Предсказание тематических категорий новостей по их содержанию

Дипломный проект студента группы DSU-60 Трофимова Павла

ООО «Нетология», 2025 год

Введение

Объём новостных источников и статей растёт с каждым днём. Помимо официальных новостных источников в сети появляется всё больше людей, которые ведут собственные каналы в соцсетях и мессенджерах, делятся новостями, аналитикой и личными мнениями. В таких условиях пользователю становится все сложнее быстро находить действительно важную и интересующую его информацию.

Предложение: Перспективным решением для эффективной навигации видится поиск новостей по тегам, проставленным на основании содержания.

Применение: Системы категоризации используются в новостных агрегаторах (Яндекс.Новости, Лента, РБК), для персонализированных рекомендаций и модерации контента.

Достоинства и недостатки

Достоинства использования тегов:

- Возможность группировать публикации вне зависимости от источника (крупные новостные платформы, пользовательские медиа) по темам, событиям или ключевым персонам
- Обеспечение доступа к полному списку материалов в рамках интересующего вопроса
- Упрощенный поиск и фильтрация контента
- Снижение возможности упустить важные детали
- Повышение роста вовлеченности аудитории
- Поддержка актуальности новостного потока, обусловленная быстрой ориентацией даже в самых насыщенных информационных лентах

Основные недостатки использования тегов:

- Сложность назначения тегов для составителя, ввиду при написании публикаций и необходимости продумывания ключевых тем, описываемых в публикациях
- Сложность разметки / назначения тегов сторонним человеком, ввиду необходимости погружения в контекст, постоянных концентрации и внимания на темах, описываемых в публикациях
- Высокие трудозатраты на назначение тегов при больших объёмах и количестве публикаций
- Пересечение тегов для публикаций из разных тематических областей

Задача

Основная задача проекта включает в себя несколько этапов:

1. Подготовка данных по новостным публикациям для дальнейшего обучения моделей.
2. Разработка модели машинного обучения для решения задачи многоклассовой классификации, в частности автоматического назначения тематических категорий для новостей на основе их содержания
3. Проведение оценки и сравнения результатов по обученным моделям с использованием стандартных метрик

Входные данные

Система обрабатывает необработанный текст новостей. Текст может быть разной длины и лексики.

Цель

Основная цель — разработка и выбор наилучшего решения по автоматическому назначению корректных тегов для новостных публикаций для упрощения поиска и фильтрации публикаций по интересующим пользователя вопросам/темам

Выходные данные

Каждая новость должна быть отнесена к одной из предопределённых категорий: «Политика», «Экономика», «Спорт», «Культура», «Общество», «Происшествия» и др.

Описание исходных данных

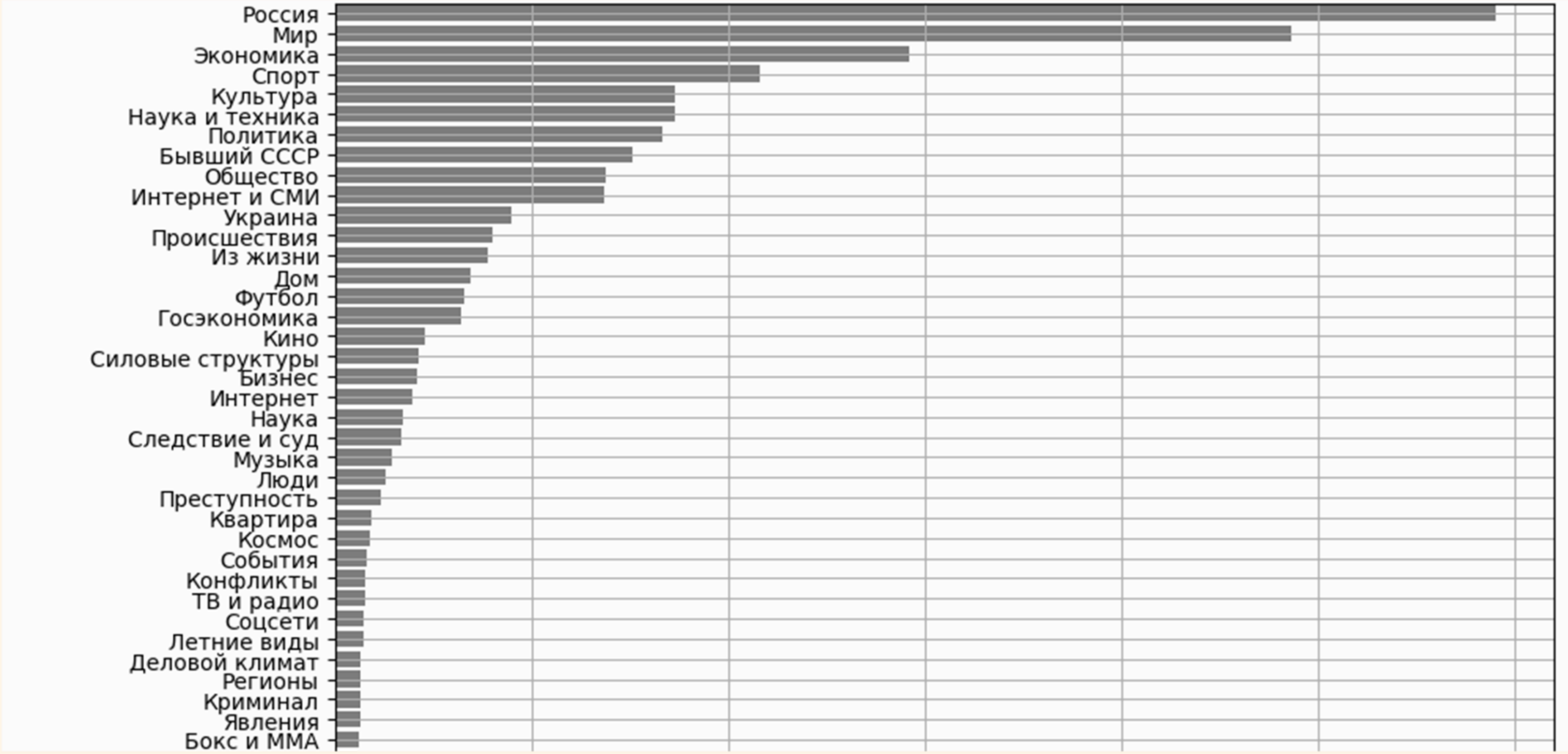
- Источник: Открытый датасет новостей Lenta.ru
- Представленный период публикаций:

  Сентябрь 1999 года - Декабрь 2019 года
- Количество уникальных записей:

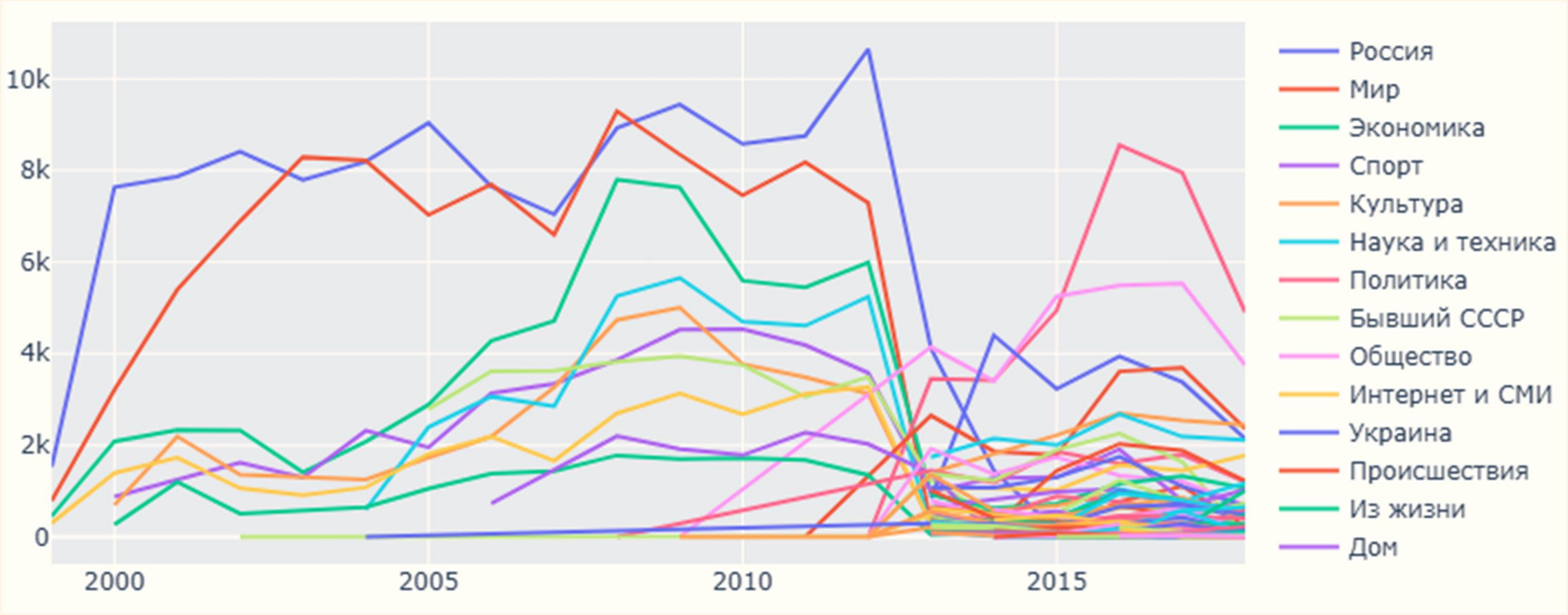
  Более 800 000
- Размер:

  Более 2 Гб
- Поля и описание:
  - url (Ссылка на публикацию, уникальные значения)
  - title (Заголовок публикации)
  - text (Содержание публикации)
  - topic (Тема публикации)
  - tags (Теги (подтемы) публикации)
  - date (Дата публикации)
- Пример категорий: "Политика", "Спорт"

Сравнение количества новостей по тегам



Сравнение количества новостей по годам для тегов



Описание данных для исследования

- Источник: Открытый датасет новостей Lenta.ru
- Представленный период публикаций:

  Сентябрь 1999 года - Декабрь 2019 года
- Количество уникальных записей: 70 889
- Поля и описание:
  - id (Идентификатор публикации, уникальные значения)
  - text (Содержание публикации)
  - колонки по каждому значению tags (Теги (подтемы) публикации)

Датасет для исследования

	text	Соцсети	Авто	Автобизнес	Белоруссия	Бизнес	Бокс и ММА	Бывший СССР	Вещи	Вирусные ролики	...	Туризм	Украина	Финансы компаний	Футбол	Хоккей	Ценности	Часы	Экология	Экономика	Явления
0	Вице-премьер Владислав Сурков заявил, что не и...	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Владислав Цыплухин, занимавший пост руководите...	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Сервис микроблогов Twitter откроет представите...	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Ежемесячная аудитория фотосервиса Instagram со...	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Пользователи Facebook загрузили на Новый год в...	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
70884	Американская супермодель Джиджи Хадид разработ...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
70885	Американский предприниматель Марк Фарезе (Mark...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
70886	Новой темой ежегодного Met Gala в музее Метроп...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
70887	Французская актриса Марион Котийяр приняла уча...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
70888	Британский модельер Пол Смит стал приглашенным...	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

70889 rows × 90 columns

Обработка данных

1 Очистка текста

Удаление пунктуации, чисел и ссылок для унификации данных.

2 Лемматизация

Приведение слов к начальной форме для уменьшения размерности словаря с использованием PyMystem3.

3 Токенизация

Разбиение текста на слова с использованием библиотек NLTK

4 Удаление стоп-слов

Фильтрация общих слов (например, "и", "но"), не несущих смысловой нагрузки.

5 Извлечение признаков

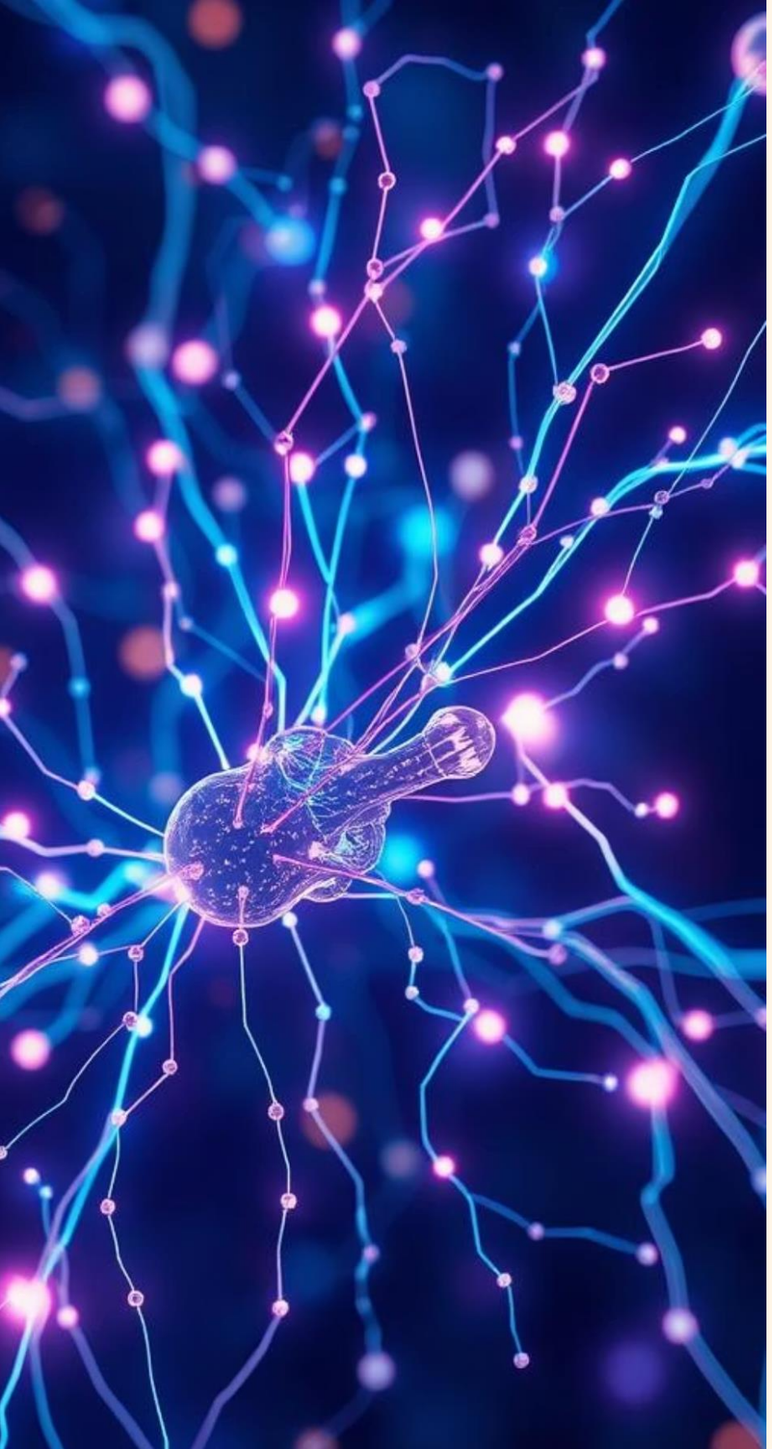
Использование TF-IDF для создания векторного представления текста

Обучение классических моделей

LogisticRegression с TfidfVectorizer на униграммах
Accuracy: 0.36987
Hamming distance: 0.58684
Precision = 0.50, Recall = 0.87, F1 = 0.63

LogisticRegression с TfidfVectorizer на униграммах и биграммах
Accuracy: 0.36754
Hamming distance: 0.58182
Precision = 0.50, Recall = 0.86, F1 = 0.63

CatBoostClassifier с TfidfVectorizer на униграммах
Accuracy: 0.47447
Hamming distance: 0.49301
Precision = 0.81, Recall = 0.51, F1 = 0.63



Обучение нейросетевых моделей

ruBERT tiny2
Accuracy: 0.68557
Hamming distance: 0.69075
Precision = 0.76, Recall = 0.70, F1 = 0.73

ruBERT base
Accuracy: 0.75801
Hamming distance: 0.76153
Precision = 0.80, Recall = 0.77, F1 = 0.78

DistilBERT
Accuracy: 0.70976
Hamming distance: 0.71629
Precision = 0.75, Recall = 0.72, F1 = 0.74

XLM-RoBERTa
Accuracy: 0.74312
Hamming distance: 0.74774
Precision = 0.78, Recall = 0.75, F1 = 0.77

Сравнение моделей

	model_name	Accuracy	Hamming_distance	F1_micro	F1_macro	Recall_micro	Recall_macro	Precision_micro	Precision_macro	Train time, min.	Size, MB
0	tfidf_logreg_ngram_1_1	0.3674	0.4790	0.5238	0.5169	0.6250	0.6180	0.4508	0.4969	2	110
1	tfidf_logreg_ngram_1_2	0.3826	0.4946	0.5526	0.5327	0.6364	0.6292	0.4884	0.5150	30	2360
2	tfidf_catboost_ngram_1_1	0.4242	0.4280	0.5672	0.4639	0.4318	0.4270	0.8261	0.5590	169	661
3	ai-forever-rubert-base	0.5606	0.5606	0.5932	0.5179	0.5606	0.5543	0.6298	0.5449	365	680
4	cointegrated-rubert-tiny2	0.5038	0.5057	0.5654	0.4823	0.5076	0.5019	0.6381	0.5216	59	111
5	FacebookAI-xlm-roberta-base	0.5455	0.5474	0.5812	0.4938	0.5492	0.5431	0.6170	0.4853	778	1030
6	Geotrend-distilbert-base-ru-cased	0.5227	0.5284	0.5640	0.4931	0.5341	0.5281	0.5975	0.5059	261	208

Президент России Владимир Путин своим указом произвел в генералы 33 полковника различных ведомств — от Министерства обороны до Военной прокуратуры. Соответствующий указ в среду, 12 декабря, опубликован на официальном портале правовой информации. Кроме того, еще 16 генералам присвоены очередные воинские звания. В частности, в Минобороны России появился один новый генерал-полковник и один полный адмирал, восемь генерал-лейтенантов и два приравненных к ним вице-адмирала, 10 генерал-майоров и один контр-адмирал. В МЧС появиллось два генерал-полковника, четыре генерал-лейтенанта и семь генерал-майоров. В свою очередь, Росгвардия пополнилась одним генерал-полковником, тремя генерал-лейтенантами и девятью генерал-майорами. В Военной прокуратуре появился один генерал-лейтенант юстиции и три генерал-майора. Наконец, Федеральная служба исполнения наказаний (ФСИН) получила двух генерал-лейтенантов и четырех генерал-майоров. Так, генерал-лейтенант Евгений Устинов, командующий войсками Центрального военного округа стал генерал-полковником, а командующий Балтийским флотом вице-адмирал Александр Носатов отныне — полный адмирал. Командующий Уральским округом Росгвардии генерал-лейтенант Александр Попов получил на погоны третью шитую звезду и стал генерал-полковником. В основном генеральские звания присвоены офицерам, чей «потолок по должности» с недавнего времени увеличен — в течение ноября-декабря президент внес изменения в положения почти о всех силовых структурах, увеличив число генеральских должностей.

Результаты

LogisticRegression с TfidfVectorizer на униграммах
pred_tag: Полиция и спецслужбы, Россия, Оружие

LogisticRegression с TfidfVectorizer на униграммах и биграммах
pred_tag: Полиция и спецслужбы, Россия, Общество, Оружие

CatBoostClassifier с TfidfVectorizer на униграммах
pred_tag: Полиция и спецслужбы

ruBERT tiny2
pred_tag: Полиция и спецслужбы

ruBERT base
pred_tag: Полиция и спецслужбы

DistilBERT
pred_tag: Полиция и спецслужбы

XLM-RoBERTa
pred_tag: Полиция и спецслужбы

На примерах хорошо видно разницу в результатах работы алгоритмов. Логистическая регрессия на tfidf-векторах хорошо определяет не только целевые категории, но и смежные к ним, отсюда должно быть большее количество ложноположительных предсказаний и более низкий Precision. Нейросети более точечны в своих предсказаниях, а потому лучше подстраиваются под имеющиеся данные.

Выводы

Достижения

Успешно разработано решение по автоматическому назначению тегов для новостей на основе их содержания на базе ruBERT base

Внедрение

Интеграция в существующие новостные агрегаторы, создание новостного агрегатора на платформе Telegram

Ограничения

Требуется значительный объем размеченных данных и технических мощностей для эффективного обучения моделей

Перспективы

Исследование новых методов многоклассовой классификации и адаптация к постоянно меняющимся новостным данным.

Спасибо за внимание!