
X = tfidfed

y = train\_df['up\_class'].values

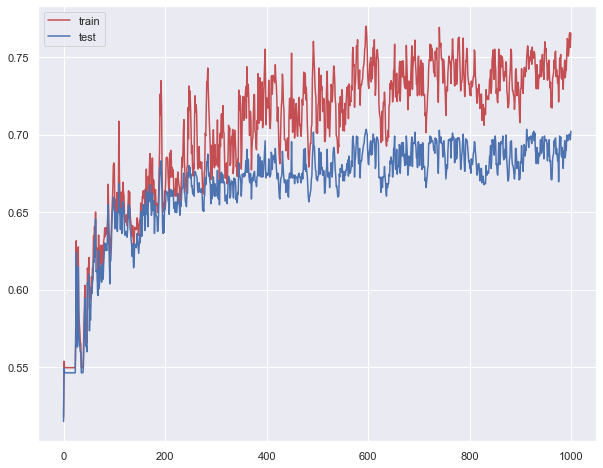
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, train\_size=0.7, random\_state=42)

Создается модель логистической регрессии и данная модель обучается на полученных данных.

Обучение производится на 1000 итераций.



Точность обучения модели



Так как модель логистической регрессия является интерпретируемой, то результаты ее обучения можно посмотреть.

Пример слов часто встречающихся в новостях при положительной динамике акций:

['xmoex', 'бизнес', 'месяца', 'европа', 'страхования', 'тинькова', 'рынках', 'цод', 'биржу', 'прибылей', 'себестоимость', 'новатэк', 'белуга', 'михельсон', 'тгк', 'дает', 'раскрытия', 'акцию', 'электроэнергию', 'апсайдом', 'продаж']

Пример слов часто встречающихся в новостях при отрицательной динамике акций:

['сургута', 'putinomics', 'ювелирных', 'обыкновенных', 'пресс', 'башнефти', 'ртс', 'мск', 'механизм', 'mmbtu', 'мсфо', 'эмитентов', 'дивидендами', 'городах', 'росте', 'интеко', 'раньше']

## Анализ.

По результатам обучения модели видно, что во множество слов, оказывающих наибольшее влияние на классификацию новости, входят названия компаний.

Результат работы модели показал точность в 70%, при оценке положительных результатов. Данная система оценивает новость о положительной или нейтральной динамике стоимости акции компании на горизонт в 1 месяц.

В процессе работы из датасета были удалены числовые значения, что позволило повысить точность прогнозирования на 15%.

По результатам можно заметить ключевые слова, которые человек воспринимает как

Положительные:

'прибылей', 'раскрытия', 'апсайдом', 'рост', 'быстрорастущих', 'повышена' 'повышать', 'расширять', 'плюс'

Негативные:

'угрожает', 'аномально', 'рискует', 'слабее', 'угрозой', 'переработки'

Что говорит о том, что модель смогла выделить ключевые закономерности.

# Вывод

Анализ соотношения информации из телеграм каналов и стоимости акций показал, что новости, которые публикуются в телеграм каналах не всегда могут оказать положительную динамику на акции. На стоимость акции компании также влияет и общее настроение на фондовом рынке. На настроение рынка также могут влиять геополитические риски, которые могут сильно изменить стоимость акций конкретных компаний, из-за чего положительные новости могут не отразиться на стоимости акции.

Результат работы показал, что новости не содержат достаточной информации для краткосрочных инвестиций.

Телеграм каналы не всегда предоставляют информацию заранее или качество информации может быть сомнительным, часто новости, опубликованные в телеграм каналах, уже заложены в стоимости акции.

По результатам работы можно увидеть, что информация из телеграм каналов может не повлиять на стоимость акции. Цена акции может зависеть от настроения фондового рынка, из-за чего не стоит принимать инвестиционные идеи, опираясь на информацию из данных источников, так как информация из телеграм каналов не всегда может показать положительную динамику портфеля инвестора.