


# Категоризация новостей в веб-приложении. Подведение итогов

## Состав команды:

Афанасьев Денис,  
Боттаева Амина,  
Гусева Софья,  
Склезнёва Ксения

# Постановка задачи

- Разработать Telegram-бот, в который пользователь загружает текст новости, а на выходе получает категорию (тему) загруженной новости
  - **Дальнейшие перспективы:**  
Telegram-бот выдает новости за указанный период по одной конкретной теме
  - **Используемые данные:** датасет, составленный из новостей с сайта Lenta.ru за 2020 год
- 
- economy: 0
  - sports: 1
  - society: 2
  - life: 3
  - entertainment: 4
  - technology: 5
  - science: 6
  - russia: 7
  - history: 8

# Этапы работы над проектом

- ✓ Сбор данных
- ✓ Поиск подходящих моделей
- ✓ Обучение моделей
- ✓ Тестирование и усовершенствование моделей
- x Создание телеграм-бота
- ✓ Развертывание проекта в Fast Api
- ✓ Тестирование полученного проекта

# Обучение моделей

## 1. Предобработка текста

- Приведение текста к нижнему регистру
- Удаление специальных символов и цифр
- Токенизация
- Удаление стоп-слов
- Приведение к нормальной форме
- Объединение токенов обратно в строку

## 2. Обученные модели

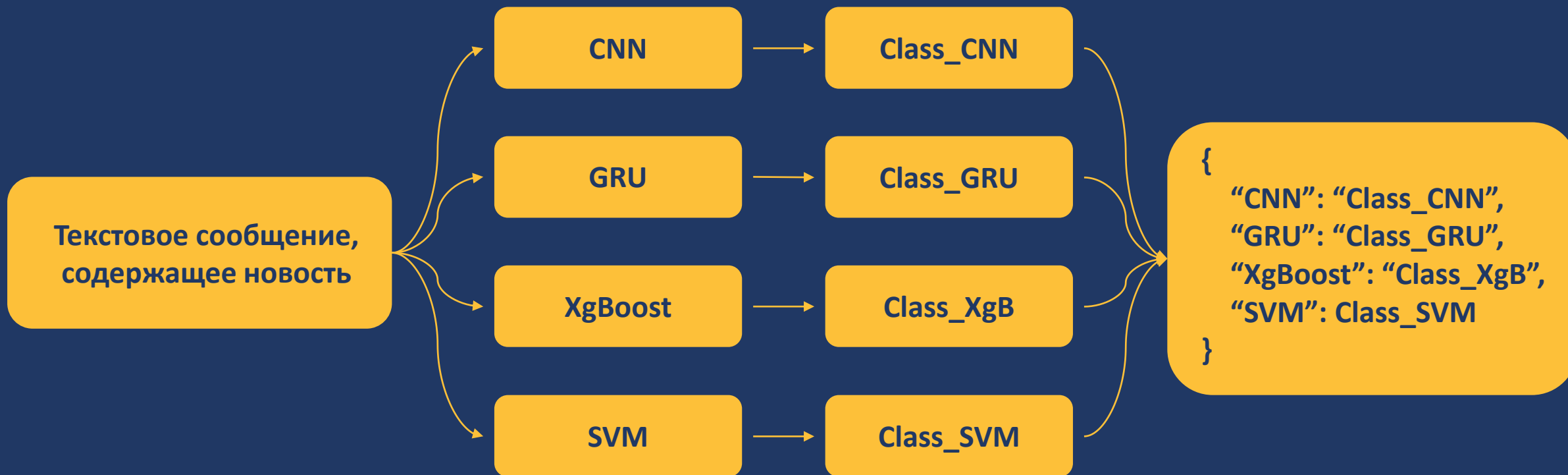
- CNN: 0.8354
- GRU: 0.8355
- XgBoost: 0.8649
- SVM: 0.8801

# Обучающий набор данных

	main_text	category
0	россия повысить зарплата россия повысить миним...	0
1	партия порошенко потребовать запретить поставк...	8
2	россия подорожать алкогольный напиток сигарета...	0
3	рогозин назвать поклонник маска свидетель илон...	6
4	россия измениться правило регистрация новый ав...	7
5	ким чен ын пообещать показать новый оружие сев...	2

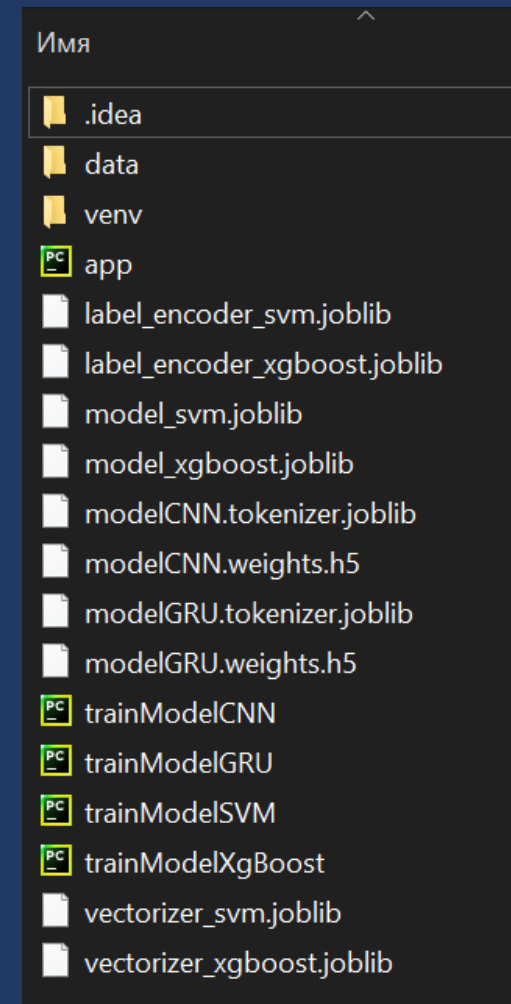
ким чен ын пообещать показать новый оружие северный корея лидер северный корея ким чен ын заявил, вскоре страна показать новый разработать вид стратегический оружие слово передавать северокорейский агентство цтак свой обращение трудовой партия корея ким чен ын заявил, сша попрехний настроить отношение северный корея враждебно, поэтому необходимо разрабатывать стратегический вооружение активно также заверил, скоро мир увидеть новый стратегический оружие кндр ранее северокорейский лидер заявил, страна необходимо принять активный наступательный мера последовательный обеспечение суверенитет безопасность северный корея планировать принять жёсткий политика отношение сша, число отказаться переговоры денуклеаризация декабрь президент сша дональд трамп предупредить лидер кндр ким чен ына риска потерять случай ухудшение отношение два страна

# Развертывание проекта. Идея



# Развертывание проекта. Реализация

- При развертывании проекта для каждой модели были созданы отдельные файлы ***trainModel\*.py***, где \* - название взятых моделей
- После обучения моделей были созданы файлы:
  - ***model\*.weights.h5*** - для моделей CNN и GRU, где хранятся веса и архитектура данных моделей
  - ***model\_\*.joblib*** – для XgBoost и SVM, где хранится обученный алгоритм
  - ***vectorizer\_\*.joblib*** – для XgBoost и SVM, где хранится векторизатор
  - ***label\_encoder\_\*.joblib*** – для XgBoost и SVM, где хранится информация о кодировке меток (категорий)
  - Был создан файл ***app.py*** для развертывания проекта с помощью библиотеки `fastapi`



# Развертывание проекта. Пример работы

Текст тестового файла economy\_new:

Response body

```
{  
  "CNN": 0,  
  "GRU": 0,  
  "XgBoost": 3,  
  "SVM": 0  
}
```

ассортимент торговый сеть измениться случай закрытие производство ростовский область покупатель заметить этот изменение такой мнение риа новость поделиться эксперт ритейл ранее газета коммерсант написать закрывать некоторый свой производственный площадка частность ростовский область говорить учёт широкий география производство маловероятный сокращение несколько производство россия значительно повлиять ассортимент наличие товар торговый сеть покупатель заметить этот изменение считать вицепрезидент союз торговый центр павел люлина объяснить несмотря позиционироваться российский производитель ретейлер значительный часть одежда производится граница страна низкий стоимость труд высокий технологичность доступ сырь например индия вьетнам производится трикотаж ткань нижний бельё пакистан денить текстиль бангладеш трикотаж денить верхний одежда нижний бельё также производство гонконг материковый китай турция узбекистан касаться фабрика россия который расположить основное юг работать весь швейный индустрия испытывать давление который серьёзно усложнять удорожать производство добавить люлина работа бренд сокращение производство юг никак отразиться очередь считать вицепрезидент союз торговый центр наталия кермедчиев предположить скорее идти оптимизация процесс бизнесконсультант управление продажа стратегический развитие дания ткачёв рассказать основной причина закрытие производство юг россия впервые нехватка персонал ростовский область южный федеральный округ регион один самый низкий уровень безработица производство функционировать нужный скорость швея должный достаточно кроме это рост арендный плата важно контролировать себестоимость производство этап производственный цикл ткачёв отметить оптимизация производство важно видеть выгодный производство свой портфель точка зрения зарплата регион точка зрения скорость производство поэтому бренд мочь переносить производство один регион внутри страна предел однако добавить крупный сеть вроде знать сколько время занимать производственный цикл логистика прочее думать это никак отразиться загрузка розничный магазин подытожить эксперт



# Развертывание проекта. Пример работы

## Response body

```
{  
  "CNN": 4,  
  "GRU": 4,  
  "XgBoost": 4,  
  "SVM": 1  
}
```

## Текст тестового entertainment\_new:

большой театр январь показать представление новогодний балет щелкунчик  
сообщаться официальный сайт габт афиша сайт театр появиться спектакль январь  
балет щелкунчик представить январь январь зритель также ждать спектакль утром  
вечером январь балет показать январь утром вечером ранее сообщаться декабрь  
большой театр представить показ щелкунчик прошлый год габт показать балет  
щелкунчик декабрь девять январь

# Развертывание проекта. Пример работы

Текст тестового файла science\_new:

Response body

```
{  
  "CNN": 6,  
  "GRU": 6,  
  "XgBoost": 6,  
  "SVM": 6  
}
```

исчезновение ледяной покров арктика ожидать середина век произойти условие быть примерно такой период год прийти фаза понижение температура воздух сообщить риа новость прессслужба арктический антарктический научноисследовательский институт ааний ранее учёный гетеборгский университет представить исследование согласно который существовать риск ледяной покров арктика изз изменение погодный условие полностью растаять лето это произойти год возможно м мнение учёный арктический антарктический научноисследовательский институт исчезновение ледяной покров северный ледовитый океан ожидать середина век произойти ледовый условие арктический море быть примерно такой самый лёгкий год акватория арктический море свободный от лёд август октябрь период год прийти фаза понижение температура воздух ход летний колебание ледовый условие арктический море быть близкий современный говориться сообщение отмечаться незначительный изменение ледяной покров арктика фиксироваться быть год начало x год начаться резкий перемена семь год площадь арктический морской лёд пик сезонный минимум сентябрь упасть сравнение аналогичный период время регулярный спутниковый наблюдение год последний год летний площадь лёд варьироваться год год целое сохраняться новый среднее уровень который примерно маленький наблюдаться год отметить институт данные учёный минимальный значение ледовитость северный морской путь сентябрь достигнуть год тысяча квадратный километр однако м площадь лёд период возрасти тысяча квадратный километр м увеличиться тысяча квадратный километр хотя температура российский арктика год выше норма градус

# Где можно посмотреть весь проект целиком

<https://github.com/leereshaus/IT-project/tree/main> - ссылка на репозиторий на GitHub, с помощью которого происходил обмен файлами между участниками команды

Репозиторий включает в себя:

- Readme файл с кратким описанием проекта
- Папку «**Описание проекта**», в которой хранятся презентации для выступления и теория по некоторым методам
- Папку «**Данные**», где есть ссылки на получившиеся датасеты
- Папку «**Разработка**», в которой находятся скрипты парсера, предобработки датасета и обучения моделей
- Папку «**Развёртывание проекта**», в которой находятся отдельные файлы с обучением конкретных моделей и реализация приложения через fast api
- Папку «**Тесты**», в которой лежат тесты для проверки корректной работы приложения

# Роли участников команды

- **Амина:** Подготовка данных: написание парсера и предобработка полученного датасета. Обучение CNN и RNN
- **Ксюша:** Подготовка данных для отдельных моделей. Обучение GRU, BiLSTM, XgBoost, SVM
- **Соня:** Организационная деятельность: распределение задач, создание репозитория на Git Hub, подготовка выступлений. Разработка приложения с помощью fast api
- **Денис:** Подготовка тестовых запросов для проверки корректной работы приложения