



МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ  
КАЗАХСТАН


Казахский национальный исследовательский технический  
университет имени К.И. Сатпаева

Институт кибернетики и информационных технологий  
Научно-образовательный центр математики и кибернетики

6M070500 – Математическое и компьютерное моделирование

**УТВЕРЖДАЮ**

Директор НОЦ МиК

 Н.С. Даирбеков

8 июня 2020 г.

**ЗАДАНИЕ**

**на выполнение магистерской диссертации**

Магистранту *Ташбаевой Арнагуль Амиргалиевне*

Тема: *Прогнозирование цены на нефть с использованием веб-анализа настроений*

Утверждена приказом Ректора Университета №\_\_-п от "\_\_"\_\_20\_\_г.

Срок сдачи законченной диссертации "8" июля 2020г.

Исходные данные к магистерской диссертации: *Статьи о предсказании цены, публикации обработки естественного языка.*

Перечень подлежащих разработке в магистерской диссертации вопросов:

- а) *Сбор и анализ информации по исходным данным*
- б) *Анализ существующих алгоритмов предсказания цен на нефть*
- в) *Изучение метода интеллектуального анализа текста*
- г) *Разработка предсказательных моделей на Питоне*
- д) *Заключение*
- е) *Список использованной литературы*

Перечень графического материала (с точным указанием обязательных чертежей):

*Презентация диссертации на 40 слайдах*

Рекомендуемая основная литература:

1 Gumus, M. and Kiran, M. S. Crude oil price forecasting using xgboost // International Conference on Computer Science and Engineering (UBMK)', IEEE, pp. 1100–1103, 2017

2 Wright, P. Knowledge Discovery in databases: Tools and Techniques // ACM, 1998

3 Wang Y. and Wang X.J. A new approach to feature selection in text classification // 4<sup>th</sup> International conference on machine learning and cybernetics, IEEE, vol 6, pp 3814-3819, 2005





## ГРАФИК

подготовки магистерской диссертации

Наименование разделов, перечень разрабатываемых вопросов	Сроки представления научному руководителю	Примечание
Сбор материала по основным понятиям рынка нефти и обзор ценовой литературы	11.03.2020	
Обзор литературы. Обзор предварительных методов и сбор базы данных	20.04.2020	
Предварительная обработка текста и модель обучения	10.05.2020	
Анализ полученных результатов и выбор модели	29.05.2020	

### Подписи

консультантов и нормоконтролера на законченную магистерскую диссертацию с указанием относящихся к ним разделов диссертации


Наименования разделов	Консультанты, И.О.Ф. (уч.степень, звание)	Дата подписания	Подпись
Сбор материала по основным понятиям рынка нефти	Р.И.Мухамедиев, профессор, д-р инж.наук	8.07.2020	
Обзор предварительных методов и сбор базы данных	Р.И.Мухамедиев, профессор, д-р инж.наук	8.07.2020	
Создание предсказательной модели и анализ результата	Р.И.Мухамедиев, профессор, д-р инж.наук	8.07.2020	
Нормоконтроль	М.С.Амирбекова, ведущий инженер	8.07.2020	

Научный руководитель \_\_\_\_\_

  
(подпись)

Р.И.Мухамедиев  
(Ф.И.О.)

Задание принял к исполнению обучающийся \_\_\_\_\_

  
(подпись)

А.А.Ташбаева  
(Ф.И.О.)

Дата "08" июля 2020 г.

## АНДАТПА

2020 жылы болған басты оқиғалардың бірі мұнай бағасының 30% -ға төмендеуі және мұнай бағасы 25 доллардың критикалық деңгейіне жетуі болды. Елдердің жинақталған валюталық резервтері алдағы бірнеше жылға жетуі мүмкін, бірақ мұнай бағасының құлдырауы экономикаға қатты әсер етеді. Нәтижесінде көптеген ғалымдар мұнай бағаларын болжаудың сенімді модельдерін жасауға тырысуда.

Магистрлік диссертацияның негізгі мақсаты - Brent маркалы мұнайдың бағасын болжау үшін мәтіндік сезім талдауы қолданылатын жаңа үлгіні жасау. Жұмыс үш бөлікке бөлінеді: 1) құрылымданбаған мәліметтерді жинау әдістемесі, атап айтқанда мұнай компанияларына қатысты жаңалықтар тақырыптары; 2) VADER көңіл-күйін талдауды қолдана отырып, жаңа функциялардың статистикалық талдауы; 3) гребневая (ридж) регрессиясын, кездейсоқ орманды және градиентті күшейтуді қоса алғанда, машиналық оқытудың үш жіктеуішін әзірлеу және бағалау.

Жұмыс нәтижелері көрсеткендей, гребневая регрессиясы ең жақсы болжам нәтижелерін береді. Сонымен қатар, деректердің жаңа түрлерін қолдана отырып, болжау дәлдігі бұрыннан бар модельдердің дәлдігімен салыстырғанда жоғары.

## АННОТАЦИЯ

Одним из главных событий, произошедших в 2020 году – это падение цен на нефть на 30% и достижение критической отметки в 25 долларов. Накопленные валютные резервы стран могут хватить на ближайшие несколько лет, однако обвал цен на нефть сильно ударит по экономике. В результате многие ученые пытаются разработать надежные модели для прогнозирования цен на нефть.

Основная цель данной магистерской диссертации – это разработка новой модели, которая использует анализ настроений текстов для предсказания цены на нефть марки Brent. Работа делится на три части: 1) методология сбора неструктурированных данных, а именно новостные заголовки, имеющих отношение к нефтяным компаниям; 2) статистический анализ новых функций с использованием анализа настроений VADER; 3) разработка и оценка трех классификаторов машинного обучения, включая гребневой (ридж) регрессии, случайного леса и градиентного бустинга.

Результаты работы показали, что гребневая регрессия обеспечивает наилучшие результаты прогнозирования. Кроме того, используя эти новые типы данных, точность прогнозирования несущественно, но превосходит точность ранее существующих моделей.

## ANNOTATION

One of the main events that took place in 2020 was a drop in oil prices by 30% and reaching a critical mark of \$ 25. The accumulated foreign exchange reserves of countries may be enough for the next few years, but a collapse in oil prices will hit the economy badly. As a result, many scientists are trying to develop reliable models for forecasting oil prices.

The main goal of this master's thesis is to develop a new model that uses text sentiment analysis to predict the price of Brent oil. The work is divided into three parts: 1) the methodology for collecting unstructured data, namely news headlines related to oil companies; 2) a statistical analysis of new functions using sentiment analysis VADER; 3) the development and evaluation of three classifiers of machine learning, including ridge regression, random forest and gradient boosting.

The results of the work showed that ridge regression provides the best forecasting results. In addition, using these new data types, prediction accuracy is superior to the accuracy of previously existing models.

## СОДЕРЖАНИЕ

1	ВВЕДЕНИЕ	10
	1.1 Введение	10
	1.2 Актуальность исследования	12
	1.3 Цель и задачи	12
	1.4 Содержание	13
2	ТЕОРИЯ ЦЕНЫ И РЫНОК НЕФТИ	14
	2.1 Введение	14
	2.2 Теоритические аспекты цены	14
	2.3 Типы нефти	15
	2.4 Факторы, влияющие на цену нефти	15
	2.5 Гипотеза эффективного рынка и случайных блужданий	19
3	ОБЗОР ЛИТЕРАТУРЫ: СТАТИСТИЧЕСКИЕ МЕТОДЫ	21
	3.1 Обзор предварительных методов	21
	3.2 Поиск знаний в базах данных	22
	3.3 Интеллектуальный анализ данных и анализ текста	23
	3.4 Классификация текста	26
	3.5 Прогнозирование цены на нефть с использованием текстового майнинга	31
4	МЕТОДОЛОГИЯ	33
	4.1 Сбор данных	33
	4.2 Анализ временных рядов	36
	4.3 Обработка текста	40
	4.4 Анализ настроений текста	41
	4.5 Модель обучения	46
5	АНАЛИЗ И РЕЗУЛЬТАТЫ	50
	5.1 Выбор модели	50
	5.2 Эффект всестороннего текстового настроения	52
	ЗАКЛЮЧЕНИЕ	55



ССЫЛКИ	57
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	60
ПРИЛОЖЕНИЯ	61

# 1 ВВЕДЕНИЕ

## 1.1 Введение

Нефть известна как «индустриальная кровь» и является необходимым стратегическим источником энергии.

Падение цен на нефть в конце 2020 года имеет значительное влияние на все государства. В то время как некоторые страны пожинают плоды низких тарифов на нефть, другие пострадали сильно. Будучи стратегическим главным источником энергии, нефть имеет большое влияние на экономический рост, рынки облигаций и национальную безопасность, поэтому не удивительно, что экономисты потратили большое количество ресурсов пытаясь предсказать его движение. Прогнозирование стоимости позволит компаниям смягчить их риски против колебаний цен.

На стоимость нефти влияют не только фундаментальные факторы спроса и предложения, но и неосновные факторы. Неосновные факторы включают в себя темпы производства, иностранные санкции, экономический рост и сезонность потребления. Ранние работы экономистов сосредоточены на исторических ценах на нефть, ценах на дополнительные товары (например, уголь, природный газ), уровень запасов и финансовые инструменты цены (фьючерсы) как метод прогнозирования движения цен. Немногие ученые смоделировали цену на нефть как функцию, зависящую не только от нефтяных переменных. Не нефтяной переменной может быть, как пример, внешняя политика. Примером может служить ситуация, когда в 2012 году была сокращена добычу иранской нефти с 2,4 млн. баррелей в день до 1,4 млн. баррелей в день из-за санкций США против Ирана, что повлияло на общую цену нефти. К этому выводу пришел и исследователь Zhao, говоря, что «предыдущие исследования по прогнозированию цен на нефть в основном были сосредоточены на области количественного анализа» [1]. В работах Hong так же были рассмотрены шесть факторов, которые влияют на прогноз цен на сырую нефть (спрос, предложение, финансовый рынок, товарный рынок,

спекуляции и геополитика) используя регрессию LASSO, чтобы обнаружить, что «прогноз восьми шагов вперед может значительно уменьшить среднеквадратичную ошибку прогноза» [2].

Однако, быстрое развитие моделей машинного обучения и нейронных сетей принесло новый бум в области моделирования. Вслед за новыми возможностями исследователи предприняли попытки использования машинного обучения в прогнозировании цен на нефть. Так, например, Minggang Wang, использовал в исследовании комбинированные многослойные сети, нейронную сеть Эльмана и эффективные функции случайных событий для прогнозирования колебаний цен и пришел к результату, что «гибридная модель имеет улучшение показателей среднеквадратичной ошибки примерно на 13% по сравнению с простыми регрессионными моделями» [3]. Несмотря на хорошие результаты, исследования в большинстве случаев зависят от публикации официальной макроэкономической статистики, которая отбирается, анализируется и агрегируется регулирующими органами, и возникают проблемы, когда данные не чувствительны к экономическим проблемам в реальном времени.

С быстрым развитием технологий больших данных, неструктурированные большие данные обеспечивают новый источник для прогнозирования. В последнее время многие исследования текстового анализа внесли значительный вклад в прогнозы рыночных цен. Одним из работ по этой области можно привести работу исследователя Ling Liu, которая «извлекла систему индикаторов из Twitter акционерной компании, чтобы проанализировать ее связь с доходностью акций, и результаты показывают, что индикаторы Twitter и цены на акции лучше связаны, чем традиционные промышленные индикаторы» [4]. Некоторые исследования посвящены анализу настроений и распознаванию тем веб-текстов, а также поиску более глубокой информации для прогнозирования. Изучением данного вопроса занимался исследователь Paul C. Tetlock. В своей работе он измерил взаимодействия между СМИ и фондовым рынком количественно, используя ежедневную

информацию из популярной колонки Wall Street Journal, и результаты показывают, что «пессимизм СМИ имеет прогностическую силу для цен на фондовом рынке» [5]. Исследователь Felix Wex так же распределял новости по темам, как ОПЕС, Crude oil, JET и NSEA и извлекал количественные показатели для различных тем и использовал линейную регрессию для проверки их способности прогнозировать доходность цен на нефть WTI. «Результаты показывают, что эффект является статистически значимым» [6].

Соответствующий анализ и исследования в основном сосредоточены на прогнозировании тенденций, а не на прогнозировании стоимости. В то же время используются только линейные модели, поэтому необходимо так же учитывать нелинейные модели для прогнозирования. На основании приведенного выше анализа в работе предлагается новая модель прогнозирования цен на нефть, основанная на текстовом анализе.

## **1.2 Актуальность работы**

По вычислительному прогнозу фондового рынка опубликовано больше статей, которые используют цифры, такие как цена акций, объемы, доход компании, стоимость компании и так далее. В настоящее время ограничено количество исследований по применению анализа настроений текста для предсказания стоимости нефти. По этой причине, важно выяснить, как фондовые рынки реагируют на последние новости связанные с нефтью.

## **1.3 Цель и задачи**

Цель: Оценка влияния новостных заголовков на прогнозирование цены на нефть.

Задачи:

- Изучение ранее существующие алгоритмы анализа ценовых трендов.

Выявление основных факторов влияния на стоимость нефти.

- Изучение метода интеллектуального анализа текста, которые могут быть использованы в попытке создания.

- Изучение и разработка разных типов моделей для предсказания цен на нефть. Сравнение и выбор наилучшей модели для прогнозирования. Оценка внедрения анализа текста при прогнозировании.

## **1.4 Содержание**

Магистерская диссертация состоит из шести глав, включая введение и заключение. Глава 2 объясняет некоторые общие сведения о нефти, о факторах, влияющих на цену нефти. Глава 3 дает общие понятия сбора и работы с текстом для дальнейшей работы. Глава 4 описывает методологию, готовится база данных, объясняется обработка данных и выбирается модель для анализа. В главе 5 описание проведенные эксперименты и конечные результаты. Диссертация завершается кратким изложением экспериментальных результатов. Так же в работу включены формулы, рисунки, использованные источники и скрипт кода на программе Python.

**Научные статьи и публикации:** По теме диссертации опубликована 1 статья в IT & M2020 The international Scientific Conference, Апрель 2020.

## 2 ТЕОРИЯ ЦЕНЫ И РЫНОК НЕФТИ

### 2.1 Введение

Как уже упоминалось в 1.3 Цель и задачи, целью диссертации является оценка воздействия новостных заголовков на цену на нефть и на рынок нефти в целом. Для углубленного понимания проблемы, нам сначала нужно понять маркетинговые цены, а так же лучше понять рынок нефти. Эта глава посвящена этим вопросам.

### 2.2 Теоретические аспекты цены

Теоретически понятие цены описывает денежную стоимость предмета.

Классическая экономическая теория использует концепцию **спроса и предложения** для определения понятия равновесной цены. В частности:

- **Спрос** – это количество товара, которое покупатели хотят купить по приемлемой цене.
- **Предложение** – это количество товара, которое продавцы хотят продать.
- **Цена** рассматривается как баланс между спросом и предложением.

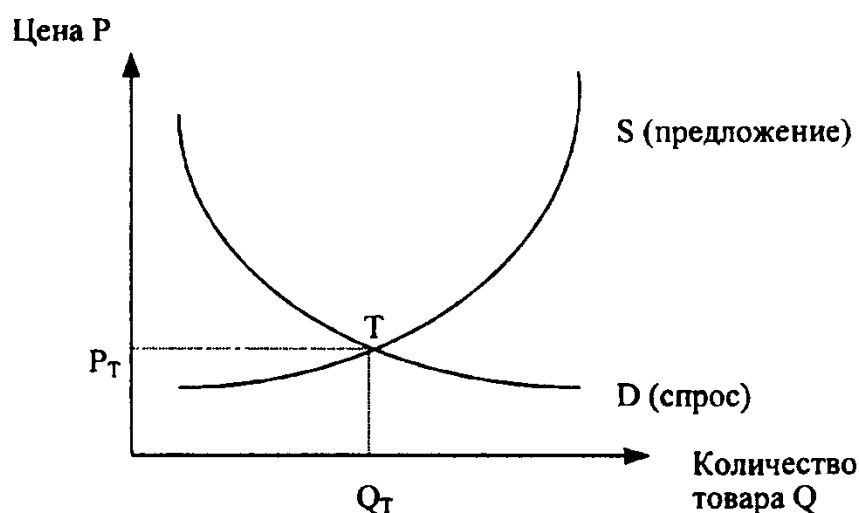


Рисунок 2.1. График спроса и предложения

Рисунок 2.1 показывает кривые спроса и предложения, точка пересечения этих кривых определяет равновесную цену или цену, по которой обмен состоялся. Эта модель особенно подходит для чистых товарных рынков с недифференцированным продуктом.

### **2.3 Типы нефти**

Нефть не является однородным товаром. Есть более 160 различных международных торгуемых видов нефти, которая варьируется с точки зрения качества и производителя. Для сырой нефти качественными характеристиками являются плотность и содержание серы. Более легкая сырая нефть обычно содержит большую долю углеводородов, и поэтому считается ценным продуктом. При добыче тяжелой сырой нефти операторы сталкиваются с характерными трудностями. Сырая тяжелая нефть имеет характерную относительную плотность, сравнимая с плотностью воды и не обладает высокой текучестью. Кроме того, она имеет высокое содержание асфальтена, тяжелых металлов, серы и азота. Качество сырой нефти определяет уровень предобработки и повторной обработки, необходимый для достижения оптимального сочетания продукта на выходе. В результате цены и различия в ценах на сырую нефть также отражает относительный случай рафинирования. Например, первичная сырая нефть, Западного Техаса (WTI) , американского эталона, или нефть марки Brent имеют относительно высокое содержание природного бензина на выходе.

### **2.4 Факторы, влияющие на цену нефти**

Для оценки цены на нефть используется концепция спроса и предложения, которая сильно зависит от таких факторов, как запасы сырой нефти и нефтепродуктов, валовое внутреннее производство, курсы валют, погодные условия и настроения рынка. Основным фактором, формирующим мировой спрос на нефть, является рост мировой экономики.

Показатели мирового спроса, мировых цен на нефть и мировой экономической динамики приведены на рисунке 2.3. Бобылев, говоря о

формировании цен на нефть, пишет, что «снижение темпов роста мировой экономики неизменно приводит к падению мировых цен на нефть» [8]. Так, Азиатский финансовый кризис стал причиной падения мировых цен на нефть. Однако за период с 2002 по 2008 году цены на нефть выросли почти в 4 раза, а именно с 22,46 доллара США до 96,04 доллара США за баррель. Это связывают с поднятием ВВП всего мира на 4%. [9].



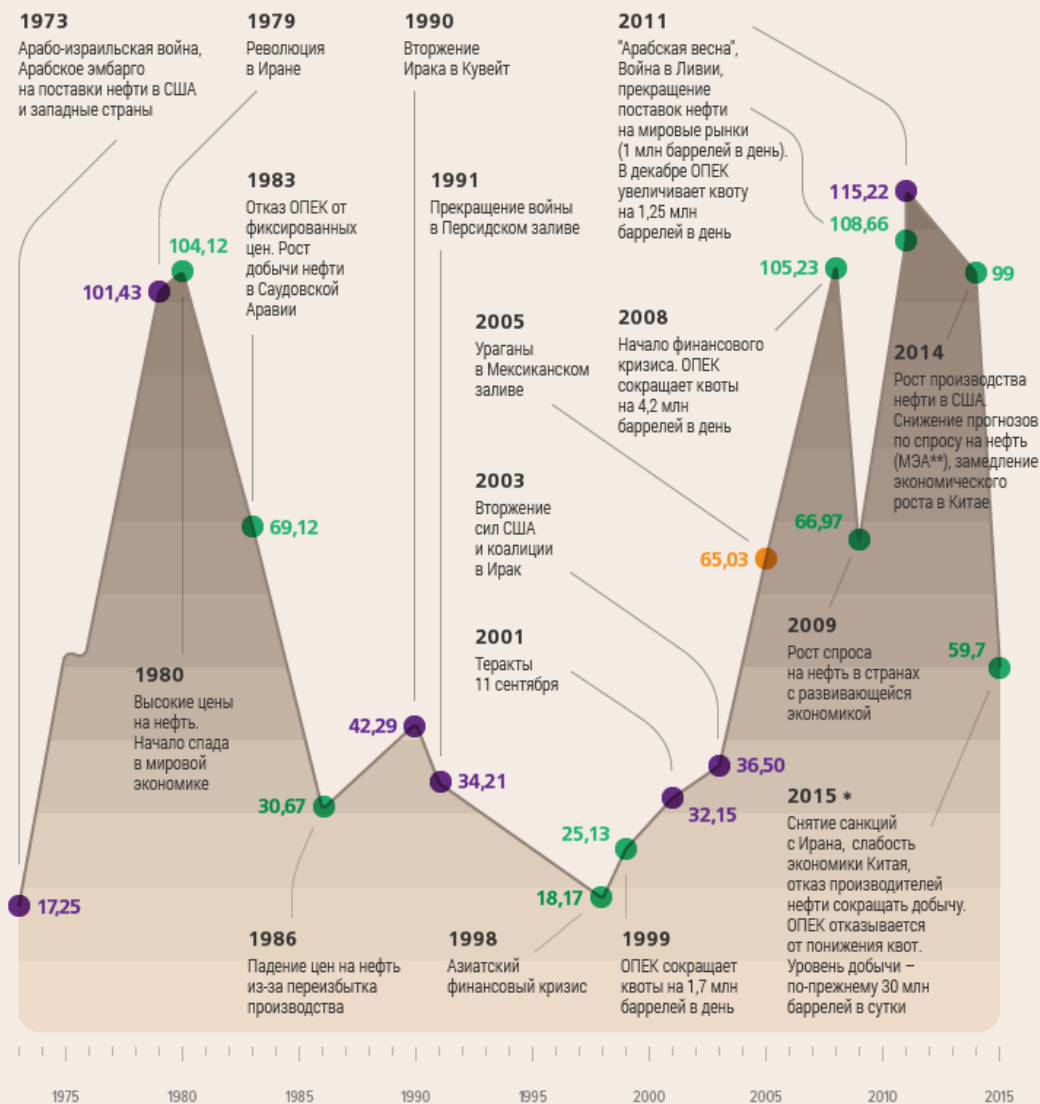
## ЦЕНЫ НА НЕФТЬ ЗА 40 ЛЕТ

В графике указана среднегодовая цена на нефть с учетом инфляции, \$ за баррель.  
1973-1984 — цены на нефть марки Arabian Light из порта Рас-Танура,  
1984-2015 — цены на нефть марки Brent.

● Политические события

● Экономические события

● Природные катаклизмы



\* Указана средняя цена за шесть месяцев 2015 года.

\*\* МЭА (International Energy Agency) — Международное энергетическое агентство.

Источник: BP Statistical review of world energy

Рисунок 2.3. Мировой спрос на нефть с 1975 – 2015 гг. Источник ТАСС

В настоящее время, глобальный экономический шок, произошедший по вине пандемии COVID – 19, привел к падению цен на нефть большинство

сырьевых товаров и, Р. Джапарова высказалась, что «в результате этого цены на энергоносители (с учетом природного газа) в 2020 году снизились на 40%»[10].

Ведущую роль в формировании мирового спроса на нефть играют промышленно развитые страны. Можно выделить три доминирующих центра мирового потребления нефти: Северная Америка (прежде всего США), Западная Европа и Азиатско-Тихоокеанский регион, прежде всего Китай и Япония как показано на рисунке 2.4.

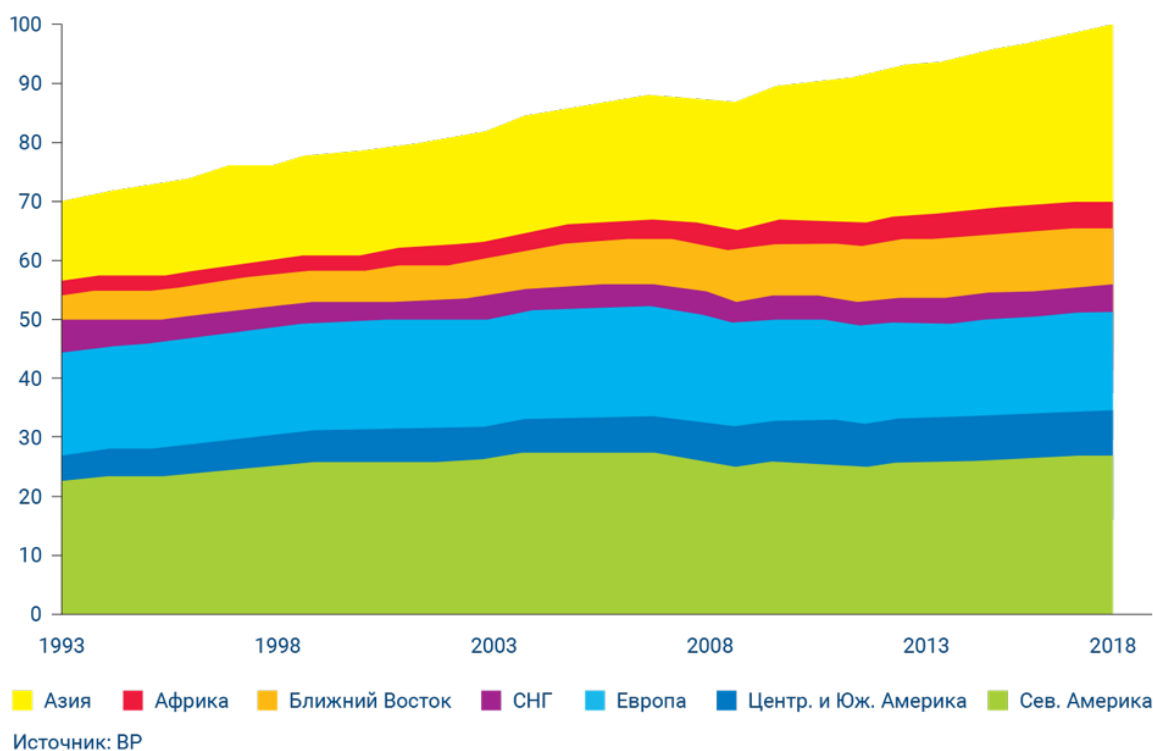


Рисунок 2.4. Потребление нефти по регионам, млн баррель/сутки. Источник журнал «Сибирская нефть». Приложение «нефтяной рынок. Просто о сложном» № 165 (ноябрь 2019)

Предложение нефти на мировом рынке определяется спросом на нефтепродукты и соответственно теми факторами, которые формируют данный спрос. В то же время на объемы предложения (добычи) нефти влияют геолого-технические факторы. На конец 2018 года запасы нефти составляет 1729,7 млрд тонн и в целом позволяют обеспечить перспективный мировой спрос на нефть.

На рисунке 2.5 можно увидеть список стран, обладающих самыми большими запасами нефти по состоянию 2018 - 2019 года:

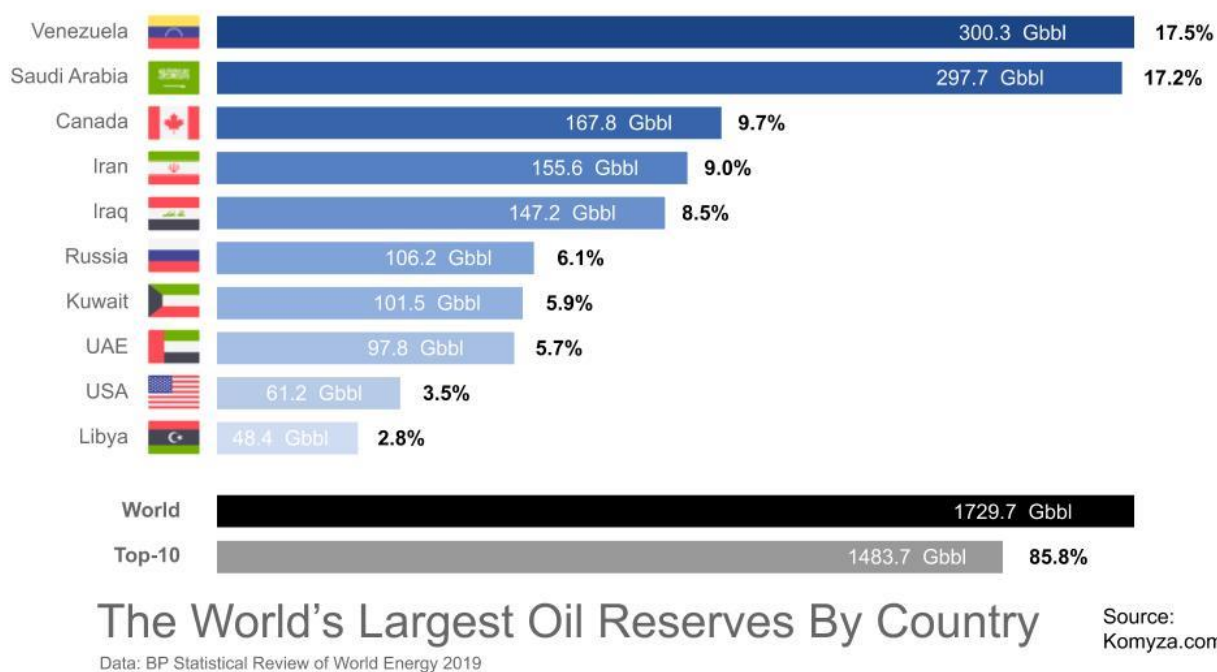


Рисунок 2.5. Страны, обладающие самыми большими резервами нефти.

Источник <https://komyza.com/>

## 2.5 Гипотеза эффективного рынка и теория случайных блужданий

«Гипотеза эффективного рынка – гипотеза, согласно которой вся существенная информация немедленно и в полной мере отображается на рыночной курсовой стоимости ценных бумаг» [11]. Другими словами, цены акций отражают всю доступную информацию о них. Ниже перечислены «три основные гипотезы эффективного рынка» [26]:

- **Слабая эффективная форма:** Стоимость рыночного актива отражает только прошлую информацию.
- **Средняя эффективная форма:** Стоимость рыночного актива полностью отражает прошлую, но и публичную информацию (отчеты компании, выступление государственных служащих, аналитические прогнозы).

- **Сильная эффективная форма:** стоимость рыночного актива полностью отражает всю информацию, а именно прошлую, публичную и внутреннюю информацию (инсайдерская информация).

Нюансы и противоречия гипотезы эффективного рынка:

- Наличие человеческого фактора (случайные ошибки, предвзятость) негативно влияют на эффективность рынка, мешая долгосрочному прогнозу.
- Рыночные пузыри и обвалы – явления, обесценивающие гипотезу.
- Единичные примеры победы эффективного рынка на практике.

Гипотеза свободного блуждания говорит, что цены финансовых инструментов испытывают случайные колебания, то есть цена актива, которая возникает на рынке в следующий момент времени, не зависит от его цены в предшествующий момент. Таким образом, существует 50% вероятность, что цена акций пойдет как вверх, так и вниз.

Теория свободного блуждания аналогична с гипотезой эффективного рынка в слабой форме, поскольку, как и в первой, так и во второй гипотезе говорят о независимости будущей цены актива от его предыдущей цены. Но гипотеза эффективного рынка более масштабная модель, чем теория свободного блуждания, так как включает в себе не только прошлую информацию о цене, но и всю прошлую информацию касающегося данного актива. Стоит отметить, что если гипотеза эффективного рынка верна, то и теория свободного блуждания будет так же верна. На эффективном рынке, цена актива будет находиться в состоянии равновесия и будет меняться только в момент получения новой информации. В зависимости от настроения информации будет меняться и цена, как в положительную, так и в отрицательную сторону. Если же гипотеза эффективного рынка не подтвердилась, то цены активов испытают определенные тренды. Но и в таком случае можно утверждать, что теория свободного блуждания имеет место быть. В результате конкретное соотношение спроса и предложения данного актива на

рынке может вызвать понижение или повышение цены в следующий момент времени.

## 3 ОБЗОР ЛИТЕРАТУРЫ: СТАТИСТИЧЕСКИЕ МЕТОДЫ

### 3.1 Обзор предварительных методов

Существуют множество опубликованных статей и исследований о добыче данных, временных рядах, прогнозировании цен, однако мало статей охватывающий текстовый анализ для рыночных прогнозов. Самые ранние работы, которые начали использовать текстовых данных для финансового прогнозирования относятся к 1991 году.

Authors	Paper Title	Year
Abramson and Finizza	"Using belief networks to forecast oil prices"	1991
Abramson and Finizza	"Probabilistic forecasts from probabilistic models: a case study in the oil market"	1995
Morana	"A semi parametric approach to short-term oil price forecasting"	2001
Ye et al.	"Forecasting Crude Oil spot price using OECD inventory levels"	2002
Abosedraa and Baghestani	"On the predictive accuracy of crude oil futures prices"	2004
Ye et al.	"A monthly crude oil spot price forecasting model using relative inventories"	2005
Fang et al.	"A generalized pattern matching approach for multi step prediction of crude oil prices"	2006
Sadorsky	"Modeling and forecasting petroleum futures volatility"	2006
Ye et al.	"Forecasting short-run crude oil price using high- and low-inventory variables"	2006
Zhang et al.	"A new approach for crude oil price analysis based on Empirical Mode Decomposition"	2007
Deesa et al.	"Modeling the world oil market- Assessment of a quarterly econometric model"	2007
Gori et al.	"Forecast of oil price and consumption in the short term under three scenarios: Parabolic, linear and chaotic behavior"	2007

Таблица 3.1. Статьи по прогнозированию цен нефти

Общая структура ранее опубликованных работ имеет много общего между собой. Множество работ построены вокруг алгоритма обучения, в основном на классификаторе для прогнозирования настроения новостных статей. Для качественной работы алгоритма нужен качественный

тренировочный набор. Генерация этих тренировочных наборов осуществляется разными способами.

### 3.2 Поиск знаний в базах данных

Как отмечает в своей работе Wright: «количество данных, собираемых из базы данных, сегодня намного превышает наши возможности сокращать и анализировать без использования методов аналитического анализа» [12]. Новое поколение вычислительных методов и инструментов требуется для извлечения полезных знаний из быстро растущих данных. Эти методы являются предметом расширяющейся области открытия знаний в базе данных (KDD – Knowledge Discovery in Databases) и интеллектуального анализа данных [13].

Открытие знаний в базе данных (KDD) является нетривиальным процессом извлечения потенциально полезного и понятного шаблона данных. Иными словами, это процесс поиска полезных знаний в сырых данных. Поиск полезных данных включает в себе вопросы подготовки данных, выбора признаков, очистки данных, применение методов сбора данных, постобработка данных и анализ полученных результатов. Из выше перечисленных процессов, сбор данных (майнинг) является самым основным и важным методом, позволяющий добыть полезные знания.

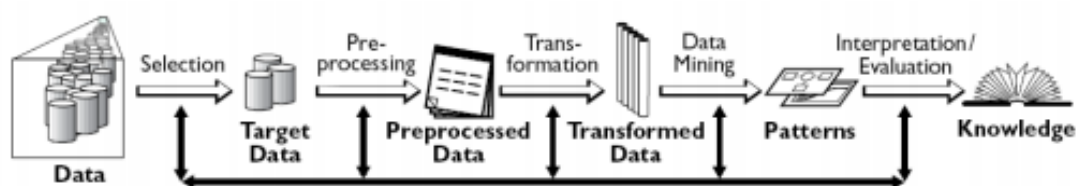


Рисунок 3.1. Процесс поиска знаний в базах данных (Fayyad, U.M., 1996)

Процесс KDD поэтапно:

- **Подготовка исходного набора данных.** Этот этап заключается в создании набора данных из различных источников.
- **Предобработка данных.** Для эффективного анализа данных, следует обратить внимание на предобработку данных. Редко встречаются готовые

и качественные данные, обычно данные содержат в себе шумы, пропуски, некорректные значения. Иногда данные могут быть не полноценными, и требуется дополнить данные некоторой информацией. Предобработка данных – это важнейший этап, который нельзя пропустить, так как без этого этапа мы не получим на выходе качественные данные для исследования.

- **Трансформация, нормализация данных.** Этот этап также необходим, так как приводит данные в пригодную форму для последующего анализа. На данном этапе данные приводятся в определенный вид. К примеру, нейронные сети работают только с данными числового формата, и данные обязаны быть нормализованы.
- **Data mining.** Главный этап в поиске полезной информации в базах данных. На этом шаге применяются различные алгоритмы для нахождения знаний. Примерами алгоритмов могут быть: нейронные сети, деревья решений, алгоритмы кластеризации и так далее.
- **Постобработка данных.** Финальный шаг процесса поиска знаний в базах данных. На этом шаге применяется интерпретация результатов и применение в работах.

Вышеперечисленные этапы поиска знаний в базах данных должны быть выполнены последовательно.

### **3.3 Интеллектуальный анализ данных и анализ текста**

#### **Интеллектуальный анализ текста (Data mining)**

Поиск полезных шаблонов в данных известен под разными именами, включая интеллектуальный анализ данных, также в разных отраслях применяется термины, как извлечение знаний, раскрытие информации, археология данных. Термин «интеллектуальный анализ данных» чаще используются статистиками, исследователями в области базы данных, и в последнее время термином пользуются в бизнес сообществе. Как говорилось



ранее, интеллектуальный анализ данных (data mining) является важным шагом в процессе поиска знаний в базе данных (KDD).

Задачи, решаемые с использованием сбора данных:

- Классификация - это отнесение входного данного к одному из заранее известных классов.
- Кластеризация – это разделение множества входных данных на группы (кластеры) по степени «схожести».
- Сокращение описания – это сжатие объемов собираемой информации.
- Ассоциация – это поиск повторяющихся образцов.
- Прогнозирование – нахождение будущего состояния объекта на основании предыдущих событий.
- Анализ отклонений
- Визуализация данных

Некоторые авторы определяют интеллектуальный анализ данных как процесс, с помощью которого открывают знания, которые они не знают. В случае больших баз данных, пользователи спрашивают невозможное: «скажи мне что-то, чего я не знал, но хотел бы знать» [14].

### **Анализ текста (Text mining)**

Сеть представляет собой большую и растущую коллекцию текстов. Это количество текста становится ценным источником информации и знания. Чтобы помочь извлечь информацию из текста появилась новая область под названием «открытие знаний в текстах» (KDT), который связан с понятием процесса открытия знания в базе данных (KDD).

Анализ текста является одним из новейших дисциплин, и обычно занимает много времени и академических обсуждений, прежде чем понять концепцию анализа текста и дать точное определение. Так как в начале исследования, ученые не могли определиться и дали 15 разных определений

для этого термина. Таким образом, нет твердой границы между понятиями анализа данных, поиска знаний в тексте, поиск информации и извлечение информации.

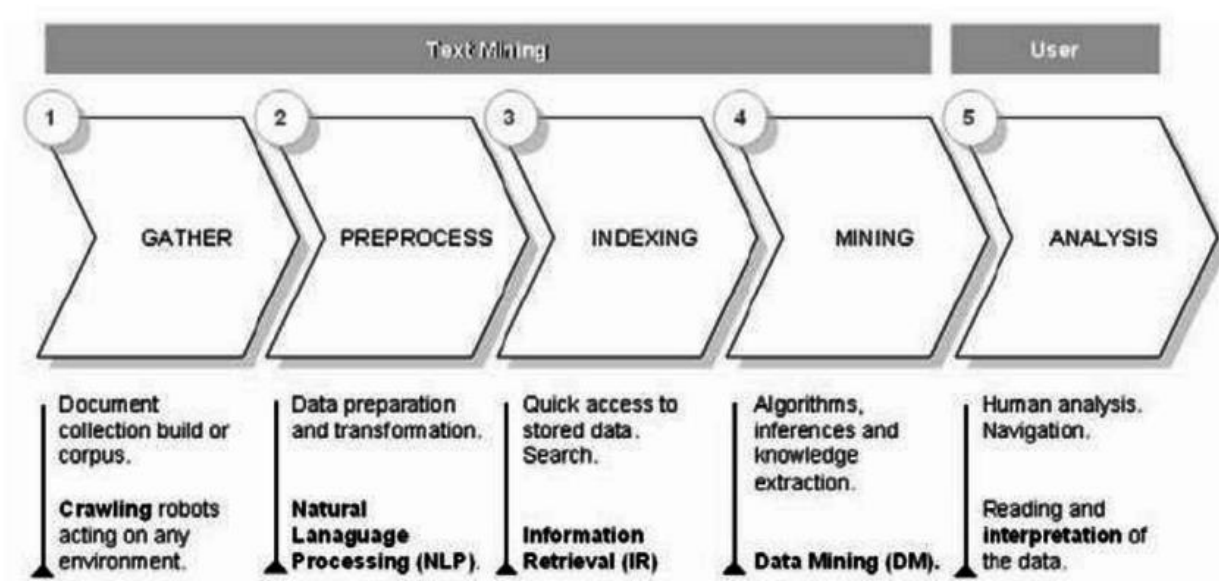


Рисунок 3.2. Этапы анализа текста

Как показано на рисунке 3.2, процесс анализа текста содержит следующие этапы:

- Сбор и идентификация текстовых источников из Интернета, файлов, документов и так далее.
- Распознавание именованных объектов – этот этап используется для идентификации именованных текстовых объектов, таких как людей, организации, географических названий и так далее.
- Устранение неоднозначностей – на этом этапе используется контекстные подсказки для интерпретации неоднозначных объектов, как например машина – это транспортное средство и так же может иметь значение механизма.
- Распознавание объектов, идентифицированных по шаблону.
- Класстеризация документов – идентификация похожих данных.

- Идентификация имен существительных и других терминов, относящихся к одному и тому же объекту.
- Анализ настроений включает в себе распознавание субъективного аспекта

### 3.4 Классификация текста

Т.В. Батура в своей работе «Методы автоматической классификации текстов» дал определение классификации текстов, как «одной из главных задач компьютерной лингвистики, в связи с тем, что к ней сводится ряд остальных задач, к примеру определение тематической принадлежности текстов. Для обеспечения информационной и публичной безопасности важное значение имеет изучение в телекоммуникационных сетях информации, который содержит незаконные данные (также данных, связанных с терроризмом, наркототиками, сетевым экстремизмом, подготовкой протестных движений или массовых беспорядков)» [15].

Следует понимать, что классификация текстов отличается от кластеризации. При классификации документов категории, классы, уже заранее определены, в то время как при кластеризации они не заданы.

Формально задачу классификации документов можно описать скудеющим образом:

Имеются множества документов  $D = \{d_1, \dots, d_{|D|}\}$  и множество возможных категорий (классов)  $C = \{c_1, \dots, c_{|C|}\}$ . Неизвестная целевая функция  $\Phi: D \times C \rightarrow \{0,1\}$  задается формулой

$$\Phi(d_j, c_i) = \begin{cases} 0, & \text{если } d_j \text{ не является элементом } c_i \\ 1, & \text{если } d_j \in c_i \end{cases} \quad (1)$$

Автор работы «Методы автоматической кластеризации текстов», Батура, выделил четыре решения задач классификации [15]:

- Предобработка и индексация документов

- Уменьшение размерности
- Построение и обучение классификатора с помощью методов машинного обучения
- Оценка качества классификации

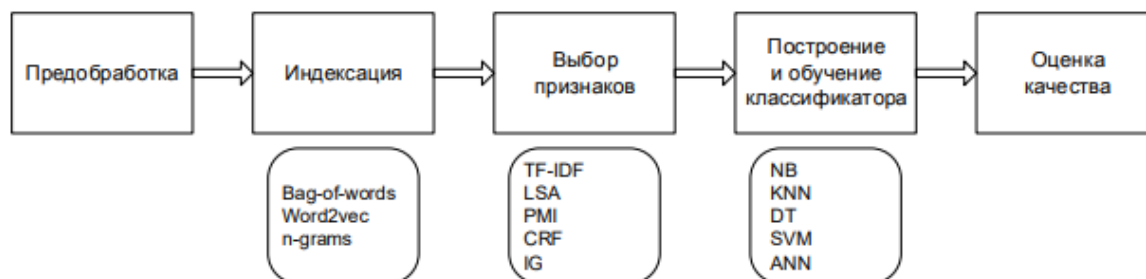


Рисунок 3.3. Этапы процесса автоматической классификации текстов

### Предобработка и индексация документов.

Предварительная обработка текста включает в себя **токенизацию**. Токенизация – это процесс разделения текстового потока на символы, слова, фразы или другие значимые элементы называемые токенами.

Существуют много способов токенизации текстовых потоков в токены. Простой способ будет просто разделить текст на пустых местах, но более улучшенным методом является разделение с помощью знаков препинания. Например, попробуем токенизировать следующую строку:

**«Hello! This is test number 11. It tests the word\_punct-tokenizer!@test66»**

В первую очередь разделяем по пробелам, затем разделяется по символам. На выходе мы получаем токены такого вида:

```
['Hello', '!', 'This', 'is', 'test', 'number', '11', '.', 'It', 'tests', 'the', 'word_punct', '-', 'tokenizer', '!@', 'test66']
```

Следующим шагом предобработки текста является удаление **стоп слов**. Стоп слова – это слова, которые не имеют важной информации в тексте, но при анализе текста стоп слова мешают для качественного анализа. Примером стоп слов является союзы.

В лингвистической морфологии **стемминг** – это сокращение слова от его склоненной формы. Стемминг нужен для уменьшения размерности элементов, что делает данные менее большими и способствует лучшей работе классификатора. Примером стемминга является:

Cats, catty, catlike → cat

Stemmer, stemming, stemmed → stem

Fishing, fishes, fisher → fish

### Индексация документов

Индексация документов - это построение числовой модели текста, которая переводит текст в удобный формат для дальнейшей обработки. Несмотря на развитие машинного обучения, компьютер не будет понимать сам текст на входе для анализа. По этой причине, используется одна из модель «мешка слов» (bag-of-words). Данная модель позволяет представить текст в виде многомерного вектора слов и его весов.

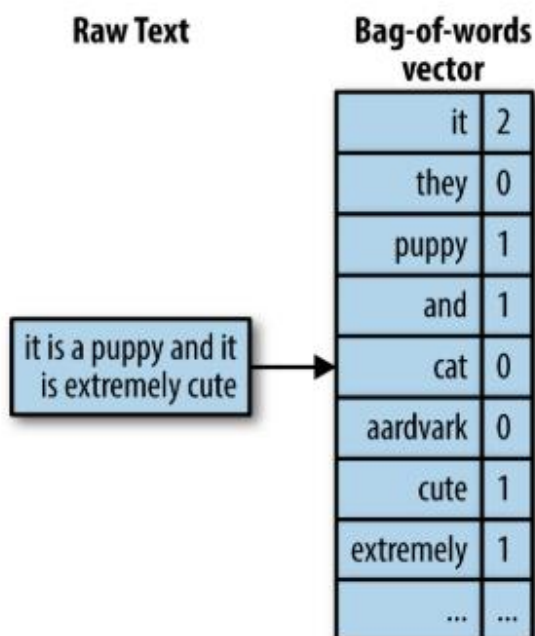


Рисунок 3.4. Модель «мешка слов»

Другая распространенная модель – это **Word2Vec**. Модель представляет каждое слово в виде вектора, который содержит информацию о контекстных словах. Возьмем пример:

**«The cat pushed the glass off the table»**

Данные полученные через эту модель показаны на Рисунке 10. Каждая скобка обозначает единичное контекстное окно. Синее поле означает входной вектор, красное же поле это выходной вектор. Из одного контекстного окна получается два элемента, то есть на одно входное слово приходится два контекстных слова. Размер окна обычно определяется самим пользователем. Чем больше размер окна, тем лучше модель, однако это может замедлить модель.

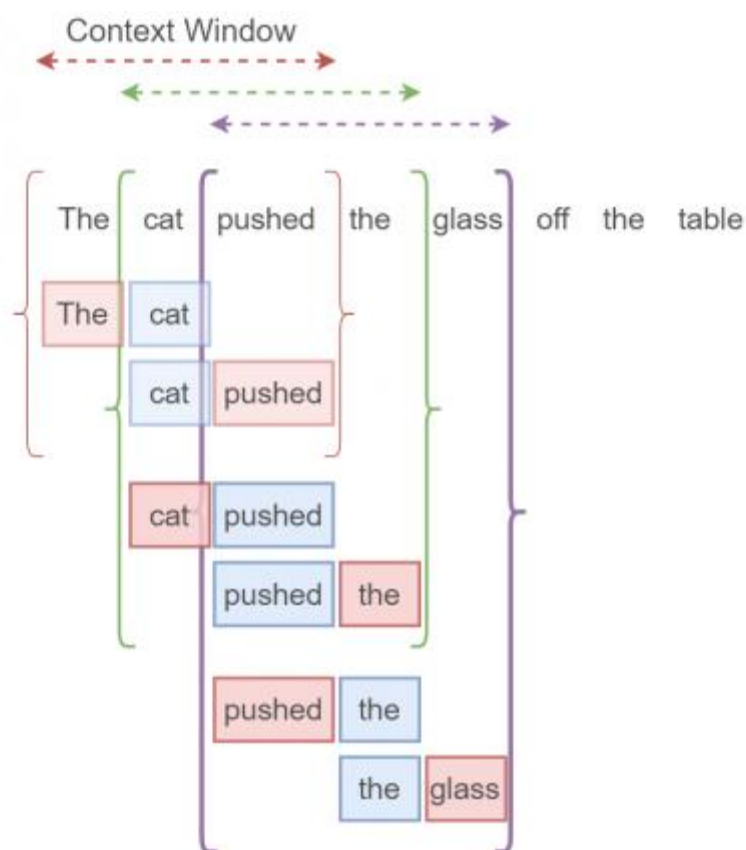


Рисунок 3.5. Word2Vec

## Выбор признаков

Существует несколько способов определения веса признаков документа. Наиболее популярным является функция **TF-IDF** [16]. Его цель состоит в том, чтобы больший вес получали слова которые имеют высокую частоту в пределах конкретного документа и с низкой частотой употреблений в других документах. Форма оценки важности слова в пределах одного документа вычисляется так:

$$TF = \frac{n_{t,d}}{n_d}, \quad (2)$$

где  $n_{t,d}$  - количество употреблений слова  $t$  в документе  $d$ ,  $n_d$  - общее число слов в документе.

Обратная частота IDF:  $IDF = \log(|D|/D_t)$ , (3)

$|D|$  - общее количество документов в коллекции

$D_t$  – количество всех документов, в которых встречается слово  $t$ .

Итоговый вес термина в документе относительно всей коллекции вычисляется по формуле:

$$V_{t,d} = TF * ID \quad (4)$$

## Построение и обучение классификатора

Существуют множества методов классификатора:

- вероятностные (Метод Байеса)
- метрические (Метод к ближайших соседей)
- логические (Дерево решений)
- линейные (логистическая регрессия)
- методы на основе нейронных сетей (RNN, CNN)

### 3.5 Прогнозирование цены на нефть с использованием текстового майнинга

За основу этой диссертационной работы взята работа исследователя Yu и соавторов [17], где использовались методы анализа текста из новостей для прогнозирования цен на нефть. Они заметили, что во время разработки имеет место быть 3 основные проблемы. Первая проблема заключается в том, как найти переменные модели, то есть факторы воздействия, для конкретной модели. В нашем случае, на сырую нефть влияют множества факторов и, следовательно, все факторы должны быть рассмотрены в модели. Вторая проблема это извлечение этих факторов. Третья проблема – это как справиться с несостоятельностью, когда правила извлечения и шаблоны конфликтуют.

В виду первых двух проблем, техника добычи текста ввела поиск и работу с различными модельными переменными. С учетом третьей проблемы может справиться грубая теория множеств, чтобы решить проблему несостоятельностью.

Общий процесс работы Yu изображен на рисунке 3.6.

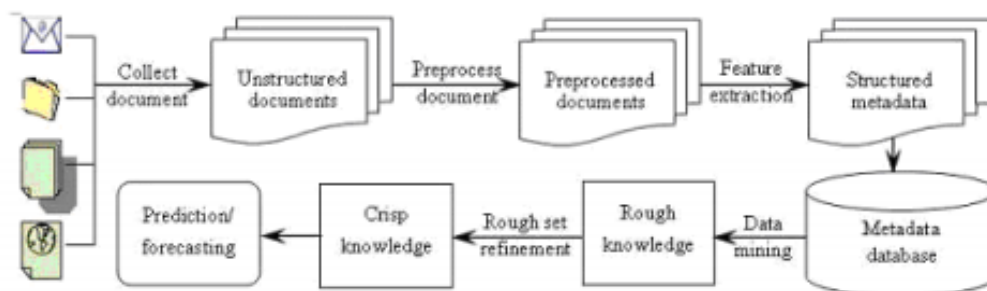


Рисунок 3.6. Основной процесс систем прогнозирования (Yu et al., 2005)

Так же этим автором было опубликовано дополненная модель для прогнозирования. Они изменили модель классификатора и использовали нейронные сети ANN и назвали новую систему «Гибридная модель систем искусственного интеллекта для предсказания цен на сырую нефть» [18]. Система состоит из пяти компонентов, как видно на рисунке 3.7.



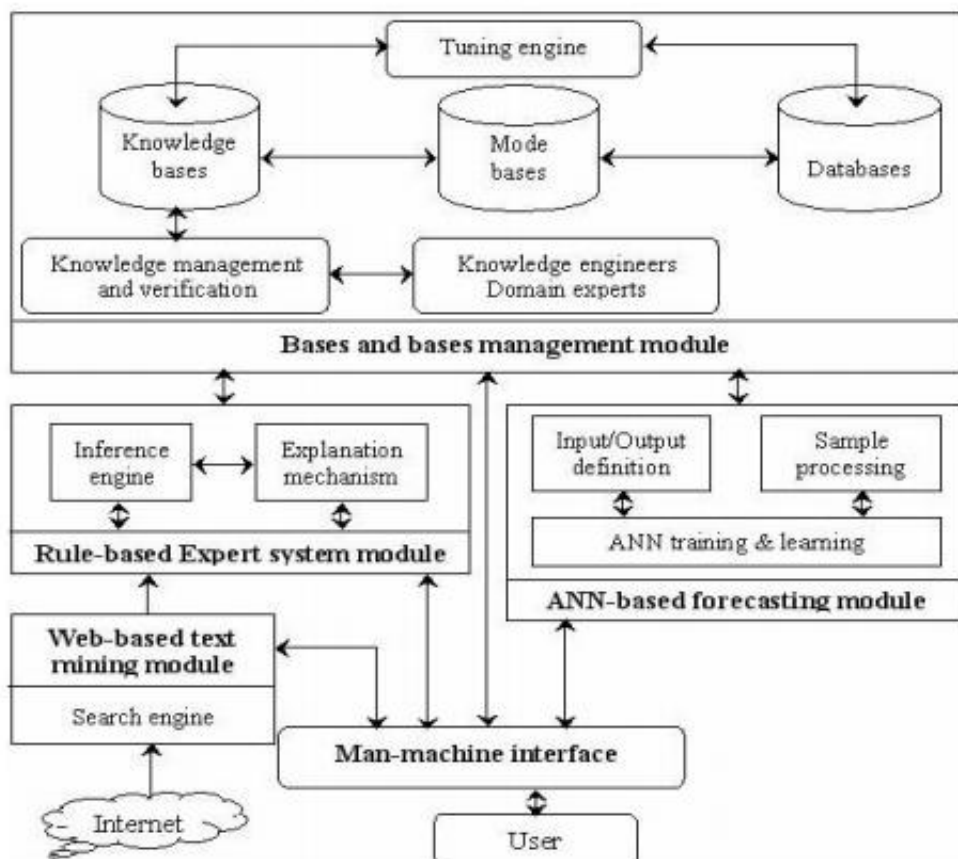


Рисунок 3.7. Основные компоненты системы прогнозирования на основе гибридного ИИ

«ANN, используемый в этом исследовании, представляет собой трехслойную нейронную сеть обратного распространения (BPNN), включающий еще алгоритм Levenberg – Marquardt для обучения» [18]. В работе нейронная сеть обратного распространения отслеживает предыдущие и текущие знания и прогнозирует будущие значения, используя исторические статистические данные. В работе предсказывалась цена нефти Западного Техаса (WTI).

## 4 МЕТОДОЛОГИЯ

### 4.1 Сбор данных

Reddit.com – это всемирно известный социальный новостной сайт, который представляет информацию в реальном времени и хранит сотни тысяч новостей о финансовых инвестиционных продуктах, включая глобальные акции, иностранные валюты, фьючерсы, облигации, фонды и цифровые валюты и многое другое.

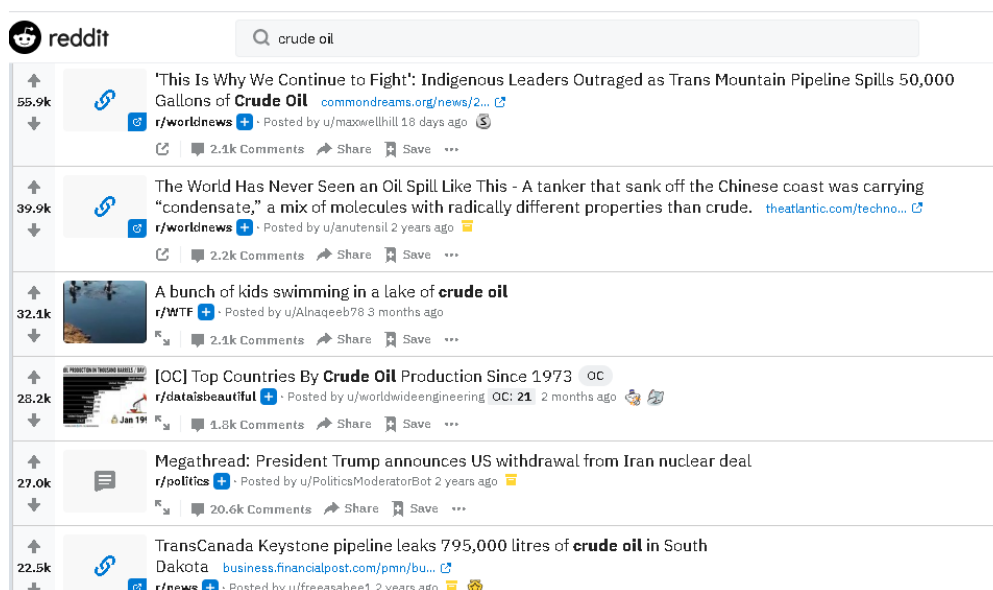


Рисунок 4.1. Reddit.com

Нужно отметить три пункта:

- **Почему новостные заголовки вместо самих новостей?** Новостные заголовки описывают в себе краткую мысль самого текста. Так как мы собираем данные с официального сайта, можно не бояться за некорректные новостные заголовки. Так же новостные заголовки уже хранят в себе все ключевые слова, которые привлекают пользователей.
- **Почему новости фондового рынка вместо новостей о нефти?** Здесь есть две причины такого выбора. Первая причина в том, что новостных данных касающейся только нефти может собраться только на 2000 данных, что очень мало для прогнозирования. Вторая причина

состоит в том, что как говорилось ранее, сырая нефть имеет в себе множество факторов воздействия, и на цену акции может влиять также добыча газа, золота и так далее. По этой причине, нужно брать во внимание все касающиеся данные фондового рынка для качественной модели.

- **Как осуществляется сбор новостных заголовков?** Для сбора текста будет использоваться веб-скрапинг. Веб-скрапинг – это сбор данных с различных интернет-ресурсов. Существует множество готовых программ для веб-скрапинга, однако за основу всех взято текстовый язык HTML. Ниже приведена визуализация работы веб скрапига.

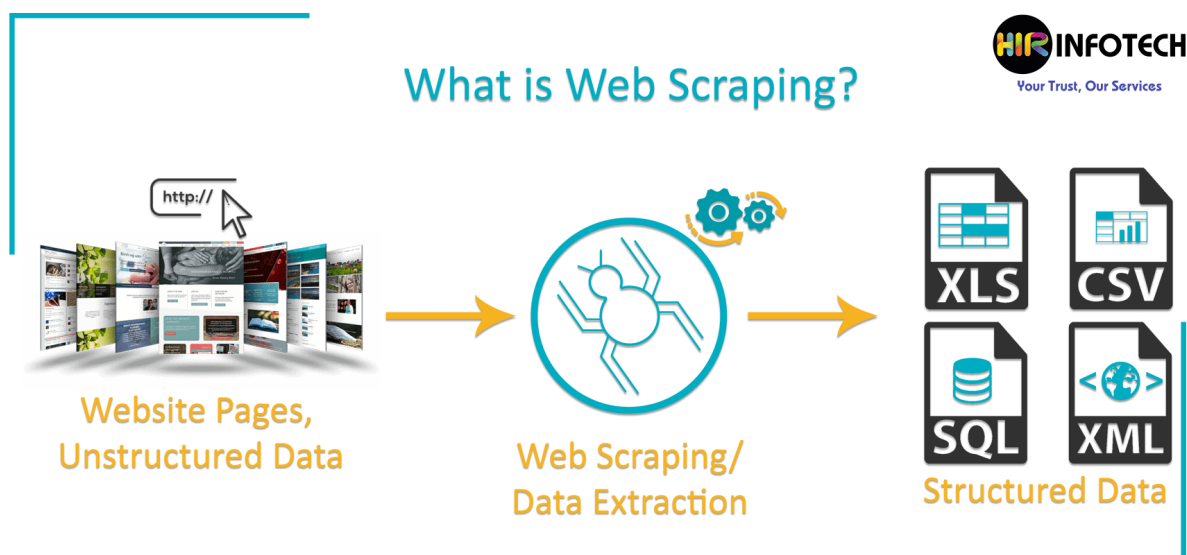


Рисунок 4.3. Веб-скрапинг.

В работе собраны 49725 новостных заголовков. Новостные заголовки были выбраны путем ранжирования по голосам пользователей Reddit.com и только 25 лучших заголовков рассматриваются на одну дату. Новостные данные собраны с 12 августа 2012 года по 12 июня 2020 года.

Данные новостных заголовков выглядят в таком виде (рисунок 4.3):

Date	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top17	Top18	
2012-08-16	b"Georgia 'downs two Russian warplanes' as cou...	b"BREAKING: Musharraf to be impeached.'	b"Russia Today: Columns of troops roll into So...	b"Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b"150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...	...	b"Al-Qaeda Faces Islamist Backlash'	b"Condoleezza Rice: "The US would not act to p...	b"This i c E Unio
2012-08-17	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict	b'Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	...	b"Do not believe TV, neither Russian nor Geor...	b'Riots are still going on in Montreal (Canada...)	b'I overtal manu
2012-08-20	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b"'"If we had no sexual harassment we would hav...	b"Al-Qaeda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...	b'Stratfor: The Russo-Georgian War and the Bal...	b'I'm Trying to Get a Sense of This Whole Geor...	...	b'Why Russias response to Georgia was right'	b'Gorbachev accuses U.S. of making a "serious ...	t Geor NAT W
2012-08-21	b' U.S. refuses Israel weapons to attack Iran...	b"'"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain's policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 "million" quake survivors...	b"'"Bush announces Operation Get All Up In Russi...	b'Russian forces sink Georgian ships'	...	b'US humanitarian missions soon in Georgia - i...	b"Georgia's DDOS came from US sources'	b' convc into vic
2012-08-22	b'All the experts admit that we should legalis...	b'War in South Ossetia - 89 pictures made by a ...	b'Swedish wrestler Ara Abrahamian throws away ...	b'Russia exaggerated the death toll in South O...	b'Missile That Killed 9 Inside Pakistan May Ha...	b"Rushdie Condemns Random House's Refusal to P...	b'Poland and US agree to missile defense deal. ...	b'Will the Russians conquer Tblisi? Bet on it,...	...	b"Georgia conflict could set back Russia's US f...	b'War in the Caucasus is as much the product ...	b"Non photos Ossetia/

Рисунок 4.3 Новостные заголовки с Reddit.com

Что касается числовых данных, в работе предсказывается цена на нефть марки Brent Crude Oil. Brent – эталонная марка нефти, добываемая в Северном Море. С 2007 года является смесью нескольких сортов нефти [25].

Данные были собраны с сайта finance.yahoo.com.

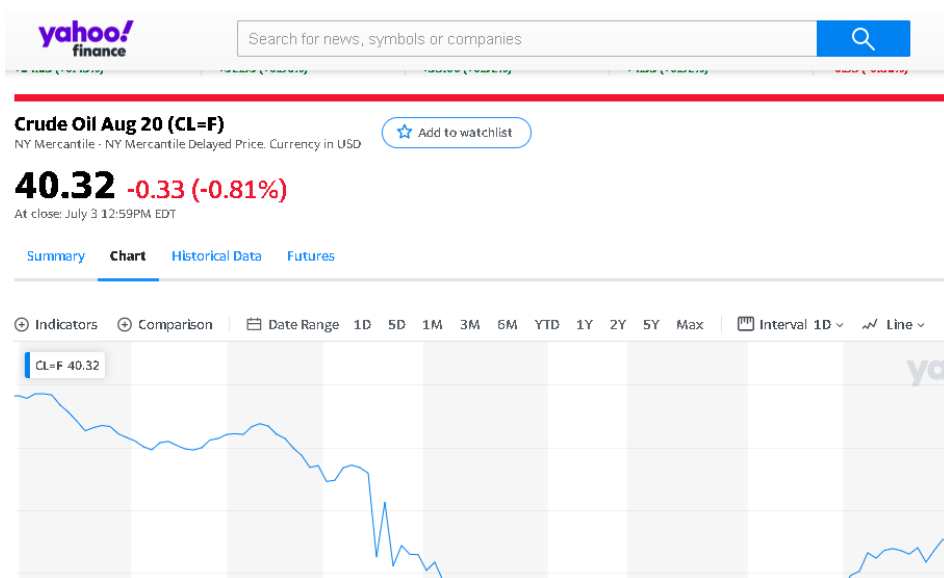


Рисунок 4.4. Нефть марки Brent в finance.yahoo.com

Числовые данные так же собраны с 12 августа 2012 года по 12 июня 2020 года. Всего 1989 наблюдений. Стоит отметить, что торги акции закрыты в праздничные и выходные дни. Данные представлены ниже:

Date	Open	High	Low	Close*	Adj Close**	Volume
Jul 02, 2020	39.84	40.74	39.59	40.32	40.32	130,102,415
Jul 01, 2020	39.79	40.58	39.05	39.61	39.61	171,886,332
Jun 30, 2020	39.41	40.08	38.85	39.66	39.66	143,721,086
Jun 29, 2020	37.75	39.89	37.50	39.63	39.63	154,012,945
Jun 28, 2020	37.96	38.12	37.68	37.85	37.85	631,650
Jun 26, 2020	39.19	39.35	37.79	38.16	38.16	151,232,804
Jun 25, 2020	37.88	39.24	37.08	39.14	39.14	208,026,414
Jun 24, 2020	40.16	40.53	37.31	37.95	37.95	201,095,437

Рисунок 4.5. Исторические данные нефти Brent с 12.08.2012 – 12.06.2020

## 4.2 Анализ временных рядов

**Рядом динамики** (динамическим рядом, временным рядом) называется последовательность значений статистического показателя (признака), упорядоченная в хронологическом порядке, т.е. в порядке возрастания временного параметра. Отдельные наблюдения временного ряда называются **уровнями** этого ряда [19].

Анализ временных рядов – это анализ, основанный на исходном предложении, согласно которому случившееся в прошлом служит достаточно надежным указанием на то, что происходит в будущем [20].

На основе приведенных выше ценовых данных мы рисуем диаграмму временных рядов, которая описывает, как цены на нефть колеблются во времени, показаны на рисунке 4.6.

## Development of stock values from 2012 to 2020

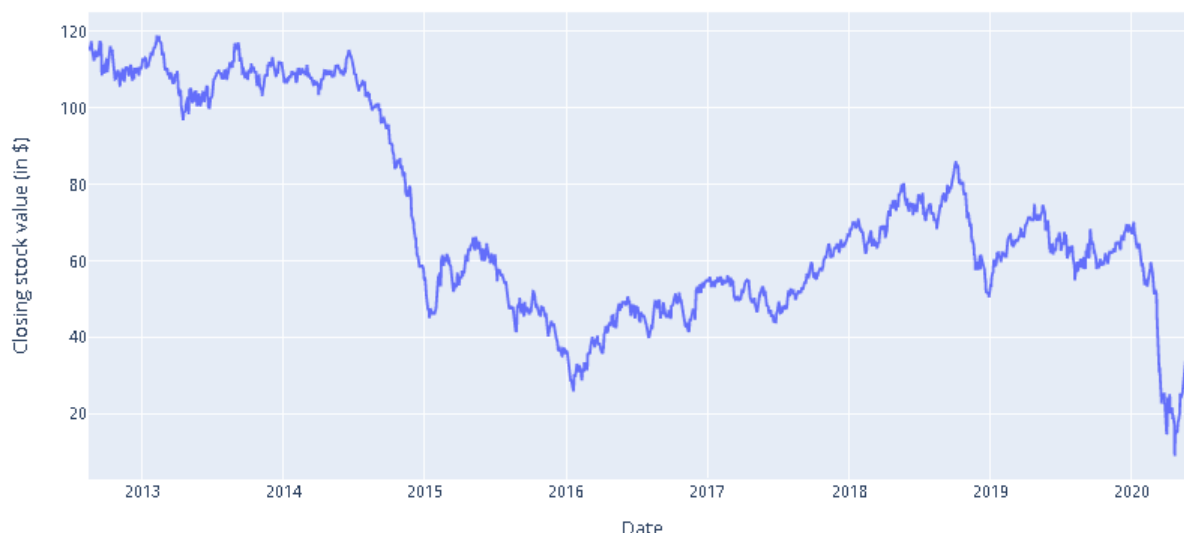


Рисунок 4.6. Ежедневные цены на нефть марки Brent

Как показано на рисунке 4.6, цены на нефть Brent претерпели значительные колебания. Поэтому очень важно прогнозировать колебания цен на нефть, используя соответствующий метод. Чтобы более точно отразить колебания цен на сырую нефть, был проведен статистический анализ данных о ценах на нефть марки Brent, обобщенных в таблице 4.1.

	Open	High	Low	Close	Adj Close	Volume
count	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000	1.971000e+03
mean	64.495885	65.351152	63.548179	64.473059	64.473059	6.250815e+06
std	22.560556	22.498068	22.668682	22.581212	22.581212	4.004763e+07
min	1.400000	13.850000	-39.439999	-2.720000	-2.720000	1.993500e+04
25%	48.489999	49.254999	47.595002	48.450001	48.450001	2.574750e+05
50%	57.299999	57.990002	56.540001	57.200001	57.200001	4.683660e+05
75%	88.889999	89.845001	87.634998	88.745002	88.745002	6.659590e+05
max	110.279999	112.239998	109.110001	110.529999	110.529999	4.599355e+08

Таблица 4.1 Сводка статистических текстов по дневным ценам на нефть марки Brent

Как показано в Таблице 4.2, средние цены на нефть марки Brent составляет 64,47, что означает, что цены на нефть колеблются в районе 63-64 – бальной стоимости. Самая высокая цена на нефть составляет 110,5, а самая

низкая – 22,59. Между максимальной и минимальной ценой существует большой разрыв, и стандартное отклонение составляет 22,4, что означает, что цены на нефть сильно колеблются. Колебания цен показывают нам о нестационарности временных рядов. Нестационарность подразумевает наличия автокорреляции во временных рядах.

Автокорреляция – это корреляционная связь между значениями одного и того же случайного процесса  $X(t)$  в моменты времени  $t_1$  и  $t_2$ . Функция, характеризующая эту связь, называется *автокорреляционной*. Корреляционная связь измеряется с помощью *коэффициента автокорреляции* ( $\rho$ ).

Коэффициент автокорреляции может использоваться для того, чтобы определить, являются ли данные случайными, имеется ли тренд (нестационарность), являются ли данные стационарными, есть ли в них сезонные колебания.

В анализе данных автокорреляция широко используется для анализа и моделирования временных рядов.

Автокорреляция затрудняет применение ряда классических методов анализа временных рядов. Статистика Дарбина – Уотсона предназначена для обнаружения автокорреляции первого порядка.

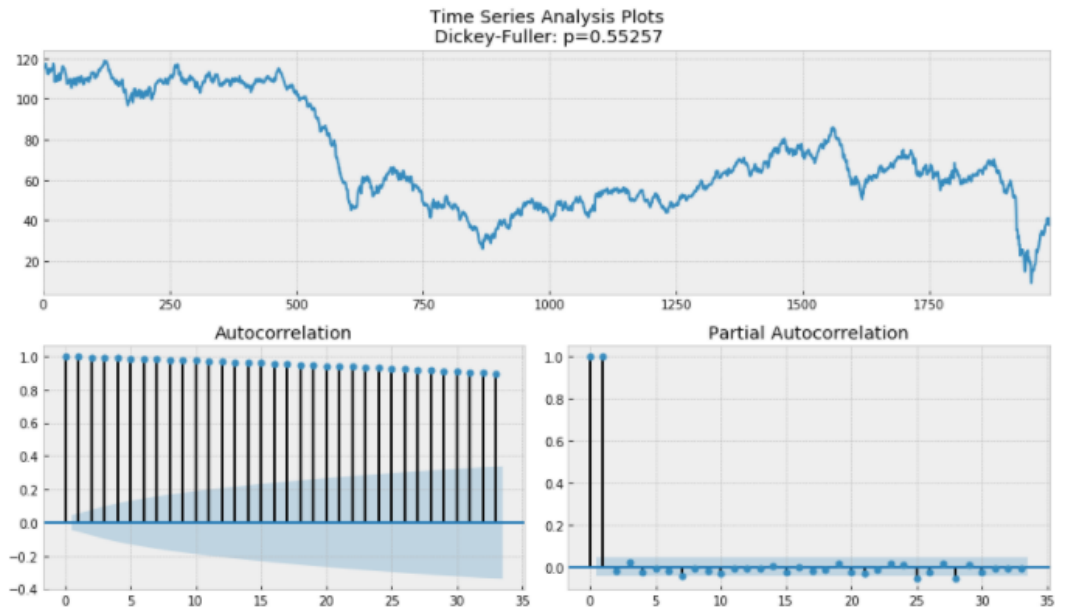


Рисунок 4.7. Анализ временных рядов на стационарность

Мы видим, что наши данные очень нестационарные. Попробуем сделать временной ряд стационарным на рисунке 4.8.

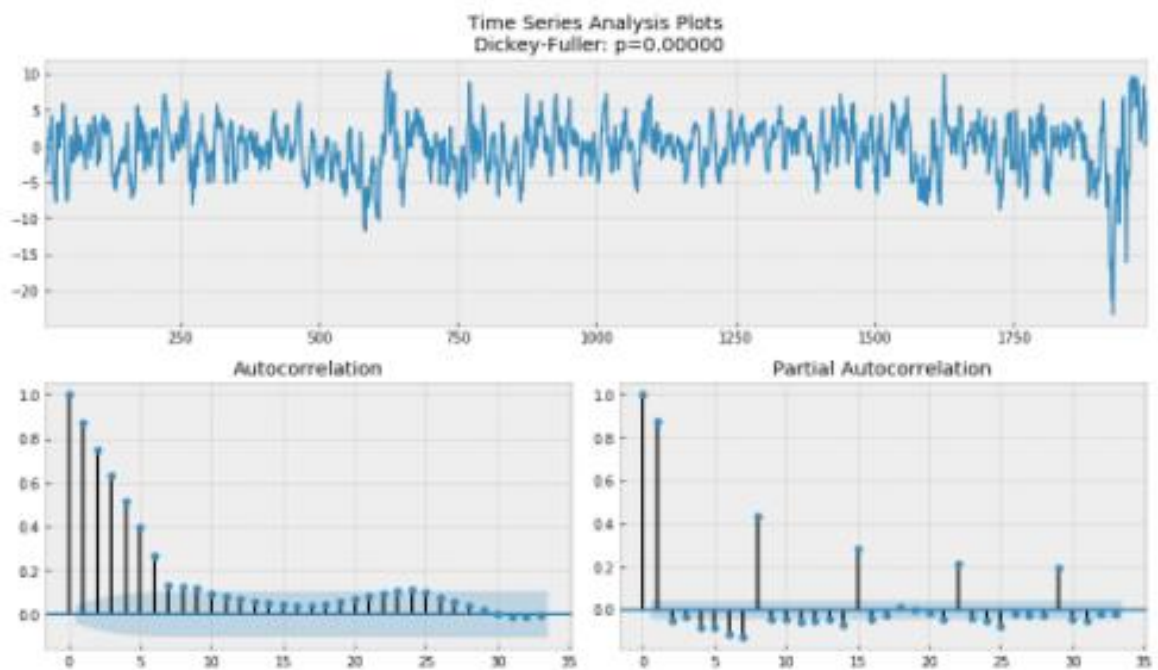


Рисунок 4.8. Анализ временных рядов

Здесь на рисунке, мы видим улучшение с помощью удаления еженедельной сезонности. Однако в автокорреляционном сюжете мы видим много существенных лагов. Попробуем убрать эти лаги.



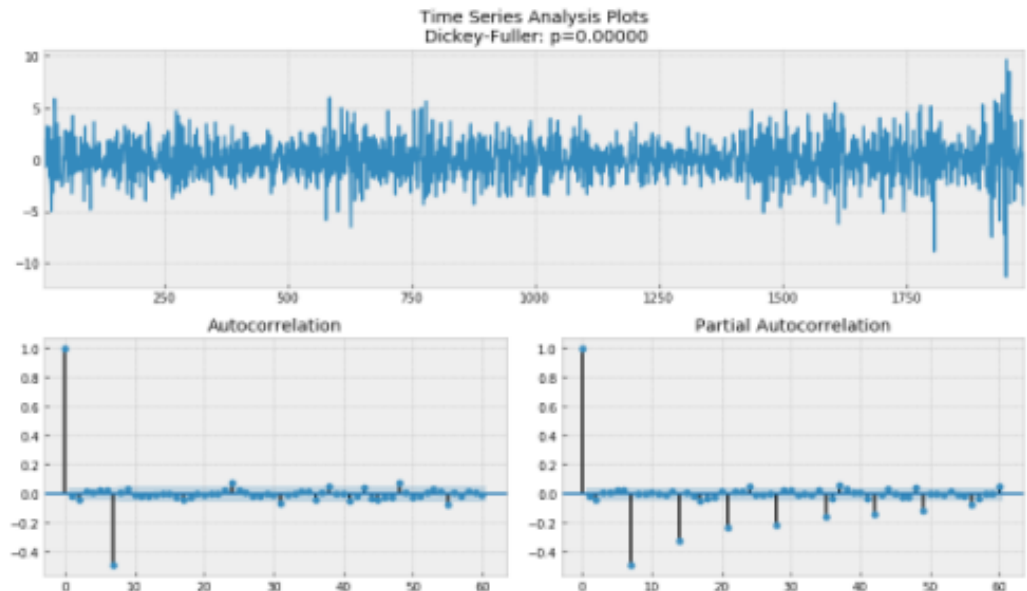


Рисунок 4.9. Анализ временных рядов

Сейчас мы можем видеть конечный результат – стационарный временной ряд, где колебания варьируются вокруг 0. Данные готовы для дальнейшей работы прогнозирования.

### 4.3. Обработка текста

После сбора неструктурированного данного, текста, мы должны провести предобработку перед анализом.

В первую очередь, мы должны преобразить все буквы в тексте в единый формат, в нашем случае, в строчный вид.

```
b"The commander of a Navy air reconnaissance squadron t
hat provides the President and the defense secretary th
e airborne ability to command the nation's nuclear weap
ons has been relieved of duty"
```

```
b"the commander of a navy air reconnaissance squadron t
hat provides the president and the defense secretary th
e airborne ability to command the nation's nuclear weap
ons has been relieved of duty"
```

Рисунок 4.11. Преобразование текста в строчный формат

Далее разделяем предложения в список слов и убираем пунктуации и бессмысленные слова. Рисунок 4.12

```
['the', 'commander', 'of', 'navy', 'air', 'reconnaissance', 'squadron', 'that', 'provides', 'the', 'president', 'and', 'the', 'defense', 'secretary', 'the', 'airborne', 'ability', 'to', 'command', 'the', 'nation', 'nuclear', 'weapons', 'has', 'been', 'relieved', 'of', 'duty']
```

Рисунок 4.12. Разделение текста в список слов

#### 4.4 Анализ настроений текста

Анализ настроений – это процесс определения, является ли фрагмент текста положительным, отрицательным или нейтральным. Для анализа настроений применяются два термина: полярность (фрагменты текста классифицируются как положительные или отрицательные) и валентность (принимается во внимание интенсивность настроения).

В работе используется метод VADER, чтобы проанализировать, и предсказать тенденцию веб-текста [21]. «VADER – это неконтролируемый метод на основе правил» [22]. Первоначально метод VADER использовался для социальных сетей. Он обладает высокой точностью и был принят большим количеством исследовательских институтов. Его основным преимуществом являются:

- Точность и экономичность
- Наличие идеального публичного словаря
- Vader может избежать огромных затрат на тегирование данных

На финансовых рынках метод VADER также применяется для прогнозирования настроения инвесторов. T. Ling в работе отметил, что «с

помощью VADER легко прогнозировать тенденции доходов от активов и выявлять настроения инвесторов по отношению к компании или бренду » [23].

Учитывая сходство между рынками нефти и обычными финансовыми рынками и преимущество VADER, мы выбрали метод VADER для анализа настроений в тексте. Таким образом, благодаря этому методу наши новостные заголовки будут иметь выходные данные как в Таблице 4.2.

Variable Name	Remark
compound <sub>score</sub>	The comprehensive text sentiment score calculated by text sentiment analysis
negative <sub>score</sub>	The negative text sentiment score calculated by text sentiment analysis
neutral <sub>score</sub>	The neutral text sentiment score obtained by text sentiment analysis
positive <sub>score</sub>	The positive text sentiment score calculated by text sentiment analysis

Таблица 4.2. Настройки параметров VADER

Анализ настроений текста с помощью метода VADER работает следующим образом:

**Входные данные:** анализируемый текст, словарь

**Выходные данные:** общий балл, негативный балл, нейтральный балл, позитивный балл

VADER основан на лексиконах слов, связанных с настроениями. Каждое из слов в лексиконе оценивается как положительное или отрицательное. Ниже можно увидеть кусочек лексикона VADER.

слово	Рейтинг настроений
трагедия	-3,4
обрадовался	2,0
ненормальный	-1,7
стихийное бедствие	-3,1
Великий	3,1

## Рисунок 4.13. Словарь VADER

**Пример:** Еда хорошая, а атмосфера приятная.

Предложение состоит из двух слов в лексиконе (хорошая и приятная) с оценками 1,9 и 1,8. Составной балл, представляет собой сумму всех рейтингов лексикона (в данном случае 1,9 и 1,8), которые были стандартизированы в диапазоне от -1 до 1.

Применим этот метод на наши новостные заголовки. Получаем в конце результат показанный на рисунке 4.14.

	Date	Close	compound_mean	compound_max	compound_min	subjectivity_mean	comp
0	2012-08-16	116.12	-0.350337	0.2144	-0.9260	0.163685	
1	2012-08-17	115.20	-0.085277	0.8156	-0.8271	0.202921	
2	2012-08-20	115.50	-0.318394	0.5423	-0.8591	0.374076	
3	2012-08-21	116.03	-0.162032	0.5106	-0.8074	0.176371	
4	2012-08-22	115.77	-0.194879	0.7177	-0.8689	0.319615	
5	2012-08-23	117.46	-0.143104	0.4404	-0.7481	0.227282	
6	2012-08-24	115.76	-0.263546	0.5106	-0.9260	0.216935	
7	2012-08-27	113.74	-0.373172	0.5574	-0.8720	0.256786	
8	2012-08-28	112.62	-0.197157	0.4847	-0.8807	0.095403	
9	2012-08-29	112.53	-0.268522	0.5719	-0.9022	0.107994	

Рисунок 4.14. Метод VADER на новостные заголовки

После предварительной обработки и анализа настроений текста необходимо проанализировать текстовые настроения всех 49725 текстов. После анализа настроений упомянутого выше, выполняется ежедневный процесс интеграции, который относится к усреднению значений намерений настроения для всех статей за день, чтобы обеспечить получение ежедневного общественного мнения для последующего предсказания цен на нефть, и наконец, все дневные тенденции проиллюстрированы на рисунках 4.14 – 4.1

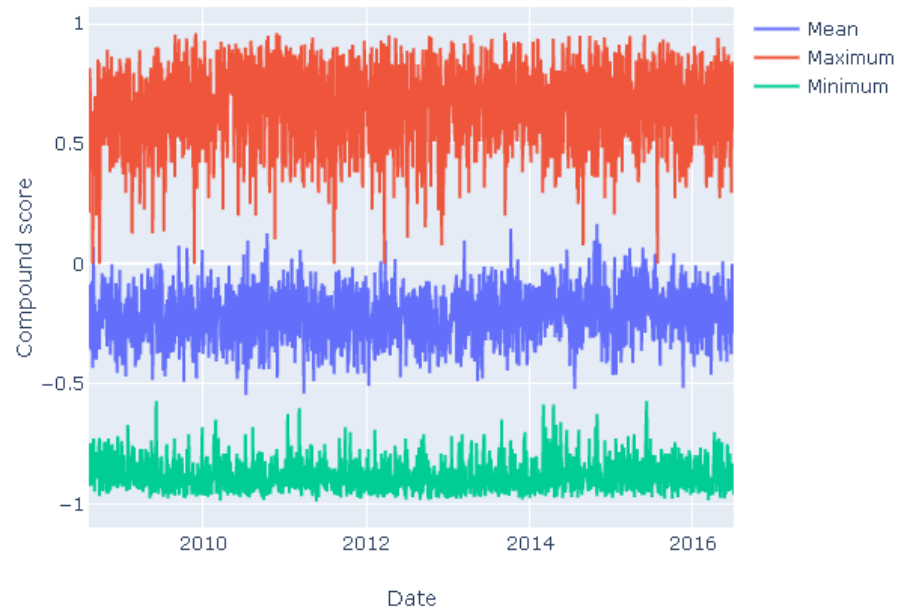


Рисунок 4.14. Ежедневные текстовые настроения для общего балла.

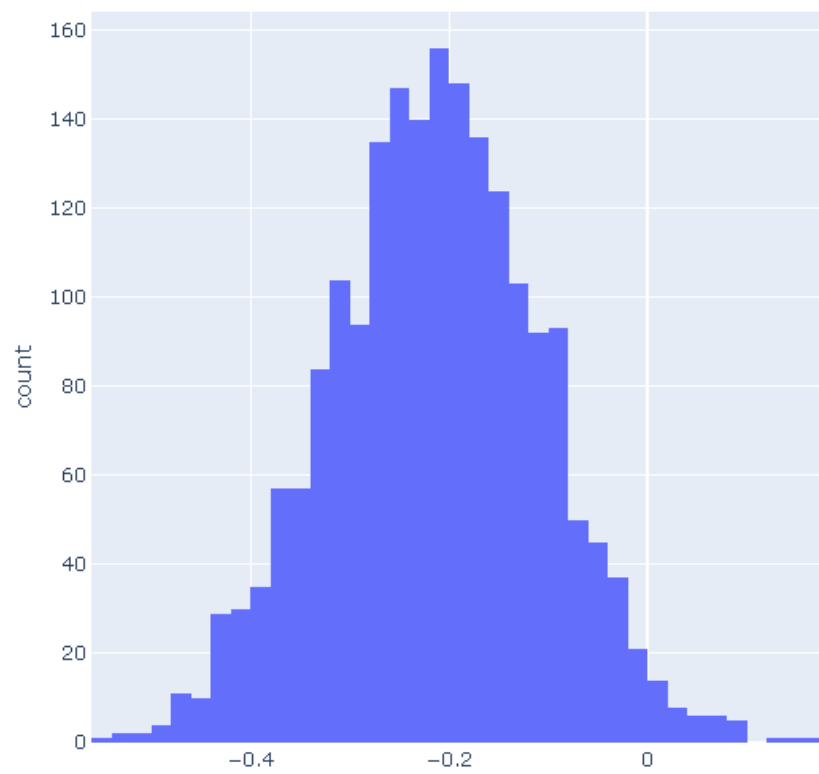


Рисунок 4.15. Распределение общего балла

## Development of subjectivity score

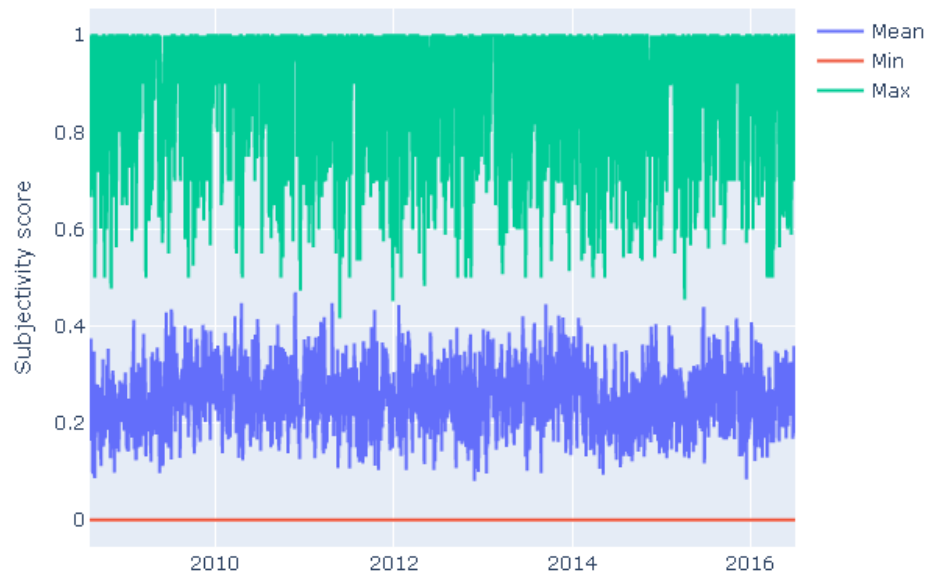


Рисунок 4.16. Ежедневные субъективные оценки текста

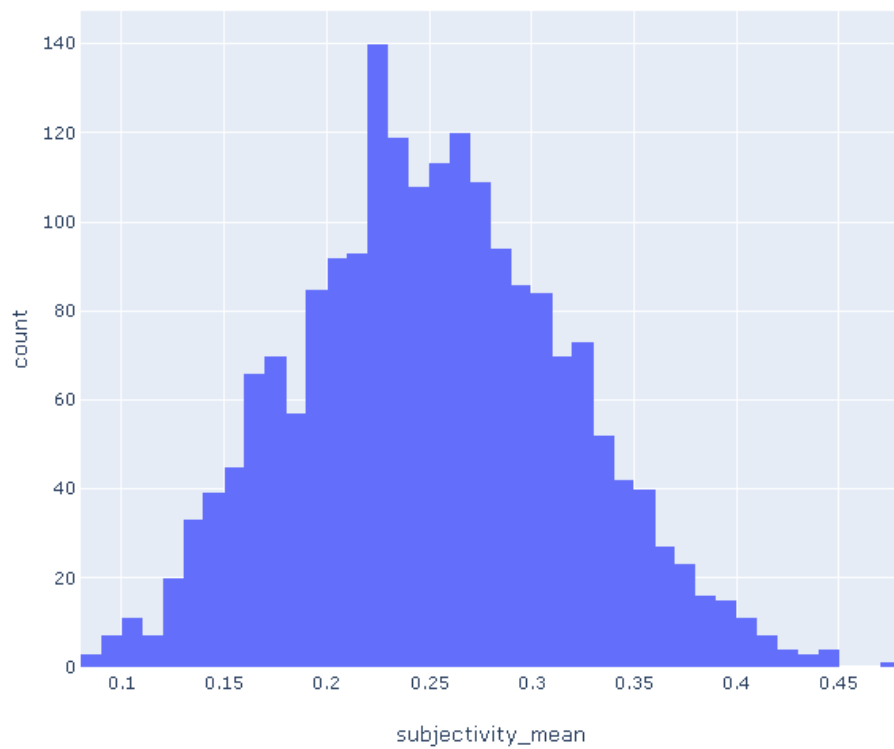


Рисунок 4.17. Распределение субъективной оценки

## 4.5 Выбор модели

В настоящее время исследования в области прогнозирования продвигаются вперед. При прогнозировании линейные модели дают стабильные результаты, однако возникают случаи, когда линейная модель не может предсказать результат. В таком случае, можно использовать нелинейные модели, которые предлагают высокую точность результатов. Учитывая все преимущества и недостатки каждой модели, были выбраны три основные модели, которые будут прогнозироваться наши данные. Модели: гребневая (ридж) регрессия, случайный лес и экстремальный градиентный бустинг. Ниже расскажем углубленно о каждой выбранной модели.

Гребневая (ридж) регрессия - это усовершенствованный метод регрессии, который специально используется для решения задач коллинеарности. Как говорили ранее, это улучшенная модель регрессии методом наименьших квадратов. С помощью регуляризации L2 модель регрессии считается практичным методом за счет потери некоторой информации и снижения точности. Модель используется во многих исследованиях, поскольку его результаты имеют более практическое значение и могут лучше влиять на относительно долгосрочные прогнозы.

Случайный лес – интегрированный метод, основанный на анализе дерева решений. Он синтезирует результаты прогнозирования каждого дерева решений для достижения окончательного прогноза и в основном используется для задач классификации: однако из-за появления дерева CART непрерывные функции могут быть разумно дискретизированы, что позволяет решать проблемы регрессии. В то же время эксперименты показывают, что он обладает сильной способностью к подгонке для сильно нелинейных задач; поскольку он в определенной степени опирается на правила, он обладает сильной способностью распознавать кусочно-нелинейные характеристики, поэтому он стал важным методом исследования для решения многих задач прогнозирования.

Бустинг — это техника построения ансамблей, в которой предсказатели построены не независимо, а последовательно. Это техника использует идею о том, что следующая модель будет учиться на ошибках предыдущей.

Экстремальный градиентный бустинг (XGBoost) – техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений.

В таблице 4.3 сравним между собой наши выбранные модели.

Модель	Преимущества	Недостатки
Ридж-регрессия	Высокая интерпретация, высокая стабильность, быстрый расчет, решение задач мультиколлинеарности.	Линейная модель, смещенная оценка, невозможно проанализировать сложные отношения, прогнозируемая производительность ограничена, легко производить переоснащение.
Случайный лес	Высокая нелинейность, высокая стабильность, высокая интерпретация, высокая точность.	Количество гиперпараметров велико, Требования к объему данных, Простота переоснащения, Длительное время обучения.
Бустинг	Нелинейный, структура модели гибкая, сильная способность добывать	Требования к объему данных, Простота



	отношения, меньше ограничений, высокая точность прогнозирования	переопределения, Низкая интерпретируемость
--	--	--

Таблица 4.3. Сравнение моделей прогнозирования.

На основании приведенного выше анализа можно обнаружить, что система моделей прогнозирования текущих цен на нефть является относительно полной и может использоваться в качестве эталонной модели для прогнозирования цен на нефть. Далее мы рассмотрим включение фактора настроения VADER, упомянутого в разделе выше, в модель прогнозирования, чтобы увидеть, значительно ли увеличена точность прогнозирования. В частности, цены на нефть и факторы ориентации на настроение вводятся в модель прогнозирования цен на нефть путем построения характеристик. Как правило, для любой из приведенных выше моделей прогнозирования  $f$  проблема прогнозирования международных цен на нефть заключается в следующем.

$$\hat{y} = f(x), \quad (5)$$

где  $\hat{y}$  предсказанной значение цен на нефть, а  $f(x)$  функция, необходимая для прогнозирования. Для общих временных рядов,  $x$  обычно историческая информация. Следовательно, форма прогнозирования может быть выражена следующим образом для временных рядов.

$$\widehat{x_{t+1}} = f(x_t, x_{t-1}, \dots, x_{t-i}), \quad (6)$$

где  $t$  - произвольный момент времени, а  $i$  - порядок запаздывания. Унифицированный, порядок отставания определяется следующей формой.

$$X_{t,i} = (x_t, x_{t-1}, \dots, x_{t-i}), \quad (7)$$

Итак, можем переписать уравнение (5) как следующее уравнение:

$$\widehat{x_{t+1}} = f(X_{t,i}), \quad (8)$$

Теперь настроение веб-текста вводится в модель  $f$  прогнозирования для обогащения информации прогнозирования, и получается следующая новая форма.

$$\widehat{x_{t+1}} = f(X_{t,i}, compound_{t,i}, negative_{t,i}, neutral_{t,i}, positive_{t,i}) \quad (9)$$

## 5 АНАЛИЗ И РЕЗУЛЬТАТЫ

### 5.1 Выбор модели прогнозирования цен на нефть

Существует много моделей прогнозирования цен на нефть, которые могут извлекать различную информацию из цен на нефть с разных точек зрения. Прежде чем приступить к анализу взаимосвязей, мы выбираем модель, которая может лучше объяснить взаимосвязь между ценами на нефть и текстовыми настроениями, оценивая ее путем прогнозирования эффективности. В соответствии с введением в разделе 4.5 Модели обучения, мы выбираем Ridge, RF и XGBoost для тестирования. Поскольку в каждом алгоритме есть гиперпараметры, ручные настройки неизбежны. После более чем 2000 попыток лучшие результаты отбираются для сравнения и анализа. Что касается текстовых особенностей, мы выбираем `compound_t`, который выражает всестороннее настроение статьи как особенность текстовых настроений.

Как видно из рисунка 5.1, эти алгоритмы обладают высокой точностью, имеют высокую степень соответствия между ценами на нефть и обеспечивают хорошую надежность.

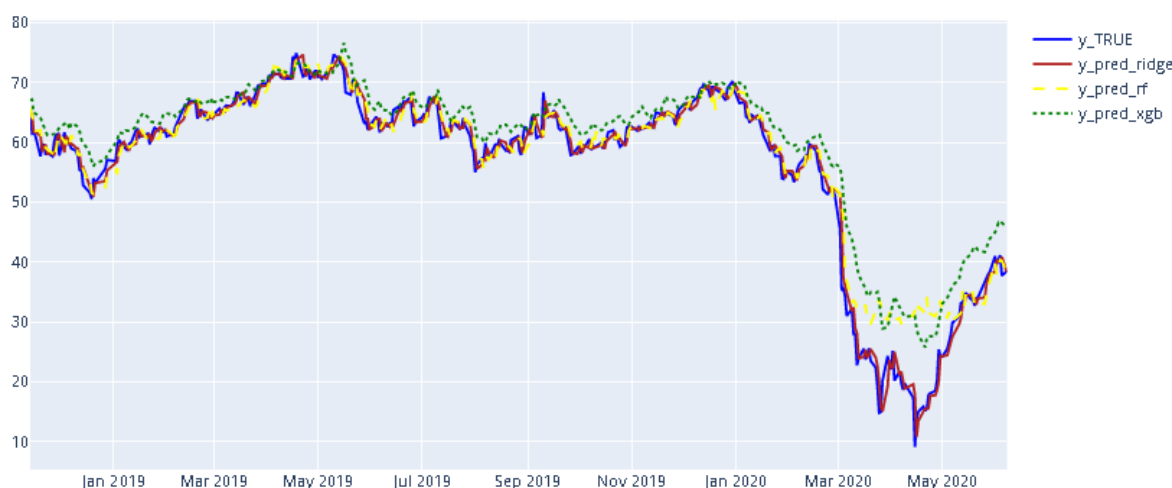


Рисунок 5.1. Прогноз цен на нефть.

Чтобы сравнить результаты этих алгоритмов, ошибку измеряют как RMSE (среднеквадратичная ошибка), MAPE (средняя абсолютная ошибка в процентах), и таким образом оценивается точность. EV (дисперсия ошибки)

используется для измерения стабильности предсказанных результатов [24] . Три статистические величины определены в уравнениях (10) – (12)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - f_i)^2}, \quad (10)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - f_i|}{y_i} * 100 \quad (11)$$

$$EV = \frac{\sum_{i=1}^N (e_i - \bar{e})^2}{N} \quad (12)$$

где N количество образцов,  $Y_i$  реальная цена на нефть,  $f_i$  является прогнозируемой ценой на нефть,  $e_i$  разница между реальными и прогнозируемыми ценами на нефть и  $\bar{e}$  это среднее  $e_i$  для всех образцов.

Модель	MSE	SD	MAPE	RMSE
Ридж	1,559522	0,482421	1,193895	1,401880
Случайный лес	52,707211	87,710394	2,381990	7,259973
XGBoost	39,261970	9,49166	3,779190	6,265937

Таблица 5.1. Прогноз ошибки статистических характеристик

Таблица 5.1 показывает сравнение нескольких алгоритмов на RMSE, MAPE. Исходя из числового значения, можно обнаружить, что погрешность и дисперсия погрешности RF и XGBoost относительно велики, и это не обеспечивает хорошую эффективность прогнозирования. У Ridge есть определенное преимущество: его RMSE может быть ниже 1,19, показывая более высокую точность, в то время как более низкое SD указывает на более высокую стабильность в прогнозировании.

Кроме того, из соотношения между характером модели и прогнозирующей эффективностью видно, что взаимосвязь между ценой на нефть и информацией о тенденциях в веб-тексте является квазилинейной: модели с высокой степенью нелинейности, RF, не предлагают хорошие прогнозы, в то время как модифицированные линейные модели Ridge лучше.

Гребневая регрессия оснащена гибкой веб-формой и прогнозы цен на нее превосходны. Поэтому последующий анализ взаимосвязи между веб-информацией и ценами на нефть выполняется с использованием Ridge в качестве прогностической модели.

## **5.2.Эффект всестороннего текстового настроения**

В этом разделе анализируется оценка текста *comround* на прогноз цен на нефть.

Хорошо известно, что новости чувствительны ко времени, а познание людьми событий также чувствительно ко времени. Требуется время, чтобы переварить отчет о его влиянии на цены на нефть. После усвоения информация не будет иметь очевидного долгосрочного влияния, если не будет времени для созревания, поэтому здесь необходимо учитывать прогноз цены на нефть. Лучше использовать новостные настроения предыдущих дней. В таком временном ряду необходимо знать, сколько шагов задержки являются оптимальными: здесь первый шаг задерживается, что указывает на настроение использовать текст со вчерашнего дня, второй порядок представляет настроение на использование текста со вчерашнего дня и позавчера вчера и так далее. Здесь RMSE выбирается в качестве индикатора для измерения точности, и различные лаги веб-информации имеют тенденцию поддерживать эффективность прогнозирования цен на нефть.

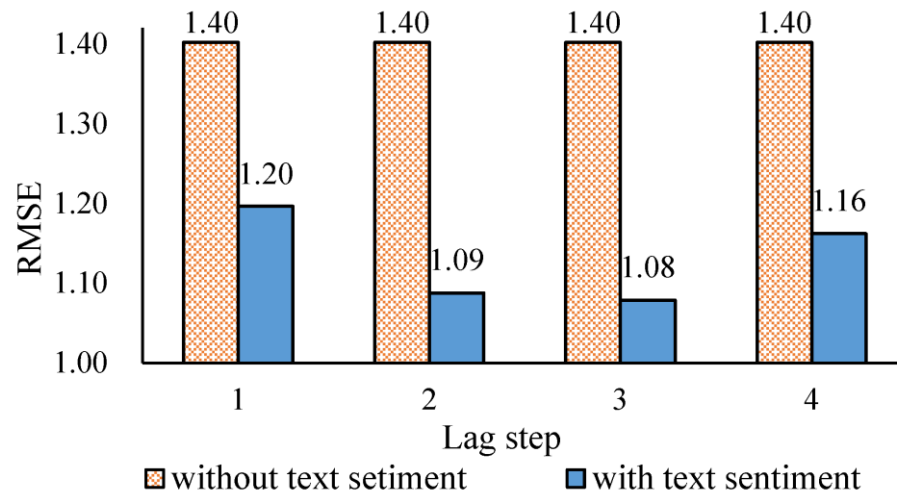


Рисунок 5.2. Сравнение RMSE без текстовых настроений и с различными шагами задержки.

Первое сравнение - RMSE: согласно рисунку 5.2, когда текст не используется, независимо от порядка задержки веб-информации, RMSE составляет 1,40. Напротив, как только веб-информация имеет тенденцию к использованию, среднее квадратическое отклонение значительно уменьшается с падением по меньшей мере на 0,2. В разных порядках ошибка предсказания также имеет определенную разницу. После третьего порядка он достигает самого низкого уровня и может упасть до 1,08. В четвертом порядке точность уменьшится, а СКО увеличится на 0,08 по сравнению с третьим порядком. Причина в том, что информация перегружена, и информация, полученная четыре дня назад, будет мешать прогнозу цен на нефть.

Таким образом, после использования настроения веб-текста точность и стабильность прогнозов цен на нефть могут быть дополнительно улучшены. Среднее квадратическое отклонение может быть уменьшено на 0,4. Используя различную текстовую информацию уровня запаздывания, точность будет отличаться. Преимущество состоит в том, что использование текстового настроения лага третьего порядка для прогнозирования цен может максимизировать точность прогнозирования; тем не менее, корректировка

различных шагов задержки текстового настроения не может привести к дальнейшим изменениям в стабильности прогноза цен на нефть

## ЗАКЛЮЧЕНИЕ

Поскольку рынок нефти очень чувствителен к нерыночным факторам, большинство из которых трудно определить количественно и всесторонне рассмотреть, что затрудняет прогнозирование и расчет. Развитие различных технологий, таких как технология обработки естественного языка, методы интеллектуального анализа текста и технологии больших данных, позволяет маркетинговым исследованиям получать и извлекать информацию из Интернета. Следовательно, внедрение этих концепций может повысить эффективность прогнозирования цен на нефть. Основываясь на рынке нефти, мы оцениваем текущее состояние исследований рынка нефти, используя соответствующие методы, и разрабатываем гибридную модель прогнозирования цен на нефть на основе текстового майнинга. С точки зрения прогнозирования цен на нефть, мы исследуем связь между веб-текстом и ценой на нефть. Мы анализируем влияние внедрения веб-текста в прогнозирование цен на нефть, влияние текстов различных типов склонности на прогнозирование стоимости нефти и влияние текстовых настроений с различными преимуществами на эффективность прогнозирования. Благодаря этим отношениям текстовая информация в Интернете может быть лучше использована в исследованиях по прогнозированию цен на нефть. На основании вышеупомянутых исследований можно сделать следующие выводы:

Взаимосвязь между ценами на нефть и настроениями в веб-тексте квазилинейна. Использование высоко нелинейных прогностических моделей, таких как RF и XGBoost, не дают хороших результатов. Лучшие результаты, получены с помощью регрессии Ridge. Гребневая (ридж) регрессия работают лучше всего благодаря их врожденной гибкости в сети.

После добавления текстовых настроений в модель прогнозирования цен на нефть, модель работает немного лучше. RMSE уменьшается примерно на 0,2, что указывает на небольшое улучшение точности и стабильности.



Таким образом, веб-текстовая информация дает преимущества при прогнозировании цен на нефть, но, учитывая взаимосвязь между веб-информацией и ценами на нефть, необходимо быть более осторожным при прогнозировании. Корректирующий эффект веб-информации появляется только тогда, когда текстовые настроения достаточно сильны. Поэтому важно определить силу настроений и использовать текстовые настроения, когда они хорошо работают. Это дает важный опыт для лучшего использования оперативной текстовой информации для прогнозирования цен на нефть в будущем.

## ССЫЛКИ

- [1] Zhao L. T. et al. Oil price risk evaluation using a novel hybrid model based on time-varying long memory //Energy Economics. – 2019. – Т. 81. – С. 70-78.
- [2] Miao H. et al. Influential factors in crude oil price forecasting //Energy Economics. – 2017. – Т. 68. – С. 77-88.
- [3] Wang M. et al. A novel hybrid method of forecasting crude oil prices using complex network science and artificial intelligence algorithms //Applied Energy. – 2018. – Т. 220. – С. 480-495.
- [4] Liu L. et al. A social-media-based approach to predicting stock comovement //Expert Systems with Applications. – 2015. – Т. 42. – №. 8. – С. 3893-3901.
- [5] Tetlock P. C. Giving content to investor sentiment: The role of media in the stock market //The Journal of finance. – 2007. – Т. 62. – №. 3. – С. 1139-1168.
- [6] Wex F. et al. Early warning of impending oil crises using the predictive power of online news stories //2013 46th Hawaii International Conference on System Sciences. – IEEE, 2013. – С. 1512-1521.
- [7] Окладников Д. Е. Ценовая политика фирмы //Маркетинг и маркетинговые исследования. – 2006. – Т. 4. – С. 344-352.
- [8] Бобылев Ю. Н. и др. Факторы формирования цен на нефть //М.: Институт экономики переходного периода. – 2006. – С. 3-33.
- [9] Поливанов А. Цены на нефть и ВВП: велика ли зависимость //Ведомости. – 2014. – Т. 7.
- [10] Джапарова Р. Всемирный банк ждет снижение цены на нефть в 2020 году.
- [11] Емельянова В. Р., Наганов А. С. Математическое моделирование экономических и социальных процессов //Математические модели техники, технологий и экономики: материа. – С. 73.

- [12] Wright P. Knowledge discovery in databases: tools and techniques //XRDS: Crossroads, The ACM Magazine for Students. – 1998. – Т. 5. – №. 2. – С. 23-26.
- [13] Fayyad U., Piatetsky-Shapiro G., Smyth P. The KDD process for extracting useful knowledge from volumes of data //Communications of the ACM. – 1996. – Т. 39. – №. 11. – С. 27-34.
- [14] Kroeze J. H., Matthee M. C., Bothma T. J. D. Differentiating between data-mining and text-mining terminology //SA Journal of Information Management. – 2004. – Т. 6. – №. 4.
- [15] Батура Т. В. Методы автоматической классификации текстов //Программные продукты и системы. – 2017. – Т. 30. – №. 1.
- [16] Jones K. S. A statistical interpretation of term specificity and its application in retrieval //Journal of documentation. – 1972.
- [17] Yu L., Wang S., Lai K. K. A rough-set-refined text mining approach for crude oil market tendency forecasting //International journal of knowledge and systems sciences. – 2005. – Т. 2. – №. 1. – С. 33-46.
- [18] Wang S., Yu L., Lai K. K. A novel hybrid AI system framework for crude oil price forecasting //Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management. – Springer, Berlin, Heidelberg, 2004. – С. 233-242.
- [19] Андерсон Т. Статистический анализ временных рядов. – Мир, 1976. – С. 757.
- [20] Кендэл М. Временные ряды: Пер. с англ. – Финансы и статистика, 1981.
- [21] Hutto C. J., Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text //Eighth international AAAI conference on weblogs and social media. – 2014.

- [22] Song K. et al. Build emotion lexicon from microblogs by combining effects of seed words and emoticons in a heterogeneous graph //Proceedings of the 26th ACM conference on hypertext & social media. – 2015. – С. 283-292.
- [23] Tang L., Wu Y., Yu L. A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting //Applied Soft Computing. – 2018. – Т. 70. – С. 1097-1108.
- [24] Zhao L. T. et al. Forecasting short-term oil price with a generalised pattern matching model based on empirical genetic algorithm //Computational Economics. – 2018. – С. 1-19.
- [25] Brent // <https://ru.wikipedia.org/wiki/Brent>
- [26] Емельянова В. Р., Наганов А. С. Математическое моделирование экономических и социальных процессов //Математические модели техники, технологий и экономики: материа. – С. 73.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Zhao, L.T., Liu, K., Duan, X.L., Li, M.F. “Oil Price Risk Evaluation Using a Novel Hybrid Model Based on Time-varying Long Memory”, 2019 // Электронная версия на сайте  
<https://www.sciencedirect.com/science/article/pii/S0140988319300982>
2. Hong, M., Ramchander, S., Wang, T., Yang, D. “Influential Factors in Crude Oil Price Forecasting”, 2017 // Электронная версия на сайте  
[https://www.researchgate.net/publication/320010166\\_Influential\\_Factors\\_in\\_Crude\\_Oil\\_Price\\_Forecasting](https://www.researchgate.net/publication/320010166_Influential_Factors_in_Crude_Oil_Price_Forecasting)
3. Wang, M., Zhao, L., Du, R., Wang, C., Chen, L., Tian, L. “A novel hybrid method of forecasting crude oil prices using complex network science and artificial intelligence algorithms”, 2019 // Электронная версия на сайте  
[https://www.researchgate.net/publication/335488284\\_A\\_hybrid\\_model\\_of\\_dynamic\\_time\\_wrapping\\_and\\_hidden\\_Markov\\_model\\_for\\_forecasting\\_and\\_trading\\_in\\_crude\\_oil\\_market](https://www.researchgate.net/publication/335488284_A_hybrid_model_of_dynamic_time_wrapping_and_hidden_Markov_model_for_forecasting_and_trading_in_crude_oil_market)
4. Liu, L., Wu, J., Li, P., Li, Q. “A social-media-based approach to predicting stock co movement”, 2014 // Электронная версия на сайте  
<https://www.sciencedirect.com/science/article/abs/pii/S0957417414008288?via%3Dihub1>
5. Tetlock, P.C. “Giving content to investor sentiment: The role of media in the stock market”, 2007// Электронная версия на сайте  
[https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3097/Tetlock\\_Media\\_Sentiment\\_JF.pdf](https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/3097/Tetlock_Media_Sentiment_JF.pdf)
6. Wex, F., Widder, N., Liebmann, M., Neumann, D. “Early warning of impending oil crises using the predictive power of online news stories”, 2013//  
<https://www.sciencedirect.com/science/article/pii/S0169207018301110>

## ПРИЛОЖЕНИЕ

```
import time
start = time.time()
import os

import numpy as np # Linear Algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

from sklearn.model_selection import cross_val_score
from sklearn.feature_selection import VarianceThreshold
from sklearn.preprocessing import StandardScaler
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from textblob import TextBlob

#plotting
import plotly.express as px
import plotly.graph_objects as go
import seaborn as sns
import matplotlib.pyplot as plt

#statistics & econometrics
import statsmodels.tsa.api as smt
import statsmodels.api as sm

#model fitting and selection
from sklearn.metrics import mean_squared_error
from sklearn.metrics import make_scorer
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import TimeSeriesSplit
from sklearn.linear_model import Lasso, Ridge
from sklearn.ensemble import RandomForestRegressor
from xgboost.sklearn import XGBRegressor

df = pd.read_csv("Combined_News_DJIA.csv",low_memory=False,
                parse_dates=[0])

full_stock = pd.read_csv("upload_DJIA_table.csv",low_memory=False,
                        parse_dates=[0])

#add the closing stock value to the df - this will be the y variable
df["Close"]=full_stock.Close

#drop the Label column
df = df.drop(["Label"], axis=1)

#check for NAN
df.isnull().sum()
df = df.replace(np.nan, ' ', regex=True)

#sanity check
df.isnull().sum().sum()

#Sentiment and subjectivity score extraction

Anakin = SentimentIntensityAnalyzer()

Anakin.polarity_scores(" ")

def detect_subjectivity(text):
    return TextBlob(text).sentiment.subjectivity

detect_subjectivity(" ") #should return 0

start_vect=time.time()
print("ANAKIN: 'Intializing the process..")

#get the name of the headline columns
cols = []
for i in range(1,26):
    col = ("Top{}".format(i))
    cols.append(col)

for col in cols:
    df[col] = df[col].astype(str) # Make sure data is treated as a string
    df[col+'_comp'] = df[col].apply(lambda x:Anakin.polarity_scores(x)['compound'])
    df[col+'_sub'] = df[col].apply(detect_subjectivity)
    print("{} Done".format(col))

print("VADER: Vaderization completed after %0.2f Minutes"%((time.time() - start_vect)/60))
```

```

for col in cols:
    comp_col = col + "_comp"
    comp_cols.append(comp_col)

w = np.arange(1,26,1).tolist()
w.reverse()

weighted_comp = []
max_comp = []
min_comp = []
for i in range(0,len(df)):
    a = df.loc[i,comp_cols].tolist()
    weighted_comp.append(np.average(a, weights=w))
    max_comp.append(max(a))
    min_comp.append(min(a))

df['compound_mean'] = weighted_comp
df['compound_max'] = max_comp
df['compound_min'] = min_comp

sub_cols = []
for col in cols:
    sub_col = col + "_sub"
    sub_cols.append(sub_col)

weighted_sub = []
max_sub = []
min_sub = []
for i in range(0,len(df)):
    a = df.loc[i,sub_cols].tolist()
    weighted_sub.append(np.average(a, weights=w))
    max_sub.append(max(a))
    min_sub.append(min(a))

df['subjectivity_mean'] = weighted_sub
df['subjectivity_max'] = max_sub
df['subjectivity_min'] = min_sub

to_drop = sub_cols+comp_cols
df = df.drop(to_drop, axis=1)

```

#### #Explorative Data Analysis

```

fig1 = go.Figure()
fig1.add_trace(go.Scatter(x=df.Date, y=df.Close,
                        mode='lines'))

title = []
title.append(dict(xref='paper', yref='paper', x=0.0, y=1.05,
                 xanchor='left', yanchor='bottom',
                 text='Development of stock values from Aug, 2008 to Jun, 2016',
                 font=dict(family='Arial',
                           size=30,
                           color='rgb(37,37,37)'),
                 showarrow=False))

fig1.update_layout(xaxis_title='Date',
                  yaxis_title='Closing stock value (in $)',
                  annotations=title)

fig1.show()

```

#### #function for quick plotting and testing of stationarity

```

def stationary_plot(y, lags=None, figsize=(12, 7), style='bmh'):
    """
    Plot time series, its ACF and PACF, calculate Dickey-Fuller test

    y - timeseries
    lags - how many lags to include in ACF, PACF calculation
    """
    if not isinstance(y, pd.Series):
        y = pd.Series(y)

    with plt.style.context(style):
        fig = plt.figure(figsize=figsize)
        layout = (2, 2)
        ts_ax = plt.subplot2grid(layout, (0, 0), colspan=2)
        acf_ax = plt.subplot2grid(layout, (1, 0))
        pacf_ax = plt.subplot2grid(layout, (1, 1))

        y.plot(ax=ts_ax)
        p_value = sm.tsa.stattools.adfuller(y)[1]
        ts_ax.set_title('Time Series Analysis Plots\n Dickey-Fuller: p={0:5f}'.format(p_value))
        smt.graphics.plot_acf(y, lags=lags, ax=acf_ax)
        smt.graphics.plot_pacf(y, lags=lags, ax=pacf_ax)
        plt.tight_layout()

```

```

#Model training
#3 ML models: Ridge, Random forest, XGBoost

def ts_train_test_split(X, y, test_size):
    """
    Perform train-test split with respect to time series structure
    """

    # get the index after which test set starts
    test_index = int(len(X)*(1-test_size))

    X_train = X.iloc[:test_index]
    y_train = y.iloc[:test_index]
    X_test = X.iloc[test_index:]
    y_test = y.iloc[test_index:]

    return X_train, X_test, y_train, y_test

```

```

X = lag_df.drop(['Close'],axis=1)
X.index = X["Date"]
X = X.drop(['Date'],axis=1)
y = lag_df.Close

X_train, X_test, y_train, y_test = ts_train_test_split(X, y, test_size = 0.2)

#sanity check
(len(X_train)+len(X_test))==len(X)

```

```

# ridge

ridge_param = {'model__alpha': list(np.arange(0.001,1,0.001))}
ridge = Ridge(max_iter=5000)
pipe = Pipeline([
    ('scale', scaler),
    ('model', ridge)])
search_ridge = GridSearchCV(estimator=pipe,
                             param_grid = ridge_param,
                             scoring = scorer,
                             cv = tscv,
                             n_jobs=4
                             )
search_ridge.fit(X_train_e, y_train_e)

ridge_e = search_ridge.best_estimator_

#get cv results of the best model + confidence intervals
from sklearn.model_selection import cross_val_score
cv_score = cross_val_score(ridge_e, X_train_e, y_train_e, cv=tscv, scoring=scorer)
econ_perf = econ_perf.append({'Model':'Ridge', 'MSE':np.mean(cv_score), 'SD':(np.std(cv_score))}, ignore_index=True)
ridge_e

coefs = ridge_e['model'].coef_
ridge_coefs = pd.DataFrame({'Coef': coefs,
                           'Name': list(X_train_e.columns)})
ridge_coefs["abs"] = ridge_coefs.Coeff.apply(np.abs)
ridge_coefs = ridge_coefs.sort_values(by="abs", ascending=False).drop(["abs"], axis=1)
ridge_coefs

```

```

# random forest

rf = RandomForestRegressor()
pipe = Pipeline([
    ('scale', scaler),
    ('model', rf)])
gridsearch_rf = GridSearchCV(estimator=pipe,
                              param_grid = rf_param,
                              scoring = scorer,
                              cv = tscv,
                              n_jobs=4,
                              verbose=3
                              )

rf_e = gridsearch_rf.best_estimator_

#get cv results of the best model + confidence intervals
cv_score = cross_val_score(rf_e, X_train_e, y_train_e, cv=tscv, scoring=scorer)
econ_perf = econ_perf.append({'Model':'RF', 'MSE':np.mean(cv_score), 'SD':(np.std(cv_score))}, ignore_index=True)

```

```

#XGBoost

xgb_param = {'model__lambda': list(np.arange(0.1,3, 0.1)), #L2 regularisation
             'model__alpha': list(np.arange(0.1,3, 0.1)), #L1 regularisation
             }

xgb = XGBRegressor(booster='gblinear', feature_selector='shuffle', objective='reg:squarederror')

pipe = Pipeline([
    ('scale', scaler),
    ('model', xgb)])
gridsearch_xgb = GridSearchCV(estimator=pipe,
                              param_grid = xgb_param,
                              scoring = scorer,
                              cv = tscv,
                              n_jobs=4,
                              verbose=3
                              )
gridsearch_xgb.fit(X_train_e, y_train_e)

```



```

#NLP models
X_train_n = X_train.drop(econ_cols, axis=1)
X_test_n = X_test.drop(econ_cols, axis=1)
y_train_n = y_train
y_test_n = y_test

nlp_perf = pd.DataFrame(columns=['Model', 'MSE', 'SD'])
nlp_perf

```

```

##ridge

ridge_param = {'model__alpha': list(np.arange(1,10,0.1))}
ridge = Ridge(max_iter=5000)
pipe = Pipeline([
    ('scale', scaler),
    ('model', ridge)
])
search_ridge = GridSearchCV(estimator=pipe,
                            param_grid = ridge_param,
                            scoring = scorer,
                            cv = tscv,
                            n_jobs=4
                            )
search_ridge.fit(X_train_n, y_train_n)

ridge_n = search_ridge.best_estimator_

#get cv results of the best model + confidence intervals
cv_score = cross_val_score(ridge_n, X_train_n, y_train_n, cv=tscv, scoring=scorer)
nlp_perf = nlp_perf.append({'Model': 'Ridge', 'MSE': np.mean(cv_score), 'SD': (np.std(cv_score))}, ignore_index=True)
ridge_n

plotcoef(ridge_n['model'], X_train_n)

coefs = ridge_n['model'].coef_
ridge_coefs = pd.DataFrame({'Coef': coefs,
                           'Name': list(X_train_n.columns)})
ridge_coefs["abs"] = ridge_coefs.Coef.apply(np.abs)
ridge_coefs = ridge_coefs.sort_values(by="abs", ascending=False).drop(["abs"], axis=1)
ridge_coefs

```

```

#Random Forest

rf_param = {'model__n_estimators': [10, 100, 300],
            'model__max_depth': [10, 20, 30, 40],
            'model__min_samples_split': [2, 5, 10],
            'model__min_samples_leaf': [1, 2, 3],
            'model__max_features': ["auto", 'sqrt']}
rf = RandomForestRegressor()
pipe = Pipeline([
    ('scale', scaler),
    ('model', rf)])
gridsearch_rf = GridSearchCV(estimator=pipe,
                             param_grid = rf_param,
                             scoring = scorer,
                             cv = tscv,
                             n_jobs=4,
                             verbose=3
                             )
gridsearch_rf.fit(X_train_n, y_train_n)

rf_n = gridsearch_rf.best_estimator_

#get cv results of the best model + confidence intervals
cv_score = cross_val_score(rf_n, X_train_n, y_train_n, cv=tscv, scoring=scorer)
nlp_perf = nlp_perf.append({'Model': 'RF', 'MSE': np.mean(cv_score), 'SD': (np.std(cv_score))}, ignore_index=True)

```

```

# XGBoost

xgb_param = {'model__lambda': list(np.arange(1,10, 1)), #L2 regularisation
            'model__alpha': list(np.arange(1,10, 1)), #L1 regularisation
            }
xgb = XGBRegressor(booster='gblinear', feature_selector='shuffle', objective='reg:squarederror')

pipe = Pipeline([
    ('scale', scaler),
    ('model', xgb)])
gridsearch_xgb = GridSearchCV(estimator=pipe,
                              param_grid = xgb_param,
                              scoring = scorer,
                              cv = tscv,
                              n_jobs=4,
                              verbose=3
                              )
gridsearch_xgb.fit(X_train_n, y_train_n)

```