ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР

Программа профессиональной переподготовки «Data Science»

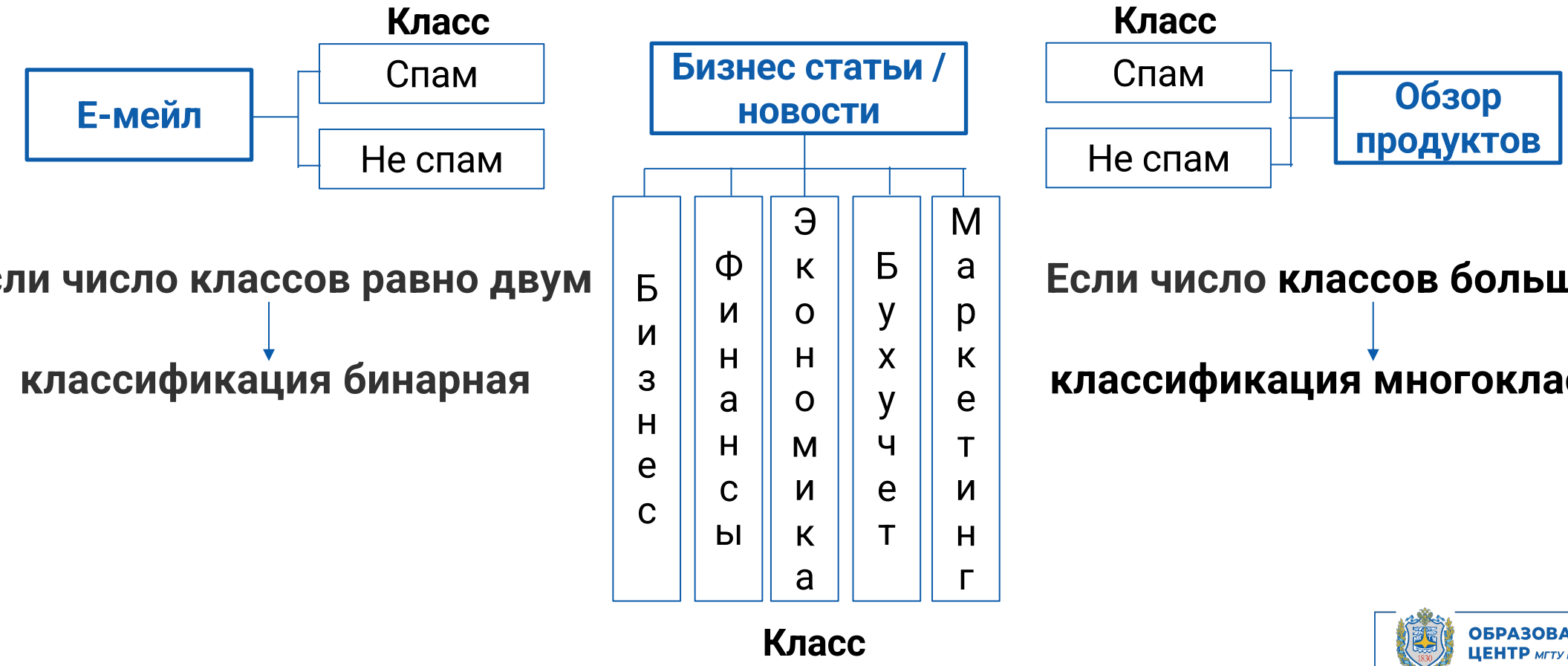
ИТОГОВАЯ РАБОТА
КЛАССИФИКАЦИЯ ТЕКСТОВ НА МАЛЕНЬКИХ ДАТАСЕТАХ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Мустяца Алексей Григорьевич

Москва 2022

Что такое классификация текста?

Классификация текстов (документов) — задача, заключающаяся в отнесении текста (документа) к одной из нескольких категорий на основании содержания текста (документа), известной как класс



Актуальность классификации текстов

- Огромное количество текстового контента (новостные статьи, обзоры продуктов, и публикации в социальных сетях и т.п) публикуемого каждый день
- Рост неструктурированных текстовых данных в виде коротких текстов активно стимулируемый социальными сетями
- Все больше времени пользователи вынуждены тратить на прочтение и отсеивание ненужной информации.
- Современные подходы традиционной классификации текстов плохо работают с короткими текстами, если их применять напрямую
- Многообразие имеющихся алгоритмов и отсутствие универсального решения
- Классификация текста — это задача с несколькими классами, где каждый документ может принадлежать к одной, многим или ни к одной из категорий.

Подходы к решению задачи классификации

- Имеются множество документов $D = \{d_1, \dots, d_{|D|}\}$ и множество возможных категорий (классов) $C = \{c_1, \dots, c_{|C|}\}$
- Неизвестная целевая функция $\Phi: D \times C \to \{0,1\}$ задается формулой

$$\Phi(d_j, c_i) = \begin{cases} 0, \text{если } d_j \notin c_i \\ 1, \text{если } d_j \notin c_i \end{cases} \quad (1)$$

- Требуется построить классификатор Φ', максимально близкий к Φ.
- Документ d_j называется положительным примером категории c_i, если $\Phi(d_j, c_i) = 1$, и отрицательным в противном случае.
- Если классификатор выдает точный ответ: $\Phi': D \times C \to \{0,1\}$ (2)
то классификация называется точной.
Если классификатор определяет степень подобия документа
$CSV: D \times C \to \{0,1\}$ (3)
то классификация называется пороговой.

Рис. 1 Обучение с учителем *vs* частичное обучение с учителем *vs* обучение без учителя

Актуальность работы

ЦЕЛЬ

ПРОДЕМОНСТРИРОВАТЬ, ЧТО ИСПОЛЬЗОВАНИЕ НЕРАЗМЕЧЕННЫХ ДАННЫХ ПОЗВОЛЯЕТ ПОВЫСИТЬ ТОЧНОСТЬ КЛАССИФИКАТОРА ТЕКСТОВ

ЗАДАЧА

ПОСТРОИТЬ SSL-МОДЕЛЬ, КОТОРАЯ БУДЕТ КЛАССИФИЦИРОВАТЬ ТЕКСТЫ ЛУЧШЕ МОДЕЛИ, ОБУЧЕННОЙ ТОЛЬКО НА РАЗМЕЧЕННЫХ ДАННЫХ

ПРОБЛЕМА

РАЗМЕТКА ДАННЫХ: МЕДЛЕННО ДОРОГО ТРУДОЗАТРАТНО НЕ ВСЕГДА НАДЕЖНО

РЕШЕНИЕ

НЕМНОГО РАЗМЕЧЕННЫХ ДАННЫХ + МНОГО НЕРАЗМЕЧЕННЫХ ДАННЫХ

МОДЕЛЬ ЧАСТИЧНОГО ОБУЧЕНИЯ (SSL)

Схема обучения модели в концепции SSL

1-2 этапы

ИЗУЧЕНИЕ ДАТАСЕТОВ НОВОСТЕЙ / ПРЕДОБРАБОТКА ДАННЫХ

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Этап 1 Анализ новостного датасета

- 4 корпуса неразмеченных новостных данных (2016-2017гг) из открытых источников
- Датасет с корпусом из 576 383 новостей
- Примеры новостей в датасете

```
[ ] df=pd.read_csv('/content/drive/MyDrive/Colab Notebooks/20170301.csv')

[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 576383 entries, 0 to 576382
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype 
---  ------       --------------   ----- 
 0   u_id         576383 non-null  int64 
 1   provider     576383 non-null  object
 2   date_time    576383 non-null  object
 3   title        576383 non-null  object
 4   description  542805 non-null  object
 5   link         576292 non-null  object
 6   pubdate      576383 non-null  object
 7   numfield     576383 non-null  int64 
dtypes: int64(2), object(6)
```

	u_id	provider	date_time	title	description	link	pubdate	numfield
0	103451	washingtonpostcom_world[eng]	2017-01-17 20:15:49	"$10,000 stuffed in a diplomat-s car." Moscow ...	Russia-s foreign minister said that United Sta...	https://www.washingtonpost.com/news/worldviews...	17.01.2017 20:15	3
1	211367	vsesmiru_business	2017-02-02 06:00:33	"100 друзей" Гродненского мясокомбината	Дизайнеры агентства Fabula Branding (Минск) пр...	http://www.vsesmi.ru/business/2017/02/02/321648/	02.02.2017 01:03	4
2	13559	mailru_common	2016-11-28 13:30:23	"12-я партия - игра жизни не только Карякина, ...	<p>Победитель шахматной олимпиады 1998 гроссме...	https://sport.mail.ru/news/chess/27940207/	28.11.2016 13:22	4
3	17165	mailru_common	2016-11-28 21:15:25	"37 мне только по паспорту". Почему Лебедев не...	<p>Боксёры Денис Лебедев и Мурат Гассиев, встр...	https://sport.mail.ru/news/boxing/27946630/	28.11.2016 21:09	4
4	147533	mailru_polit	2017-01-24 10:15:32	"5 канал": Украина подала в ЕСПЧ пять исков пр...	<p>Речь идет о событиях в Крыму и Донбассе и н...	https://news.mail.ru/politics/28526323/	24.01.2017 10:04	4

- Значимые столбцы - title (краткое описание новости), description (контекст новости), link (web-источник нахождения новости)

Этап 1 Анализ и обработка новостей

- Отобраны 37 243 новостей бизнеса, экономики и финансов на русском языке

	u_id	provider	date_time	title	description	link	pubdate	numfield	Languagereveiw
0	211367	vsesmiru_business	2017-02-02 06:00:33	"100 друзей" Гродненского мясокомбината	Дизайнеры агентства Fabula Branding (Минск) пр...	http://www.vsesmi.ru/business/2017/02/02/321648/	02.02.2017 01:03	4	True
1	353361	vsesmiru_business	2017-02-24 09:00:34	"ArcelorMittal Кривой Рог" инициировал антидем...	В течение 30 дней, с даты публикации сообщения...	http://www.vsesmi.ru/business/2017/02/24/385699/	24.02.2017 08:31	4	True
2	65369	newrucom_common	2016-12-05 20:30:06	"Абсолютно никчемный аргумент": Путин прокомме...	Глава государства назвал совершенно не имеющей...	http://www.newsru.com/finance/05dec2016/gazpro...	05.12.2016 20:26	4	True
3	58879	vsesmiru_business	2017-01-11 07:45:35	"Аврора" со следующей недели открывает новый р...	Интерфакс-Россия, Новость:\nАвиакомпания "Авро...	http://www.vsesmi.ru/business/2017/01/11/256028/	11.01.2017 07:26	4	True
4	243101	vestifinance_common	2017-02-07 11:15:54	"АвтоВАЗ" наращивает продажи. Главное	Отечественный концерн по производству автомоби...	http://www.vestifinance.ru/videos/32121	07.02.2017 11:11	4	True

- Произведена предобработка вручную текстов первых 2 069 новостей из отобранных новостей бизнеса, экономики и финансов на русском языке
- Произведена разметка новостей с целью определения оттенка каждой новости для последующего прогнозирования разделения новостей на классы

Этап 2 Предобработка данных

В процессе разметки новостям вручную присваивались классы по результатам изучения контекста и содержания всего текста новости
Конкретный алгоритм выявления слов-маркеров отсутствовал

Класс 1 - реклама, позитивное ожидание чего-либо

Класс 1 **(464)** - Дизайнеры агентства Fabula Branding (Минск) провели комплексную разработку торговой марки колбасных изделий «100 друзей» (нейминг, логотип, дизайн упаковки) для ОАО «Гродненский мясокомбинат». Продукт – колбасные изделия среднего ценового сегмента: сырокопченые, сыровяленые, вареные колбасы, сосиски и сардельки. Регионы продаж – Беларусь и Россия. Ситуации потребления: дружеские и семейные застолья, пикник, гости, быстрый перекус. Целевая аудитория – мужчины и женщины 25-40 лет. Решением стал теплый и яркий желтый цвет, который хорошо отстраивает продукт на полке, служит цветовым идентификатором торговой марки и вызывает приятные ассоциации с пикником / дружескими посиделками на даче. Дружелюбную и приветливую стилистику дизайна поддерживает цифра «100», образованная колбасками (маркер категории), а также рукописный шрифт логотипа

Класс 2 - судебные дела, иски, претензии. Стоп-фактор, требует рассмотрения новости вручную

Класс 2 **(213)** - В течение 30 дней, с даты публикации сообщения в "Урядовом курьере", министерство будет проводить регистрацию заинтересованных в расследовании лиц и рассматривает требования относительно проведения слушаний. В течение 60 дней, МЭРТ рассматривает письменно изложенные комментарии и информацию относительно возбуждения расследования.

Этап 2 Предобработка данных

- **Класс 3 - информационное сообщение, нейтральное по смыслу**

  Класс 3 **(1 125)** - Глава государства назвал совершенно не имеющей под собой оснований "идейкой" расхожее мнение о том, что те, кто покупает российский газ, попадают в зависимость от Москвы. "Это глупый абсолютно, никчемный аргумент, потому что это взаимозависимость", - подчеркнул президент России.

- **Класс 4 - рост продаж, производства, поставок - позитивный характер в соответствии с контекстом и содержанием новостей**

  Класс 4 **(155)**- "АвтоВАЗ" в январе увеличил продажи на 4,6% в годовом выражении до 16,3 тыс. автомобилей, сообщает компания. Доля Lada на российском рынке составила в прошлом месяце почти 19,5%, против 19,9% в 2016 г.

- **Класс 5 - уменьшение продаж, поставок и т.д. - негативный характер в соответствии с контекстом и содержанием новостей**

  Класс 5 **(112)** - "АвтоВАЗ" сократит 740 человек (2% сотрудников). Попадающим под сокращение работникам предложены вакансии во всех подразделениях "АвтоВАЗа", в том числе в Индустриальном парке.

Данные классы были выбраны для выявления характера новостного тренда: позитивный, негативный, судебный. В случае судебного тренда необходимо было вручную определить дальнейшие действия, так как это могла быть, как положительная новость для одной из сторон, так и отрицательная.

Этап 2 Предобработка данных

- Очистка текста новостей: удалены стоп-слова, текст обработан, очищен от знаков препинания, пробелов, цифр, и переведен в нижний регистр

	Новость	Класс	Очистка	Перевод в нижний регистр
0	Дизайнеры агентства Fabula Branding (Минск) пр...	1	Дизайнеры агентства Fabula Branding Минск пров...	дизайнеры агентства fabula branding минск пров...
1	В течение 30 дней, с даты публикации сообщения...	2	В течение дней с даты публикации сообщения в У...	в течение дней с даты публикации сообщения в у...
2	Глава государства назвал совершенно не имеющей...	2	Глава государства назвал совершенно не имеющей...	глава государства назвал совершенно не имеющей...
3	Интерфакс-Россия, Новость: Авиакомпания "Аврор...	3	Интерфакс Россия Новость Авиакомпания Аврора с...	интерфакс россия новость авиакомпания аврора с...
4	Отечественный концерн по производству автомоби...	4	Отечественный концерн по производству автомоби...	отечественный концерн по производству автомоби...

- Стэмминг, токенизация, лемматизация новостей

	Новость	Класс	Очистка	Перевод в н/регистр	Стемминг	Токенизация	Лематизация
2060	Шведская компания IKEA выплатит $50 млн в каче...	2	Шведская компания IKEA выплатит млн в качестве...	шведская компания ikea выплатит млн в качестве...	шведск компан ike выплат млн качеств компенсац...	шведская компания ikea выплатит млн качестве к...	шведский компания ikea выплачивать млн качеств...
2061	Компания IRI Investments Lietuva, контролируем...	3	Компания IRI Investments Lietuva контролируема...	компания iri investments lietuva контролируема...	компан ir investments lietuv контролируем швед...	компания iri investments lietuva контролируема...	компания iri investments lietuva контролируемы...
2062	Краснинский суд Смоленской области арестовал с...	2	Краснинский суд Смоленской области арестовал с...	краснинский суд смоленской области арестовал с...	краснинск суд смоленск област арестова счет мл...	краснинский суд смоленской области арестовал с...	краснинский суд смоленский область арестовыват...
2063	Шведская компания IKEA не согласна с решением ...	2	Шведская компания IKEA не согласна с решением ...	шведская компания ikea не согласна с решением ...	шведск компан ike согласн решен суд арестова м...	шведская компания ikea согласна решением суда ...	шведский компания ikea согласный решение суд а...
2064	Шведская IKEA не будет строить торговый центр ...	3	Шведская IKEA не будет строить торговый центр ...	шведская ikea не будет строить торговый центр ...	шведск ike стро торгов центр quot мег quot мыт...	шведская ikea строить торговый центр quot мега...	шведский ikea строить торговый центр quot мега...
2065	Шведская IKEA планирует выставить права на дол...	3	Шведская IKEA планирует выставить права на дол...	шведская ikea планирует выставить права на дол...	шведск ike планир выстав прав долгосрочн аренд...	шведская ikea планирует выставить права долгос...	шведский ikea планировать выставлять право дол...
2066	Шведская компания IKEA намерена трудоустроить ...	3	Шведская компания IKEA намерена трудоустроить ...	шведская компания ikea намерена трудоустроить ...	шведск компан ike намер трудоустро окол сирийс...	шведская компания ikea намерена трудоустроить ...	шведский компания ikea намерен трудоустраивать...
2067	Шведский ритейлер IKEA направил обращение упол...	2	Шведский ритейлер IKEA направил обращение упол...	шведский ритейлер ikea направил обращение упол...	шведск ритейлер ike направ обращен уполномочен...	шведский ритейлер ikea направил обращение упол...	шведский ритейлер ikea направлять обращение уп...
2068	Каждый пятый товар сети подешевеет на 15–20%	5	Каждый пятый товар сети подешевеет на –	каждый пятый товар сети подешевеет на –	кажд пят товар сет подешевеет –	каждый пятый товар сети подешевеет –	каждый пятый товар сеть подешеветь –\n

- Преобразование текста в векторную форму осуществлялось в основном через CountVectorizer

Схема обучения модели в концепции SSL

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Этап 3 Результаты применения N-gram

- Для определения часто встречающихся сочетаний слов (n-gram) был проведен анализ частоты использования соседних слов в обучающей выборке (выделение N-gram)
- Определенных слов-маркеров, однозначно относящих новость к классу, выявлено не было
- Результаты применения алгоритма N-gram

```
[ ]  Counter(bigrams).most_common(5)

     [(('говорится', 'сообщении'), 62),
      (('уровне', 'баррель'), 62),
      (('млрд', 'рублей'), 61),
      (('deutsche', 'bank'), 48),
      (('млрд', 'руб'), 46)]
```

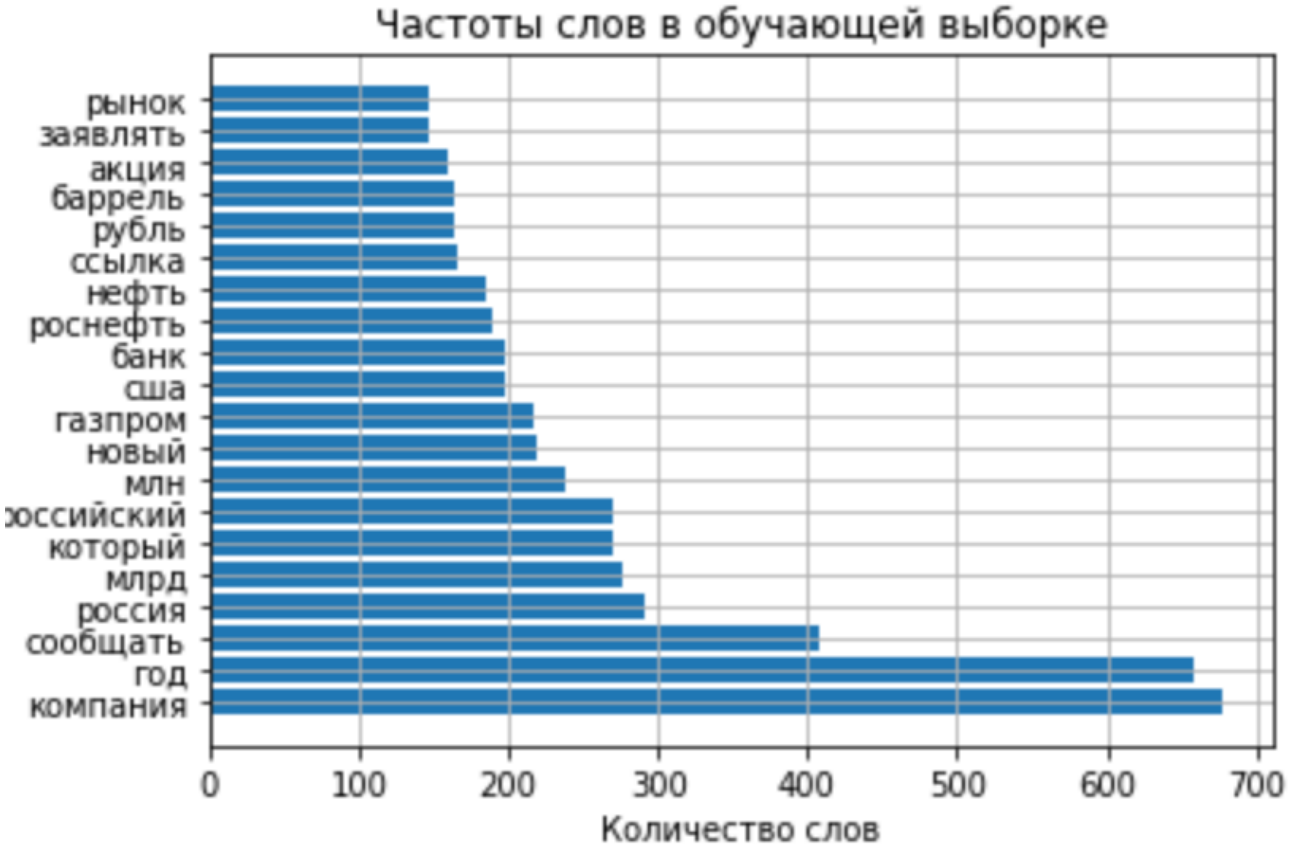
```
[ ]  Counter(trigrams).most_common(5)

     [(('фьючерсы', 'brent', 'торговались'), 35),
      (('сообщает', 'rns', 'ссылкой'), 25),
      (('бирже', 'ice', 'futures'), 24),
      (('сша', 'дональда', 'трампа'), 23),
      (('лондонской', 'бирже', 'ice'), 23)]
```

```
[ ]  Counter(fourgrams).most_common(5)

     [(('лондонской', 'бирже', 'ice', 'futures'), 23),
      (('президента', 'сша', 'дональда', 'трампа'), 22),
      (('фьючерсы', 'brent', 'торговались', 'лондоне'), 20),
      (('brent', 'торговались', 'лондоне', 'уровне'), 20),
      (('торговались', 'лондоне', 'уровне', 'баррель'), 20)]
```

```
[ ]  Counter(fivegrams).most_common(5)

     [(('фьючерсы', 'brent', 'торговались', 'лондоне', 'уровне'), 20),
      (('brent', 'торговались', 'лондоне', 'уровне', 'баррель'), 20),
      (('электронных', 'торгах', 'товарной', 'биржи', 'nymex'), 15),
      (('общий', 'объем', 'продажи', 'иностранной', 'валюты'), 14),
      (('объем', 'продажи', 'иностранной', 'валюты', 'долларовом'), 14)]
```

Этап 3 Применение BigRTM

- Проведено тематическое моделирование для выяснения наиболее часто встречающихся слов. Отсюда уже можно будет ориентироваться на категории новостей
- Частота использования слов в обучающей выборке



Этап 3 Применение BigRTM

Наиболее часто встречающиеся слова в классах

```
Topic #0: ['на', 'по', 'года', 'млрд', 'ссылкой', 'будет', 'заявил', 'передает', 'может', 'это']
Topic #1: ['компания', 'млн', '2017', 'как', 'баррель', 'газа', 'год', 'при', 'уровне', 'глава']
Topic #2: ['что', 'до', 'для', 'россии', '2016', 'роснефть', 'пишет', 'после', 'руб', 'банка']
Topic #3: ['за', 'об', 'не', 'сообщает', 'со', 'сша', 'от', 'рублей', 'говорится', 'акций']
Topic #4: ['этом', 'компании', 'из', 'году', 'газпром', 'нефти', 'января', '10', 'долларов', 'президента']
```

Применение BigARTM показало, что новости можно сгруппировать:

- 0 - новости касательно новостей заявительного характера, то есть по сути совпадают с рекламным характером
- 1 - новости касательно биржевой оптовой торговли нефтью и газом, возможно 2017 года
- 2 - новости относительно 2016 года касательно России и ПАО «Роснефть»
- 3 - новости относительно США и акций
- 4 - новости относительно ПАО «Газпром», нефти, и очевидно долларов США

Считаю, что это неудовлетворительное разделение на категории. Соответственно дальнейшая работа была выстроена с предложенными мною ранее классами новостей

Этап 3 Результаты обучения

	Байесовс-кий классифи-катор	Многослойный перцептрон (3 слоя по 8 нейронов)	Linear Support Vector Machine (SGD)	Случай-ный лес	Логисти-ческая регрессия	RNN (LTSM) (epochs = 5, batch_size = 64)	FastText supervised (lr=0.5, epoch=25, wordNgrams=2)
Accuracy	0,568438	0,616747	0,628019	0,632850	0,642512	0,825519	0,993233
Precision							
1	0,88	0,52	0,59	0,64	0,55	0,80	1,00
2	1,00	0,79	0,84	0,83	0,79	0,51	1,00
3	0,56	0,67	0,63	0,62	0,67	0,94	0,99
4	0,33	0,28	0,38	0,70	0,42	0,70	1,00
5	0,00	0,64	0,60	0,83	0,67	0,00	1,00
accuracy							
macro avg	0,55	0,58	0,61	0,72	0,62	0,59	1,00
weighted avg	0,63	0,62	0,63	0,66	0,64	0,80	0,99
Recall							
1	0,10	0,34	0,22	0,24	0,36	0,90	0,99
2	0,03	0,44	0,38	0,43	0,46	0,84	0,96
3	1,00	0,83	0,96	0,95	0,87	0,95	1,00
4	0,03	0,45	0,24	0,18	0,50	0,29	1,00
5	0,00	0,24	0,08	0,14	0,16	0,00	0,99
accuracy							
macro avg	0,55	0,46	0,38	0,39	0,57	0,60	0,99
weighted avg	0,63	0,62	0,63	0,63	0,64	0,83	0,99

Этап 3 Результаты обучения

	Байесовс-кий классифи-катор	Многослой-ный перцептрон (3 слоя по 8 нейронов)	Linear Support Vector Machine (SGD)	Случай-ный лес	Логисти-ческая регрессия	RNN (LTSM) (epochs = 5, batch_size = 64)	FastText supervised (lr=0.5, epoch=25, wordNgrams=2)
Accuracy	0,568438	0,616747	0,628019	0,632850	0,642512	0,825519	0,993233
F1-score							
1	0,19	0,41	0,32	0,35	0,43	0,85	1,00
2	0,06	0,57	0,53	0,56	0,58	0,63	0,98
3	0,72	0,74	0,76	0,75	0,76	0,95	0,99
4	0,05	0,35	0,29	0,29	0,46	0,41	1,00
5	0,00	0,35	0,14	0,23	0,26	0,00	1,00
accuracy	0,57	0,62	0,63	0,63	0,64	0,83	0,99
macro avg	0,20	0,48	0,41	0,44	0,50	0,57	0,99
weighted avg	0,44	0,60	0,57	0,58	0,62	0,80	0,99
Support							
1	143	143	143	143	143	464	464
2	68	68	68	68	68	213	213
3	335	335	335	335	335	1125	1125
4	38	38	38	38	38	155	155
5	37	37	37	37	37	112	112
accuracy	621	621	621	621	621	2069	2069
macro avg	621	621	621	621	621	2069	2069
weighted avg	621	621	621	621	621	2069	2069

Схема обучения модели в концепции SSL

4 этап

ЧАСТИЧНОЕ ОБУЧЕНИЕ С НЕРАЗМЕЧЕННЫМИ ДАННЫМИ

Загрузка датасетов 20170601.csv, 20170901.csv, 20171201.csv и ранее отобранных новостей из бизнеса, экономики и финансов на русском языке

Датасет новостей 41 555 записей

Обучение различным алгоритмам, оценка результатов

Классификация обученными алгоритмами, валидация алгоритмом FastText, отбор верно классифицированных новостей

Датасет верно клас-сифицированных новостей 66 601

Верно классифицирован-ные новости 37 503 записей – 90,2%

Точность 98,6%

Точность составила 98,6%

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана

Этап 4 Частичное обучение с неразмеченными данными (1 итерация)

- Частичное обучение было проведено на неразмеченном новостном датасете состоящим из 37 243 новостей бизнеса
- Была проведена подготовка слов, очищение всех оставшихся новостей, токенизация
- На основе этого было осуществлено предсказание логистической регрессией и валидирование FastText
- Из 37 243 новостей по бизнесу, экономике и финансам на русском языке было успешно предсказано с применением логистической регрессии 29 098 новостей (78%)
- Анализ содержимого показал, что это соответствует истине
- **Частичное обучение**: **Точность составила 97,5%**

Accuracy	0,9750859106529209			
	Precision	Recall	F1-score	Support
1	0,91	0,75	0,82	290
2	0,92	0,90	0,91	136
3	0,98	0,99	0,99	5 324
4	0,86	0,71	0,78	59
5	0,75	0,27	0,40	11
accuracy			0,98	5 820
macro avg	0,88	0,72	0,78	5 820
weighted avg	0,97	0,98	0,97	5 820

Этап 4 Частичное обучение с неразмеченными данными (1 итерация)

Результаты разнесения новостей по классам

```
[['Гендиректора сети IKEA в России Вальтер Каднар сообщил, что ритейлер планирует снизить цены на 15-20% на 1,8 тысячи видов продаваемых товаров, отметив, что на некоторые товары
  'гендиректора сети ikea россии вальтер каднар сообщил ритейлер планирует снизить цены тысячи видов продаваемых товаров отметив некоторые товары цена снижена передает риа amp quot
  '3'],
 ['IKEA не хочет платить российскому бизнесмену Константину Пономарёву 507 млн рублей, которые ему присудил Краснинский суд, и',
  'ikea хочет платить российскому бизнесмену константину пономарёву млн рублей которые присудил краснинский суд',
  '2'],
 ['Производитель сигарет Imperial Tobacco решил закрыть одну из двух своих российских фабрик из-за падения табачного рынка и непростой ситуации в экономике России.',
  'производитель сигарет imperial tobacco решил закрыть одну двух своих российских фабрик падения табачного рынка непростой ситуации экономике россии',
  '3'],
 ['Компания объясняет решение падением российского рынка сигарет и ростом табачных акцизов',
  'компания объясняет решение падением российского рынка сигарет ростом табачных акцизов',
  '3'],
 ['Журналист Шенан Молони заявил об истинной, по его мнению, причине крушения трансатлантического парохода «Титаник».\r\n\r\n\tКак сообщает газета...',
  'журналист шенан молони заявил истинной мнению причине крушения трансатлантического парохода титаник сообщает газета',
  '3'],
 ['О русских  хакерах и реальных киберугрозах Эвелина Закамская говорит с  гендиректором компании Infowatch - Натальей Касперской в студии программы "Мнение".',
  'русских хакерах реальных киберугрозах эвелина закамская говорит гендиректором компании infowatch натальей касперской студии программы мнение',
  '3'],
 ['Расхождение в денежно-кредитной политике между США и Европой приведет к паритету между долларом и евро, считают аналитики банка ING Group.',
  'расхождение денежно кредитной политике сша европой приведет паритету долларом евро считают аналитики банка ing group',
  '3'],
 ['Один из крупнейших в мире производителей компьютерных компонентов, американская компания Intel, планирует инвестировать более 7 миллиардов долларов в завершение строительства заво
  'крупнейших мире производителей компьютерных компонентов американская компания intel планирует инвестировать миллиардов долларов завершение строительства завода полупроводников fab
```

	description	text_sw	label
0	Гендиректора сети IKEA в России Вальтер Каднар...	гендиректора сети ikea россии вальтер каднар с...	3
1	IKEA не хочет платить российскому бизнесмену К...	ikea хочет платить российскому бизнесмену конс...	2
2	Производитель сигарет Imperial Tobacco решил з...	производитель сигарет imperial tobacco решил з...	3
3	Компания объясняет решение падением российског...	компания объясняет решение падением российског...	3
4	Журналист Шенан Молони заявил об истинной, по ...	журналист шенан молони заявил истинной мнению ...	3

Этап 4 Частичное обучение с неразмеченными данными (2 итерация)

- Загрузка оставшихся датасетов и проведение второй итерации
- Из корпуса новостей были отобраны 41 555 неразмеченных новостей бизнеса, экономики, финансов на русском языке
- Из 41 555 неразмеченных новостей успешно определены с применением логистической регрессии 37503 новостей (90%)
- По результатам в обучающую выборку консолидировали 66 601 новостей
- Провели предсказание логистической регрессией
- **Частичное обучение: точность повысилась на один пункт! Точность составила 98,6%**

Accuracy	0,9864124314991367			
	Precision	Recall	F1-score	Support
1	0,94	0,86	0,90	679
2	0,96	0,94	0,95	323
3	0,99	1,00	0,99	12 149
4	0,93	0,89	0,91	141
5	1,00	0,72	0,84	29
accuracy			0,99	13 321
macro avg	0,96	0,88	0,92	13 321
weighted avg	0,99	0,99	0,99	13 321

Результаты

- Разработана программа классификации текстов на маленьких датасетах с помощью частичного обучения и подтверждена применимость данного подхода
- Качество программы достаточно высокое даже при наличии небольшого количества неразмеченных примеров в случае маленьких датасетов
- Все стандартные средства классификации с частичным обучением приводят к удовлетворительным результатам. Правильные результаты классификации были получены для 60-95% классифицируемых новостей
- Среди стандартных средств нельзя выделить одно наиболее подходящее для применения для классификации текстов на маленьких датасетах
- Алгоритм FastText был выбран как валидационный.
- Присутствует субъективность разметки и восприятия новости человеком: то есть новость может быть воспринята каждым человеком по своему. Для одного нейтральная, для другого положительная, для третьего - негативная
- При работе с языковыми конструкциями, написанными на естественном языке, возникают сложности, связанные с уникальностью русского языка, а также с пониманием программы контекста

СПАСИБО ЗА ВНИМАНИЕ

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР МГТУ им. Н. Э. Баумана