# Цифровизация в ВТБ: автоматизация процессов NLP-моделями

Феодосий Котов

Лидер команды «Модели оптимизации РБ»,
Департамент анализа данных и моделирования, Банк ВТБ (ПАО)

# Поиск шаблонов в смс-нотификациях

# Поиск шаблонов в смс-нотификациях

## Цель

Необходимо найти одинаковые шаблоны текстов смс-нотификаций и разметить сообщения по этим шаблонам в зависимости от системы-отправителя, соотнести найденные шаблоны с согласованными шаблонами операторов

**Выборка для пилота:** ~1.5 млрд записей
**Выборка для регулярного расчета:** ~500 млн записей

**Примеры:**

- ***Анна Евгеньевна**, очередной платеж по Автокредиту в размере **15555.77** руб. успешно списан. Благодарим за своевременное погашение!*
- *Вы вошли в интернет-банк Банк X-Онлайн, **16.05.2024 17:12:11**. Счет **«Рога и копыта»** на **1234.56р** выставлен. Оплатите его в https://online.bankX.ru/i/bills*

## Эффект

Автоматизация поиска и оптимизации типовых шаблонов, перевод с смс на push (актуальность возросла при повышении тарифов мобильных операторов)
+Выбор оптимальных условий тарификации
***Экономия нескольких миллионов руб. в месяц***

## Идеи

1. **Кластеризация сообщений по косинусной близости** после применения эмбеддингов слов.
   После первичного анализа стало ясно, что в сообщениях высока доля текста не из шаблонов, который может зашумлять результаты, предпочтение было отдано способам 2-3 + трудозатратные вычисления

2. **Анализ коллокаций и n-грамм**
   **Гипотеза:** слова шаблонов по частоте встречаемости должны превосходить слова не из шаблонов. Например, если в данных много платежей одного и того же ИП («оплата ИП <Фамилия>»), то «Оплата» будет встречаться чаще чем конкретная фамилия.

3. **Морфологический разбор**
   Поиск шаблонов в текстах, слова которых кодируются частями речи

# Поиск шаблонов в смс-нотификациях: тесты

## N-граммы

- Гипотеза: сообщения с одинаковым шаблоном ⇔ одинаковые последовательности слов
- Удалена пунктуация, заменены суммы, даты, время, номера карт. Из русскоязычных сообщений удалены английские символы
- Частотный анализ для последовательностей длины 2-15

Пример выделенных шаблонов:

| |
|---|
| {сумма_в_руб} {номер_карты_или_счета} анастасия м |
| {сумма_в_руб} {номер_карты_или_счета} анастасия п |
| {дата} логин используйте номер |

- Выделяются слишком детальные шаблоны
- Одинаковые сообщения за вычетом различий в именах, выделяются в разные шаблоны
- Сложная интерпретация

## Коллокации

- Гипотеза: сообщения с одинаковым шаблоном ⇔ одинаковые словосочетания
- Удалена пунктуация, заменены суммы, даты, время, номера карт. Из русскоязычных сообщений удалены английские символы
- Модель **FPGrowth**

Пример выделенных шаблонов:

| |
|---|
| мобильное приложение вошли {время} онлайн {дата} втб |
| вошли {время} {дата} втб онлайн |
| втб онлайн |

- Выделяются высокоуровневые шаблоны
- 124 шаблона

# Поиск шаблонов в смс-нотификациях: тесты

## NER и POS-анализ

- Гипотеза: сообщения с одинаковым шаблоном ⇔ общая структура
- Выделение именованных сущностей и разметка слов сообщения частями речи с помощью моделей slovnet

## Результаты

- Хорошо выделяет и детальные, и высокоуровневые шаблоны
- В некоторых случаях разделяет одинаковые шаблоны (разные имена, организации, способы написания чисел и др.)
- Выделено ~100 тыс.+ шаблонов

## Примеры выделенных шаблонов

Вы вошли в мобильное приложение ВТБ-Онлайн, 16.05.2023 09:27:13.
Вы вошли в мобильное приложение ВТБ-Онлайн, 09.05.2023 18:34:51.

PRON VERB ADP ADJ NOUN NOUN PUNCT NUM NUM PUNCT NUM PUNCT NUM PUNCT

Код 555777 для документа №16 от 15.05.2023
Код 666999 для документа №323 от 12.05.2023

NOUN NUM ADP NOUN SYM NUM ADP NUM

Зачисление средств на счет 1111 от АО «РОГА И КОПЫТА N 111». Документ № 9876 на сумму 686500.00 RUB. Остаток 715284.26 RUB.

Зачисление средств на счет 1234 от ООО «ВОСТОРГ». Документ № 1234 на сумму 27692.23 RUB. Остаток 9413246.12 RUB.

PROPN NOUN ADP NOUN NUM ADP ORG PUNCT PUNCT NOUN SYM NUM ADP NOUN NUM PROPN PUNCT NOUN NUM X PUNCT

Зачисление средств на счет 0882 от РЕГИОНАЛЬНОЕ ОТДЕЛЕНИЕ N1111 БАНК X //9999999999. Документ № 123456 на сумму 13316.24 RUB. Остаток 1258045.22 RUB.

Зачисление средств на счет 0409 от РЕГИОНАЛЬНОЕ ОТДЕЛЕНИЕ N 2222 БАНК X //8888888888. Документ № 12346 на сумму 85056.10 RUB. Остаток 401803.63 RUB.

PROPN NOUN ADP NOUN NUM ADP ORG PUNCT NUM PUNCT NOUN SYM NUM ADP NOUN NUM PROPN PUNCT NOUN NUM X PUNCT

# Поиск шаблонов в смс-нотификациях: тесты

## NER и POS-анализ + Предобработка

Предобработка регулярными сообщениями для дедупликации:

- Удаление английских символов
- Объединение тегов организаций, добавление тегов сумм, дат, номеров карт, счетов, ISIN-кодов ценных бумаг, номеров телефонов, ссылок, организаций, не распознанных slovnet
- Использование словаря ФИО пользователей для объединения шаблонов с разными именами

## Примеры соотнесения шаблонов

| Сообщение | Шаблон оператора |
|---|---|
| Счет "ЖКУ ЕИРЦ" на 7692.92р выставлен. Оплатите его в https://online.bankX.ru/i/bills | Счет %w{1,5} выставлен. Оплатите его %w{1,5} |
| Счет «Балашихинская управляющая компания (Московская обл)" на 1345.45р выставлен. Оплатите его в https://online.bankX.ru/i/bills | Нет шаблона |
| Зарплата 273600р Счет*1234 Баланс 568900.25р 09:10 | зарплата %w{1,5} баланс %w{1,5} %d{1,5} |
| Зарплата12843.45р Счет*1234 Баланс 20563.83р 11:50 | Нет шаблона |

## Примеры выделенных шаблонов

Вы не прошли тест для неквалифицированных инвесторов. Пока вы не можете использовать «Акции иностранных компаний». Попробуйте пройти тест снова.

Вы не прошли тест для неквалифицированных инвесторов. Пока вы не можете использовать «Маржинальное кредитование». Попробуйте пройти тест снова.

PRON PART VERB NOUN ADP ADJ NOUN PUNCT ADV PRON PART VERB VERB ORGN PUNCT VERB VERB NOUN ADV PUNCT

Зачисление средств на счет 0123 от АО «Рога и Копыта». Документ № 1111 на сумму 46200.00 RUB. Остаток 46588.76 RUB.

Зачисление средств на счет 9876 от ОТДЕЛЕНИЕ БАНКА X//99999999999. Документ № 111111 на сумму 5486.32 руб. Остаток 205343.96 руб.

PROPN NOUN ADP NOUN VTBACCOUNT ADP ORGN PUNCT NOUN SYM X ADP NOUN SUMRUB PUNCT NOUN SUMRUB PUNCT

## Результаты

- Шаблоны дедуплицированы
- Наилучшее качество разметки шаблонов среди рассмотренных способов
- Объяснимые результаты, доступны исходные сообщения
- Выделено 25 тыс.+ шаблонов
- Добавлено соотнесение шаблонов с шаблонами операторов

# Классификация корневых причин обращений клиентов

# Классификация корневых причин обращений клиентов

## Цель

Необходимо классифицировать причины клиентских обращений, связанных со сберегательными продуктами, по экспертно размеченным категориям

## Эффект

- Автоматизация ручной разметки сотрудниками, сокращение времени на классификацию обращений
- Улучшение качества классификации в сравнении с ручной

## Подход к решению

- Очистка данных:
  - » Очистка названий классов (удаление пробелов, опечаток и т.д.)
  - » Дедупликация классов
  - » Удаление категорий с недостаточным количеством примеров (Итого ~90 классов)
  - » Удаление из текста обращений всего, кроме русских слов
- BGE-M3 на основе архитектуры XLM-Roberta для эмбеддингов + PCA для сокращения размерности + Логистическая регрессия

## Очистка и дедупликация классов

| Изначальный класс | Новый класс |
|---|---|
| Проблема при цифровом подписании | Проблема при цифровом подписании |
| Недостаточно информации. Предварительно проблема на этапе цифрового подписания | |
| Недостаточно информации Ошибка при открытии нс (тайм аут) | Недостаточно информации Ошибка при открытии нс |
| Недостаточно информации Ошибка при открытии нс | |
| Недостаточно информации Ошибки при открытии нс | |

# Классификация корневых причин обращений клиентов

## Результаты

| | F1-micro | F1-macro | F1-weighted |
|---|---|---|---|
| **Train** | 85.4% | 89.8% | 84.5% |
| **Test** | 70.4% | 61.4% | 70.3% |
| **Вся выборка** | 82.4% | 84.5% | 81.7% |

| Текст обращения | Класс, проставленный вручную | Класс, предсказанный моделью |
|---|---|---|
| Клиент хотел открыть Вклад ''Новые деньги''. Вчера обращался на ГЛ. Сотрудник сначала предложил открыть вклад в режиме звонка, клиент не смог, уточнил может ли он на следующий день обратится в офис для открытия, на что сотрудник ответил, '' да конечно''. Сегодня клиент обратился в офис, клиенту не могут открыть вклад | Некорректная консультация/обслуживание | Новые деньги. Некорректная консультация/обслуживание |
| У клиента в разделе мои договоры не отображается заявление на открытие накопительного счета ,подписывала в онлайн | Прочее | Проблема при цифровом подписании |
| При оформлении вклада указывается, что у клиента недостаточно средств. Суммы для открытия достаточно | Недостаточно информации, ошибка при открытии нс | Проблема при открытии, недостаточно средств |
| Клиент произошла миграция с Банка А на Банк Б, денежные средства были переведены в рамках миграции и сотрудник сделал заявление на открытие договора вклада на сумму , на срок  дней под ставку %. Вклад не открылся, а денежные средства поступили и денежные средства у клиента лежали без процентов. Из-за того что вклад не открылся, клиент потерял в совокупности  рублей и более | Прочее | Проблема при миграции из Банка А |

# Анализ текстов коммуникации клиентов с банком

# Анализ текстов коммуникации клиентов с банком

## Цель

Разработать решение, позволяющее на регулярной основе выявлять информацию о мнении клиентов о сберегательных продуктах на основе транскрибаций звонков КЦ

## Эффект

Оперативное выявление проблем пользователей и улучшение клиентского опыта через анализ тем взаимодействия в чатах и звонках

## Пример

<cl>да<eos>

<op>ожидайте пожалуйста на линии сейчас загружается миту-две<eos>
благодарю вас за длительное ожидание нет у вас списание по кредиту не произошло вас по счету но которые внесению производили имеется блокировка расходных операции
а поэтому сумма не отображается по по вашему счету вам необходимо будет обратиться по данному вопросу в отделение банка для того чтобы могли снять блокировку по счету и также<eos>

<cl>все я<eos>

## Проблемы с качеством в текстах:

1. Специальные символы <op>, <eos>, <cl>
2. Речь оператора (между символами <op>)
3. Опечатки
4. Иностранные слова
5. Ненормативная лексика
6. Имена, географические названия, данные клиентов

# Рассмотренные подходы

| Модель | Краткое описание | Выводы | Категория |
|---|---|---|---|
| Коллокации 1-ый этап | N-граммы (2, 3 и 4-граммы) генерируют словосочетания, выбираются наиболее значимые по метрикам | Малоинформативно: часто употребимые словосочетания словосочетания, напр. «горячая линия, плохо слышно» | Извлечение коллокаций |
| Коллокации 2-ой этап | Фильтрация стоп-слов – удалены неинформативные слова (из топ-частых и топ-редких), исключены дубликаты n-грамм с перестановкой слов | Малоинформативно: часто употребимые словосочетания словосочетания, напр. «доставленные неудобства, абонент выключить» | Извлечение коллокаций |
| Иерархические коллокации | Для каждой топовой n-граммы отбираются тексты, содержащие её, и строятся новые n-граммы | Малоинформативно: общие словосочетания, напр. «забыть запутаться, единственный вариант». Требуется составление нового списка стоп-слов и дополнительная очистка текста | Извлечение коллокаций |
| Коллокации с использованием текста оператора | Включена речь оператора в тексты | | |
| Attention на текстах | Модель bge-m3 – анализирует важность слов через attention-веса последнего слоя | | Готовые нейросетевые кодировщики |
| Коллокации на кластерах 3-ий этап | Кластеризация KMeans на эмбеддингах bge-m3, удалены опечатки, имена и географические названия из текстов | Коллокации отражают проблематику каждого кластера, напр. «оплата налог, списание денежный средство» | Кластеризация |
| Суммаризация кластеров | Модель Qwen – для каждого кластера KMeans отбираются тексты, ближайшие к центроиду по косинусному расстоянию | Суммаризация дополнила результаты по коллокациям, уточнив описание проблематики каждого кластера, напр. «Клиенты обращались по вопросам обслуживания и предоставления информации в банке» | Готовые нейросетевые кодировщики |

# Итоговое решение. Кластеризация

Исправление опечаток

Удаление спец. символов, английских слов, слов из 1 буквы

Удаление стоп-слов

Исключение текстов длиной < 6

Токенизация и приведение в нормальную форму

## Кластеризация

- LLM bge-m3 (XLM_Roberta) для построения эмбеддингов
- UMAP для понижения размерности эмбеддингов
- Для кластеризации Kmeans и HDBScan

## Построение коллокаций

1. Строятся 3, 4, 5-граммы с различными размерами окон
2. Удаляются дубликаты n-грамм
3. Отбираются итоговые коллокации с наибольшей частотой

# Итоговое решение. Суммаризация

## Модель для суммаризации

Использовалась языковая модель – Qwen2.5-7B-Instruct. Вход модели – 10 ближайших к центру кластера текстов

## Prompt модели

{"role": "system", "content": "Ты - Qwen, виртуальный ассистент, обученный для анализа обращений клиентов банка. Твоя задача - анализировать списки клиентских сообщений и кратко формулировать их основные темы в одном предложении. Используй формулировку: 'Клиенты обращались по вопросам...'."}

1. "Ниже приведены примеры обращений клиентов банка.\n"
2. "Проанализируй их и определи, какие темы их беспокоят. "
3. "Ответь кратко в одном предложении, используя формат:\n"
4. "'Клиенты обращались по вопросам...' \n\n"

# Результаты

## КЛАСТЕР 1 – Оплата и налогообложение

Топ 3-грамм:

- ('оплата', 'налог', 'выставить')
- ('приходить', 'оплата', 'налог')
- ('налог', 'выставить', 'оплатить')

Суммаризация:

- Клиенты обращались по вопросам оплаты налогов, возникновения неизвестных счётов и проблем с личным кабинетом

## КЛАСТЕР 2 – Судебные приставы и банкротство

Топ 3-грамм:

- ('проходить', 'процедура', 'банкротство')
- ('судебный', 'пристав', 'снять')
- ('судебный', 'пристав', 'арестовать')

Суммаризация:

- Клиенты обращались по вопросам арестов счетов, блокировки средств и получения информации о причинах арестов

## КЛАСТЕР 3 – Долговые обязательства и возвраты

Топ 3-грамм:

- ('сверх', 'лимит', 'задолженность')
- ('списание', 'денежный', 'средство')
- ('списать', 'денежный', 'средство')

Суммаризация:

- Клиенты обращались по вопросам списаний с их счетов, проблем с мобильным приложением банка и задолженностей по кредитам

## КЛАСТЕР 4 – Движение денежных средств и справки

Топ 3-грамм:

- ('движение', 'денежный', 'средство')
- ('заказать', 'справка', 'госслужащий')
- ('заполнять', 'справка', 'доход')

Суммаризация:

- Клиенты обращались по вопросам получения справок, закрытия и активации счетов, а также проверки открытых счетов в банке

# Спасибо

Феодосий Котов

fkotov@vtb.ru