

# Тематическое моделирование и классификация корпуса текстов

Выполнил: Андрей Орлов

Курс ДПО по направлению  
«Компьютерная лингвистика», НИУ ВШЭ

Москва, 2023 г.

## Итоги промежуточного проекта:

- сформирован корпус текстов;
- выполнено тематическое моделирование корпуса;
- каждому тексту присвоены соответствующие теги тем, имен и названий;
- выделены ключевые слова в тексте.

Ссылка на проект

## Задачи финального проекта:

- дообучить модель BERT для классификации текстов на основе тематического моделирования;
- загрузить дополнительные тексты и применить к ним модель классификации;
- провести трендовый анализ на основе тематического моделирования.

Материал для работы – новостные публикации на сайте [SecurityLab.ru](https://securitylab.ru)

Объем корпуса: 8000 текстов

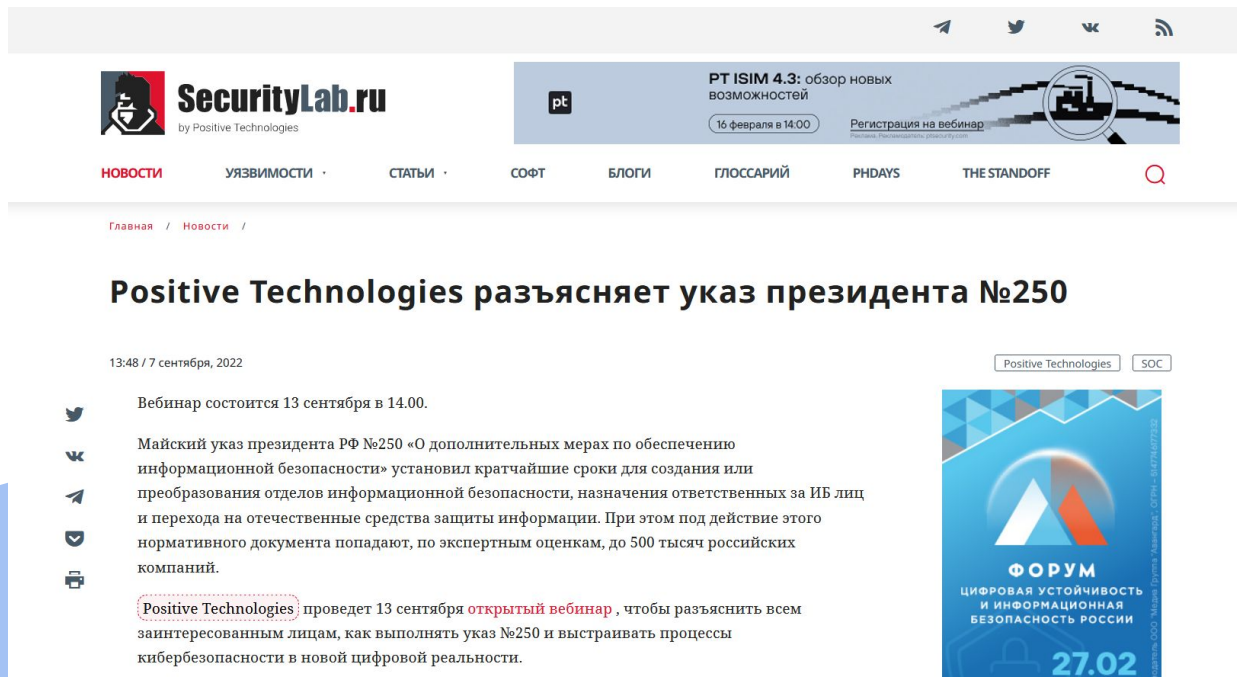
## Классификация текстов: данные

Для разметки подготовлен корпус новостных статей с сайта, посвященного информационной безопасности (<https://www.securitylab.ru/>).

Каждому тексту присвоены от одной до трех пометок с указанием содержащейся в тексте темы.

# Пример текста

## Выделен тематический тег: 'события и мероприятия'



The screenshot shows the SecurityLab.ru website. The header includes the site logo, navigation links (Новости, Уязвимости, Статьи, Софт, Блоги, Глоссарий, PHDays, The Standoff), and a search icon. A banner for 'PT ISIM 4.3' is visible. The main article is titled 'Positive Technologies разъясняет указ президента №250' and is dated '13:48 / 7 сентября, 2022'. The article text discusses the implementation of President Putin's decree №250 on information security. A sidebar on the right features a blue graphic for a forum titled 'ФОРУМ ЦИФРОВАЯ УСТОЙЧИВОСТЬ И ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ РОССИИ' with the date '27.02'.

SecurityLab.ru by Positive Technologies

PT ISIM 4.3: обзор новых возможностей  
16 февраля в 14.00  
Регистрация на вебинар

Новости · Уязвимости · Статьи · Софт · Блоги · Глоссарий · PHDays · THE STANDOFF

Главная / Новости /

### Positive Technologies разъясняет указ президента №250

13:48 / 7 сентября, 2022

Positive Technologies SOC

Вебинар состоится 13 сентября в 14.00.

Майский указ президента РФ №250 «О дополнительных мерах по обеспечению информационной безопасности» установил кратчайшие сроки для создания или преобразования отделов информационной безопасности, назначения ответственных за ИБ лиц и перехода на отечественные средства защиты информации. При этом под действие этого нормативного документа попадают, по экспертным оценкам, до 500 тысяч российских компаний.

Positive Technologies проведет 13 сентября **открытый вебинар**, чтобы разъяснить всем заинтересованным лицам, как выполнять указ №250 и выстраивать процессы кибербезопасности в новой цифровой реальности.

ФОРУМ  
ЦИФРОВАЯ УСТОЙЧИВОСТЬ  
И ИНФОРМАЦИОННАЯ  
БЕЗОПАСНОСТЬ РОССИИ  
27.02

ссылка на публикацию

## Форматирование датасета

Для удобства работы с данными датасет приведен к формату DatasetDict. Полученный объект содержит три отдельных датасета для обучения, валидации и тестирования модели

Используется библиотека  
Datasets

## Модель

- для обработки текста с помощью BERT необходимо токенизировать данные;
- для предобработки данных и для классификации используется дообученная на русскоязычных текстах модель BERT От DeepPavlov.

Используется библиотека Transformers

## Оценка полученной модели

Задан набор метрик для оценки дообученной модели и сравнения версий. В дальнейшем результаты сохраняются в репозиторий вместе с моделью.

Ссылка на репозиторий с моделью



## Оценка полученной модели на данных для тестирования

Loss: 0.24590303003787994

F1: 0.7657472210786331

ROC curve: 0.8392317829770035

Accuracy: 0.30538922155688625

Ссылка на репозиторий с моделью

## Примеры классифицированных текстов:

Исследование показало, что использование искусственного интеллекта в разработке программного обеспечения может существенно повысить производительность и инновации в мире.

'разработка',  
'программное\_обеспечение'

## Примеры классифицированных текстов:

Эксперты компании «Инфосистемы Джет» провели исследование , посвященное анализу проблем и рисков, связанных с атаками через сторонние организации, когда злоумышленники атакуют целевую компанию не напрямую, а через ее доверенных партнеров, поставщиков или подрядчиков.

'события\_и\_мероприятия',  
'исследования'

## Примеры классифицированных текстов:

Исследователи «Лаборатории Касперского» обнаружили ранее незадокументированное семейство вредоносных программ и выявили операционные ошибки, допущенные группой Andariel, фракцией северокорейской группировки Lazarus Group.

'мошенничества', 'вредоносы',  
'уязвимости'

# Статистический анализ трендов на основе тематического моделирования

выполнен общий анализ количества публикаций в корпусе на разные темы за охваченный временной период;

На основе полученных данных проведены тесты на выделение тренда.

Тест Манна-Кендалла на выделение тренда;

Тест Петтитт на выделение точки перелома тренда;

Регрессионный анализ тренда.

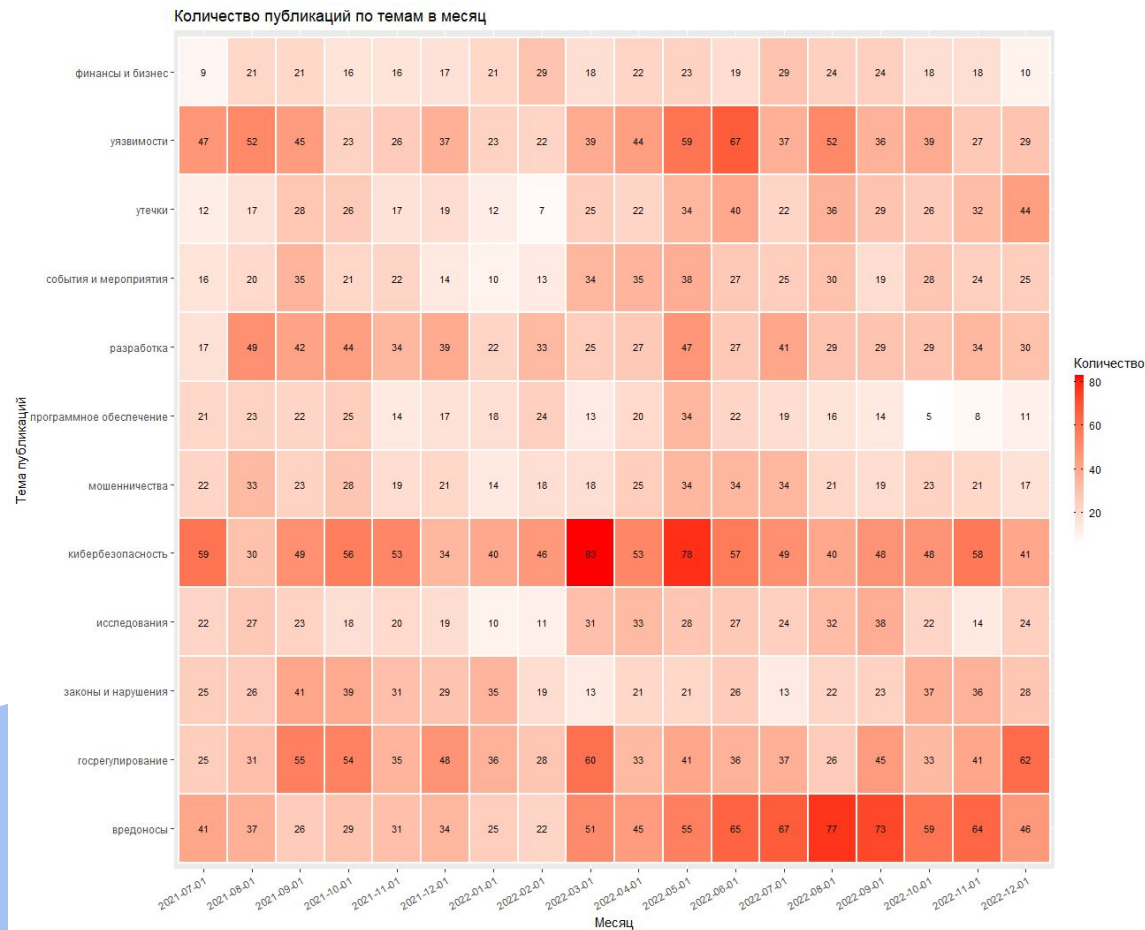


Таблица сопряженности по количеству публикаций в месяц

## Тест Манна-Кендалла:

$p\text{-value} = 0.003219$

Тест показывает высокую статистическую значимость временных изменений.

Р-значение позволяет нам отказаться от гипотезы о нормальном распределении публикаций по месяцам.

Непараметрический тест на выделение тренда в данных временного ряда

## Тест Петтитт:

p-value = 0.0098

probable change point at time K 10

Точка перелома тренда находится в 38-й позиции ряда

Непараметрический тест на  
выделение точки перелома тренда в  
данных временного ряда



## Линейная регрессия (Тема : «Утечки данных»)

Multiple R-squared: 0.4668, p-value: 0.001261

Отношение полной дисперсии к не объясненной дисперсии модели ( $R^2$ ) указывает, что модель охватывает более половины дисперсии значений

Метод линейной регрессии показывает зависимость между двумя переменными – количеством текстов на тему по месяцам

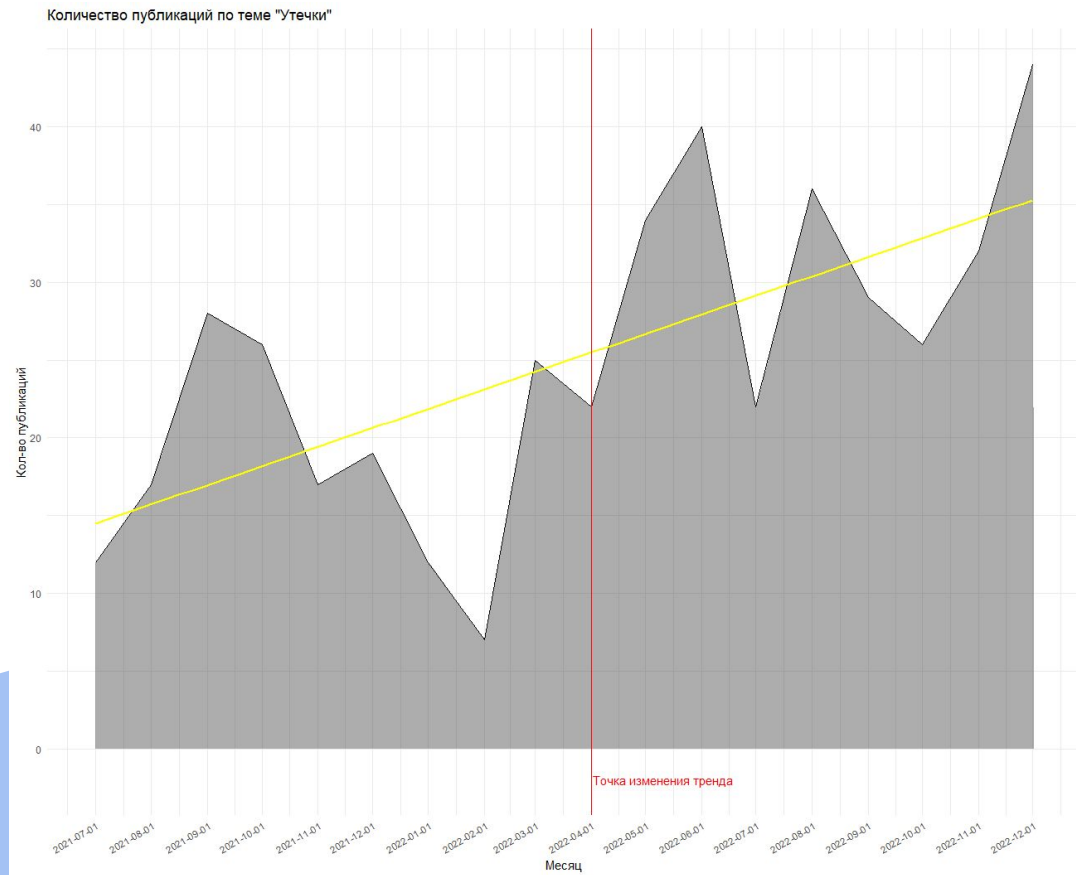


График количества публикаций по теме и выделение  
временного тренда

## Итоги проекта:

Предварительно выполненное тематическое моделирование позволяет провести разметку новых данных.

Тематическое моделирование и классификация позволяют изучать данные и делать выводы об их распределении.

Также нормализованы выделенные на промежуточном этапе ключевые слова и именованные сущности