Макроэкономическое предсказание с использованием данных социальных сетей

Елена Шуляк

Цели и задачи

Цель: выяснить, полезны ли данные из российской социальной сети «Вконтакте» для прогнозирования некоторых макроэкономических переменных.

Цели и задачи

Задачи:

1. Собрать посты и комментарии к ним из групп новостных СМИ «Вконтакте».
2. Обработать собранные тексты, выделить те, которые относятся к экономике.
3. Построить индексы настроений на основе полученных данных.
4. Использовать методы машинного обучения и построенные индексы и некоторые дополнительные данные из «Вконтакте» для прогнозирования переменных.
5. Сравнить полученные прогнозы с прогнозами простой модели.

Сбор данных

- СМИ: РБК, РИА Новости, Forbes, Meduza, ТАСС, Известия, Интерфакс, Коммерсант, Российская Газета, Ведомости.
- Период: с августа 2010 по декабрь 2020.
- 1.5 млн. новостных постов и 23 млн. комментария к ним.
- Дополнительно собраны лайки, репосты, количество просмотров, заголовки и т.д.

Сбор данных

Проблемы

- Новостные посты на разные темы (спорт, искусство, наука, развлечения и т.д.), не только об экономике.
- Средняя длина новостных сообщений из «ВКонтакте» - 17 слов, так как все СМИ размещают полные тексты новостей на своих сайтах, а в группах «ВКонтакте» только краткие описания и заголовки.

Подготовка данных

1. Удаление пунктуации, чисел, специальных символов.
2. Перевод символов в нижний регистр
3. Лемматизация – приведение слов к словарной (начальной) форме
4. Удаление стоп-слов и редких слов.

Original post	Prepared post
Титов предложил Путину ввести специальный налоговый режим для кафе и ресторанов. Бизнес-омбудсмен в письме Путину предложил установить специальный налоговый режим с прогрессивной шкалой для ресторанов и кафе. Он также предложил подумать о снижении или полной отмене НДС для отрасли	титов предлагать путин вводить специальный налоговый режим кафе ресторан бизнесомбудсмен письмо путин предлагать устанавливать специальный налоговый режим прогрессивный шкала ресторан кафе предлагать подумать снижение полный отмена ндс отрасль
Reuters узнал о попытке ЦБ предложить банк «Открытие» «Яндексу». ЦБ предлагал «Яндексу» рассмотреть возможность покупки банка «Открытие», но IT-компанию предложение не заинтересовало, рассказали Reuters госбанкиры. По их словам, после срыва сделки с Тиньковым IT-гигант решил развивать банковский бизнес самостоятельно	reuters узнавать попытка цб предлагать банк открытие яндекс цб предлагать яндекс рассматривать возможность покупка банк открытие itкомпанию предложение заинтересовывать reuters госбанкир слово срыв сделка тиньков itгигант решать развивать банковский бизнес самостоятельно

Метод GSDMM

GSDMM – метод для выделения тем в коротких текстах, у него есть 4 гиперпараметра:

K – первоначальное количество групп. Общее количество кластеров может быть меньше, чем *K*.

I – количество итераций.

α – отвечает за возможность текстов «перейти» в группу, которая на одной из итераций стала пустой. Если α = 0, то текст никогда не «выберет» группу, которая когда-то стала пустой.

β – отвечает за однородность групп. Чем меньше его значение, тем больше одинаковых слов должно быть в текстах группы.

Выбираем параметры со значениями α = 0.1, β = 0.1, I = 15, K = 500.

Примеры кластеров

Кластеры с новостями об экономике

```
Cluster 44 : [('рубль', 12237), ('доллар', 7935), ('курс', 7777), ('евро', 4763), ('ставка', 4386), ('рост', 4265), ('рынок', 3929), ('экономика', 3855), ('валюта', 3455), ('цена', 2946)]

Cluster 131 : [('россиянин', 9275), ('опрос', 2927), ('страна', 2715), ('вциом', 1796), ('зарплата', 1706), ('население', 1684), ('работа', 1623), ('уровень', 1566), ('число', 1556), ('доход', 1511)]

Cluster 494 : [('банк', 9800), ('цб', 4674), ('кредит', 3061), ('лицензия', 2025), ('компания', 1955), ('рынок', 1952), ('клиент', 1700), ('кредитный', 1674), ('рубль', 1523), ('россиянин', 1495)]
```

Кластеры с новостями на другие темы

```
Cluster 395 : [('мир', 3349), ('олимпийский', 3337), ('сборная', 3033), ('медаль', 2994), ('завоевывать', 2968), ('чемпионат', 2642), ('олимпиада', 2507), ('золото', 2300), ('игра', 2157), ('выигрывать', 2139)]

Cluster 208 : [('землетрясение', 3472), ('жертва', 1757), ('происходить', 1587), ('число', 1576), ('ураган', 1515), ('погибать', 1504), ('наводнение', 1239), ('сильный', 1231), ('результат', 1161), ('пожар', 1110)]

Cluster 398 : [('музей', 2954), ('выставка', 2561), ('искусство', 1687), ('картина', 1572), ('художник', 1501), ('открываться', 1215), ('галерея', 1066), ('работа', 1039), ('москва', 928), ('русский', 911)]
```

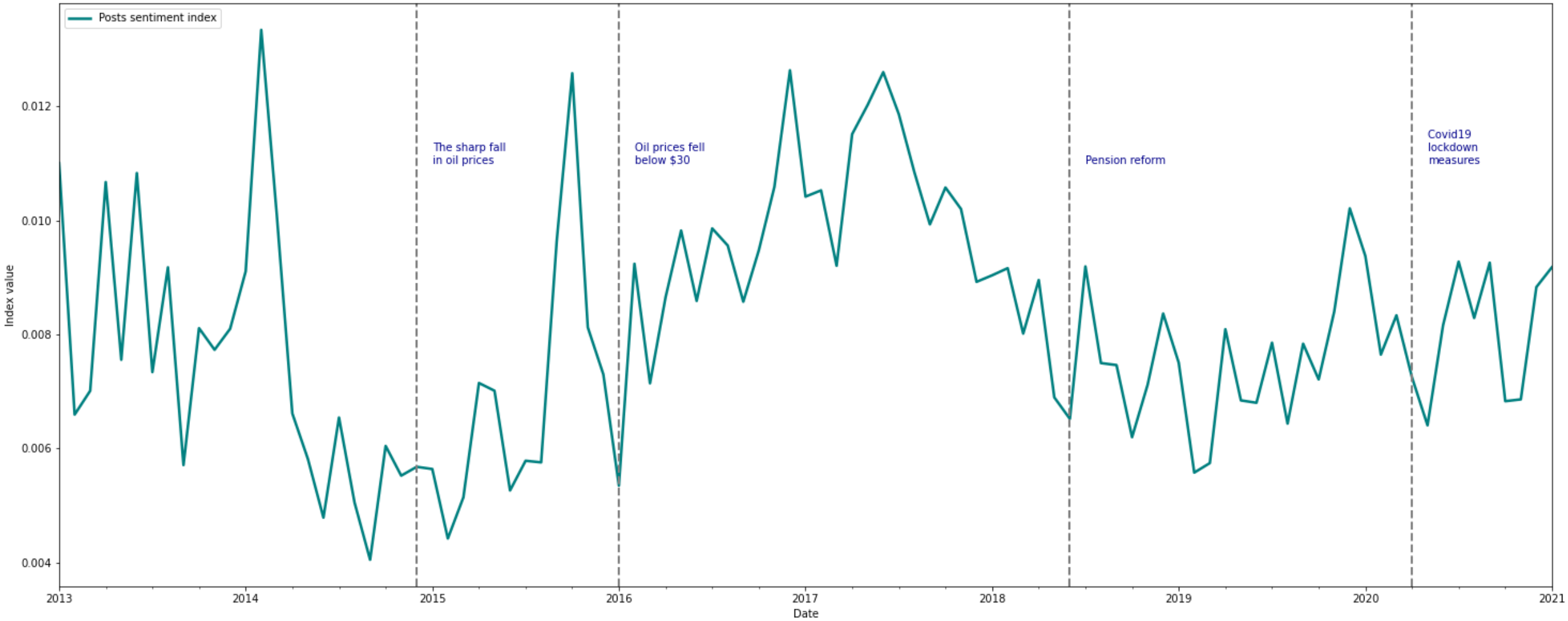
Индексы настроений

На основе собранных текстов строятся индексы настроений и используются для прогнозирования.

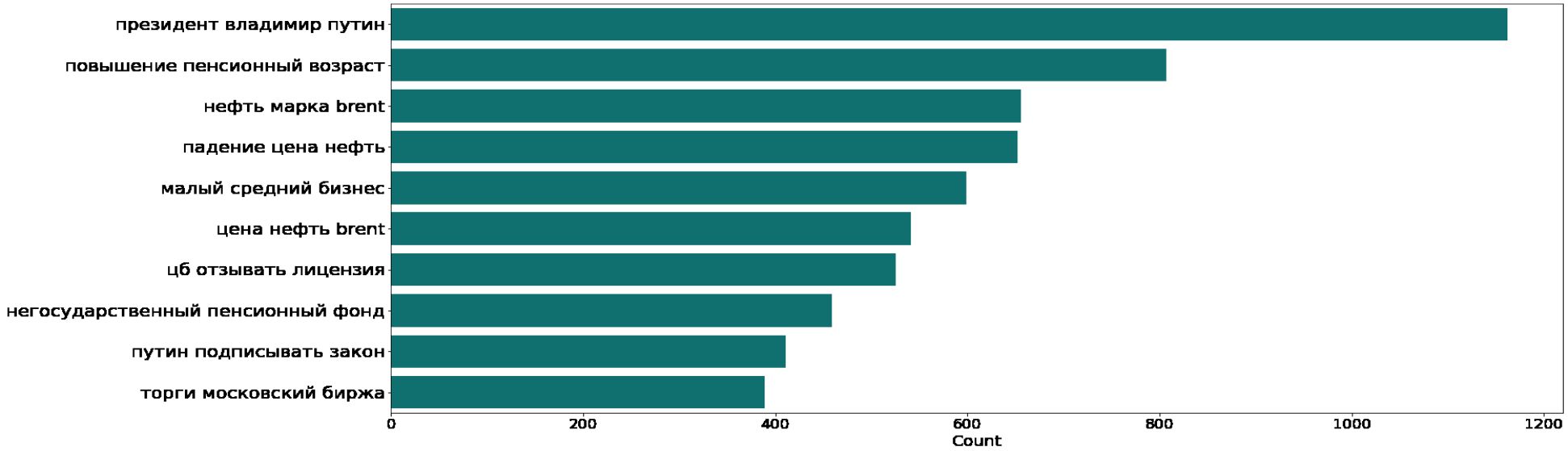
1. Индекс настроения комментариев.
2. Индекс настроения новостных постов.
3. Индекс кризисных слов.

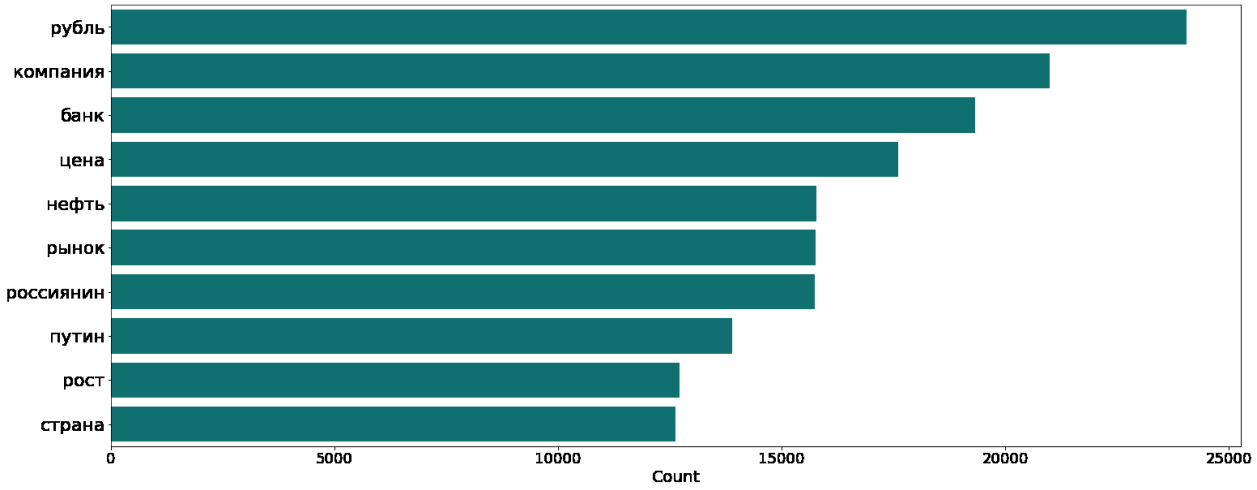
Индекс настроения комментариев

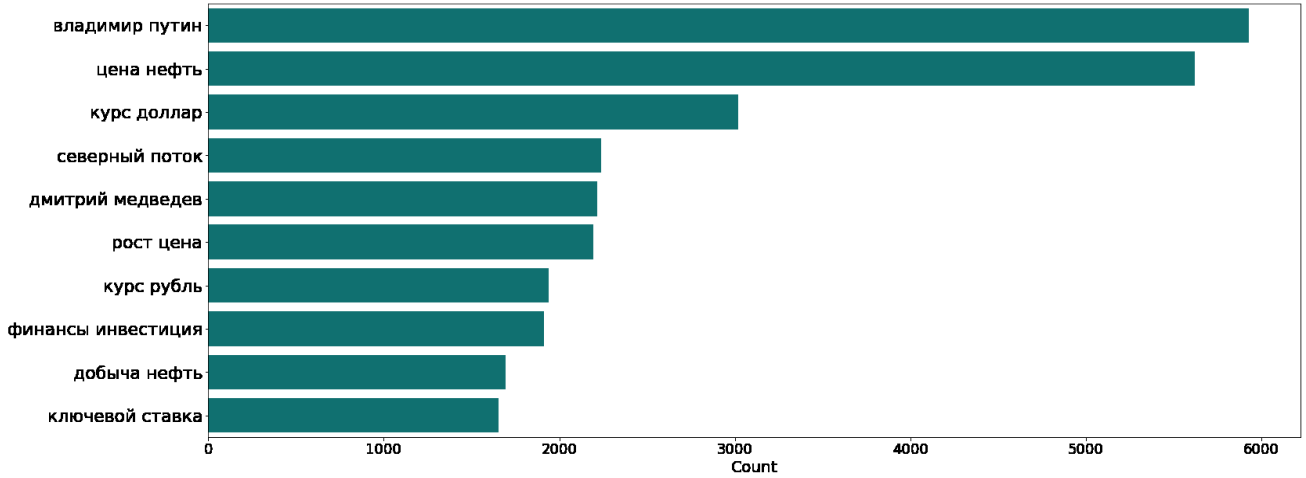
Индекс настроения постов



N-граммы

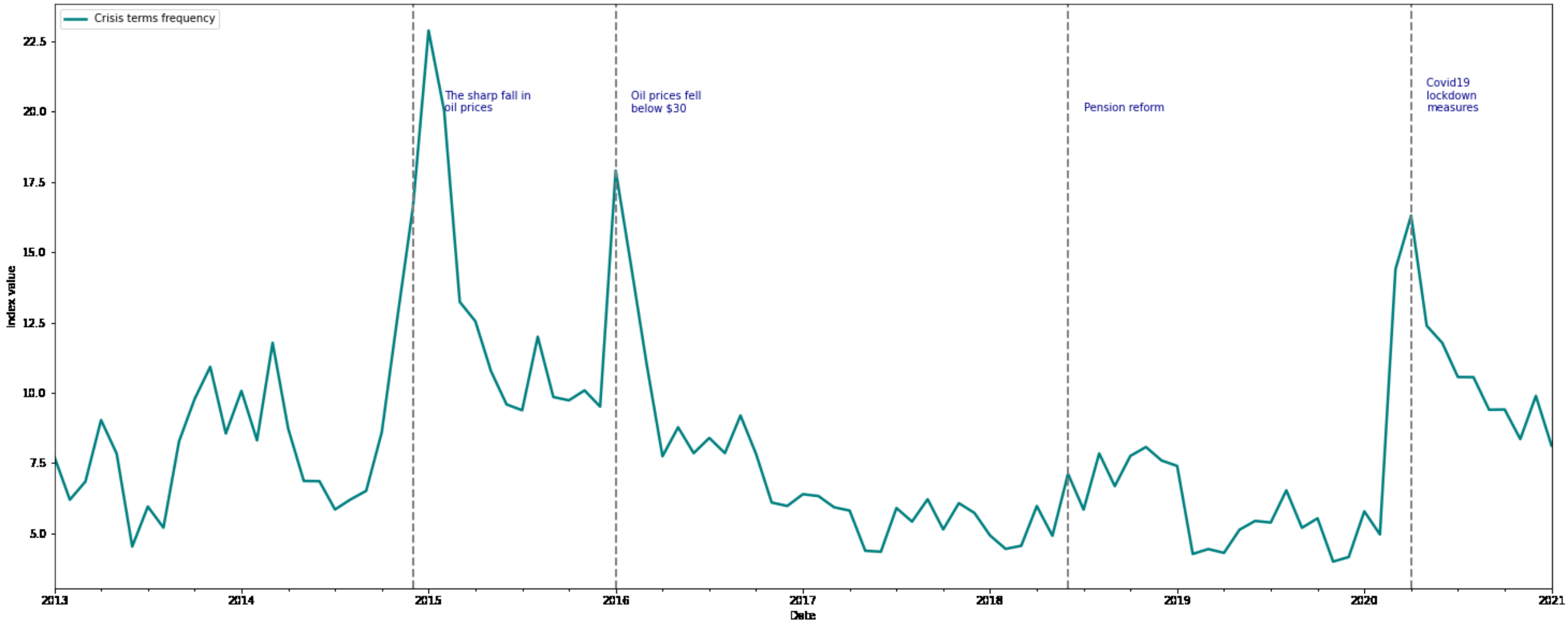

президент владимир путин
повышение пенсионный возраст
нефть марка brent
падение цена нефть
малый средний бизнес
цена нефть brent
цб отзывать лицензия
негосударственный пенсионный фонд
путин подписывать закон
торги московский биржа
0
200
400
600
800
1000
1200
Count


рубль
компания
банк
цена
нефть
рынок
россиянин
путин
рост
страна
0
5000
10000
15000
20000
25000
Count


владимир путин
цена нефть
курс доллар
северный поток
дмитрий медведев
рост цена
курс рубль
финансы инвестиция
добыча нефть
ключевой ставка
0
1000
2000
3000
4000
5000
6000
Count

Кризисные слова

Индекс кризисных слов



Переменные и признаки

1. USD/RUB, т.е. цена доллара США, выраженная в российских рублях (средняя цена закрытия за месяц).
2. EUR/RUB, т.е. цена евро, выраженная в российских рублях (средняя цена закрытия за месяц).
3. Уровень инфляции, т.е. увеличение общего уровня цен по сравнению с предыдущим месяцем.
4. Промышленное производство (месяц к месяцу), т.е. отношение текущего объема промышленного производства к объему промышленного производства в предыдущем месяце.

Признаки

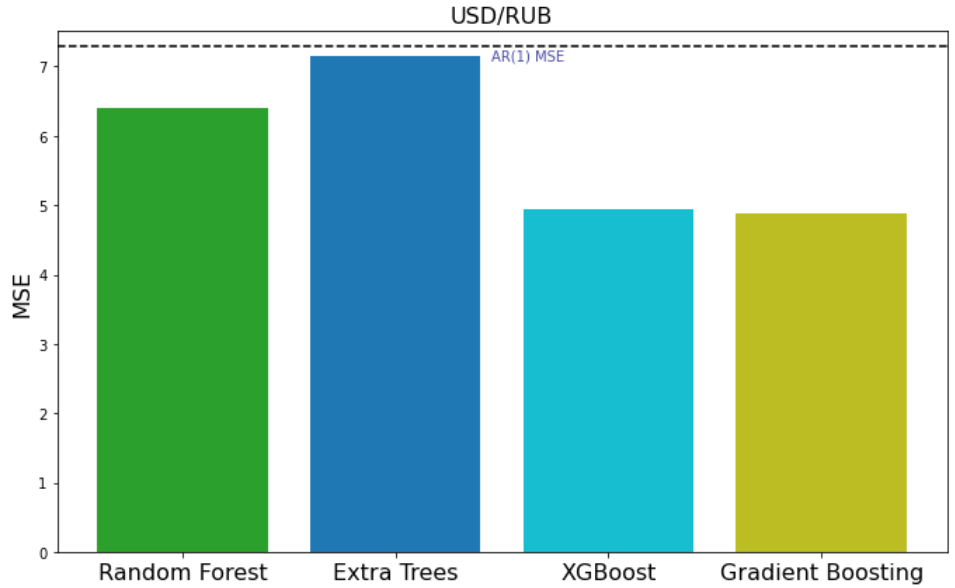
В качестве признаков используются построенные индексы, количество новостей об экономике в данном месяце, лайков, просмотров, репостов и т.д.

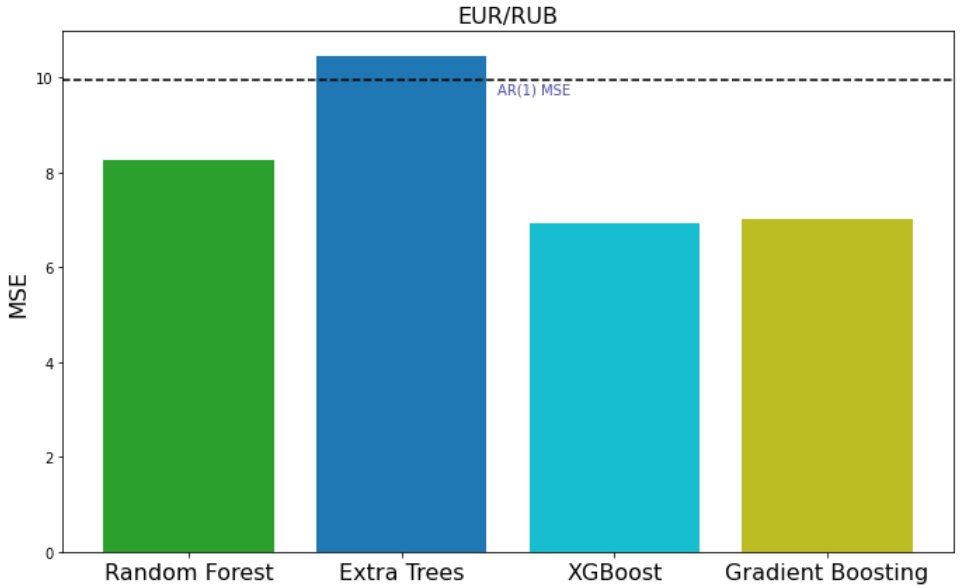
Прогнозирование

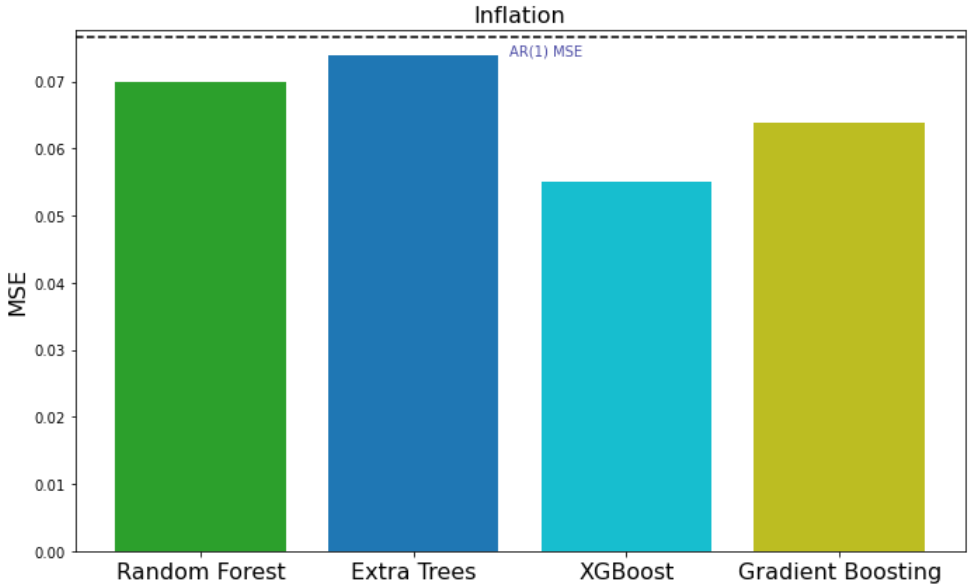
Модели: Random Forest, Extremely Randomized Trees, Gradient Boosting, XGboost.

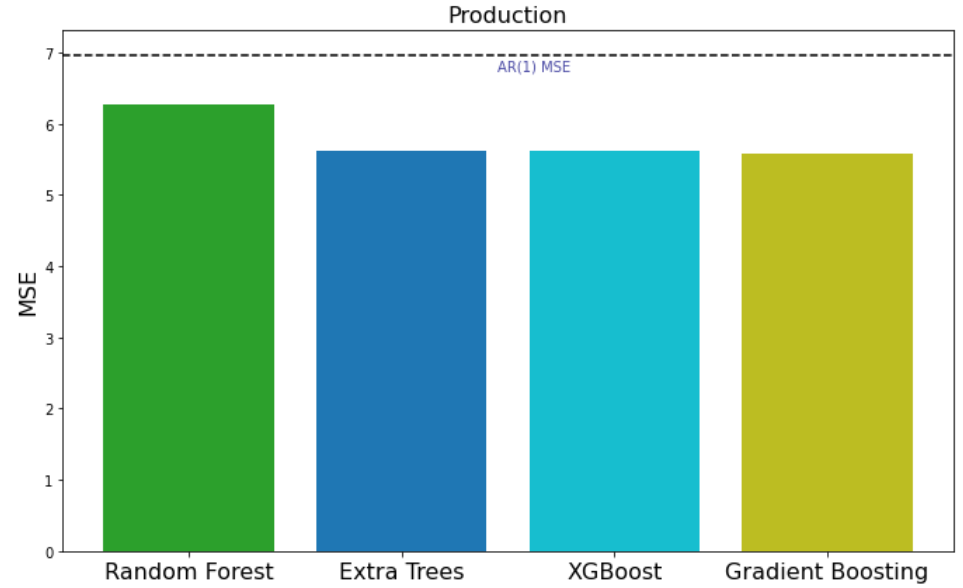
1. Разделяем временной ряд таким образом, что наблюдения из прошлого становятся обучающей выборкой, а из будущего – тестовой.
2. Модель делает прогноз на один период вперед.
3. Прогноз сравниваем с настоящим значением переменной, считаем ошибку.
4. Настоящее значение переменной добавляем к обучающей выборке, снова обучаем модель.
5. Повторяем до тех пор, пока в тестовой выборке не закончатся наблюдения.

Результаты


USD/RUB
AR(1) MSE
MSE
0
1
2
3
4
5
6
7
Random Forest
Extra Trees
XGBoost
Gradient Boosting


EUR/RUB
AR(1) MSE
MSE
0
2
4
6
8
10
Random Forest
Extra Trees
XGBoost
Gradient Boosting


Inflation
AR(1) MSE
MSE
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Random Forest
Extra Trees
XGBoost
Gradient Boosting


Production
AR(1) MSE
MSE
0
1
2
3
4
5
6
7
Random Forest
Extra Trees
XGBoost
Gradient Boosting

Заключение

1. Собраны и обработаны данные из соцсети «Вконтакте».
2. С помощью метода GSDMM выделены новости об экономике.
3. Построены индексы настроений на основе постов и комментариев.
4. Полученные индексы сравниваются с традиционными индексами.
5. Новые индексы и некоторые дополнительные данные из «Вконтакте» используются для предсказания переменных.
6. Ошибки ML-моделей меньше ошибок простой модели.

<u>Практическое значение</u>

- Полученные индексы могут использоваться в дополнение к традиционным индексам, основанным на опросах.
- Данные из «Вконтакте» легко собирать, поэтому индексы можно строить на ежедневной основе.
- Данные из социальных сетей можно использовать в макроэкономическом прогнозировании.