


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

Обзор решения задачи

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

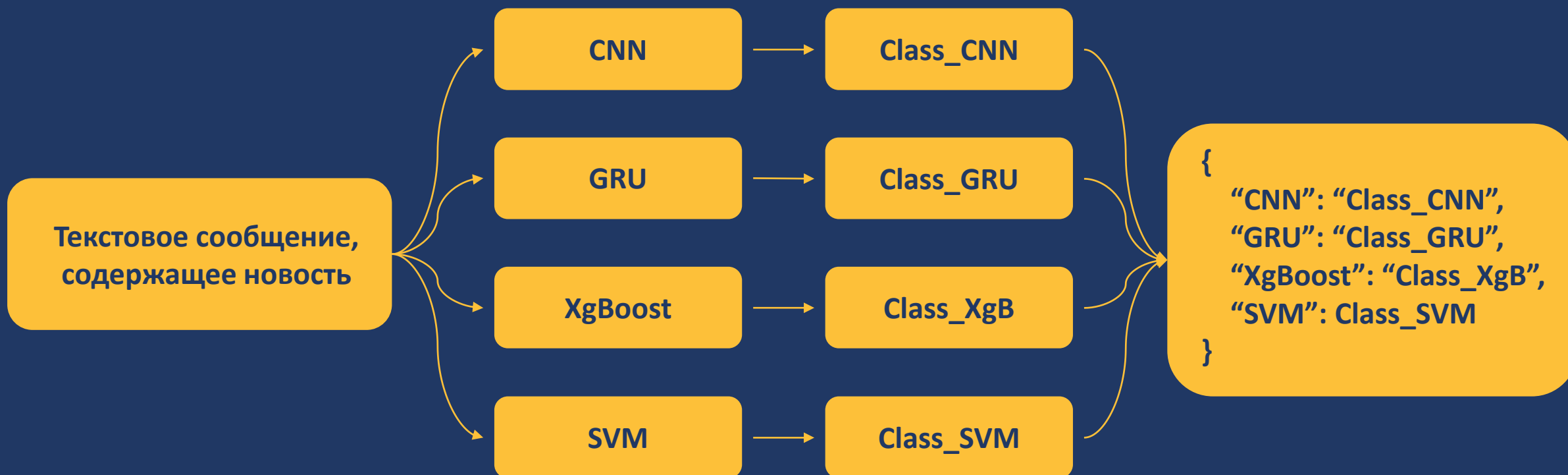
- CNN: 0.8354
- GRU: 0.8355
- XgBoost: 0.8649
- SVM: 0.8801

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

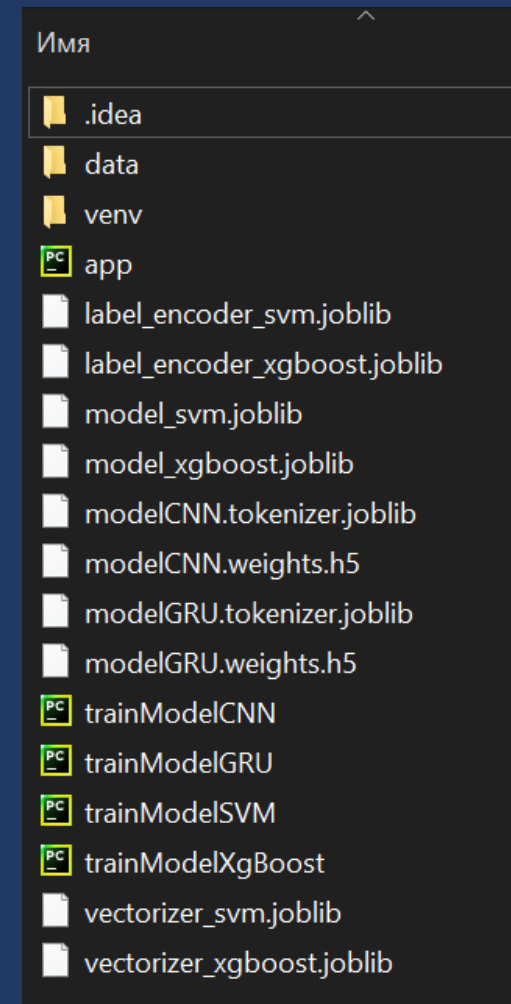
ким чен ын пообещать показать новый оружие северный корея лидер северный корея ким чен ын заявил, вскоре страна показать новый разработать вид стратегический оружие слово передавать северокорейский агентство цтак свой обращение трудовой партия корея ким чен ын заявил, сша попрехний настроить отношение 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

```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict/' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@example.txt;type=text/plain'
```

Request URL	
http://127.0.0.1:8000/predict/	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "CNN": 2, "GRU": 2, "XgBoost": 2, "SVM": 2 }</pre> <div>Response headers</div> <pre>content-length: 37 content-type: application/json date: Sat, 14 Dec 2024 21:03:29 GMT server: uvicorn</pre>

Текст тестового файла:

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 северокарейский 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 северокарейский 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. Пример работы

```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict/' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@example.txt;type=text/plain'
```

Request URL	
http://127.0.0.1:8000/predict/	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "CNN": 2, "GRU": 2, "XgBoost": 2, "SVM": 2 }</pre> <div>Response headers</div> <pre>content-length: 37 content-type: application/json date: Sat, 14 Dec 2024 21:03:29 GMT server: uvicorn</pre>

Текст тестового файла:

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 северокареийский агентство цтак свой обращение трудовой партия корея ким чен ын заявил, сша попрешний настроить отношение северный корея враждебно, поэтому необходимо разрабатывать стратегический вооружение активно также заверил, скоро мир увидеть новый стратегический оружие кндр ранее северокареийский 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. Пример работы

```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict/' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@example.txt;type=text/plain'
```

Request URL	
http://127.0.0.1:8000/predict/	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "CNN": 2, "GRU": 2, "XgBoost": 2, "SVM": 2 }</pre> <div>Response headers</div> <pre>content-length: 37 content-type: application/json date: Sat, 14 Dec 2024 21:03:29 GMT server: uvicorn</pre>

Текст тестового файла:

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 попреежний 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. Пример работы

```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict/' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@example.txt;type=text/plain'
```

Request URL	
http://127.0.0.1:8000/predict/	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "CNN": 2, "GRU": 2, "XgBoost": 2, "SVM": 2 }</pre> <div>Response headers</div> <pre>content-length: 37 content-type: application/json date: Sat, 14 Dec 2024 21:03:29 GMT server: uvicorn</pre>

Текст тестового файла:

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 попреежний 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. Пример работы

```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict/' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file=@example.txt;type=text/plain'
```

Request URL	
http://127.0.0.1:8000/predict/	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "CNN": 2, "GRU": 2, "XgBoost": 2, "SVM": 2 }</pre> <div>Response headers</div> <pre>content-length: 37 content-type: application/json date: Sat, 14 Dec 2024 21:03:29 GMT server: uvicorn</pre>

Текст тестового файла:

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 попреежний 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

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 находится скрипты парсера, предобработки датасета и обучения моделей
- Папку «Развёртывание проекты», 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