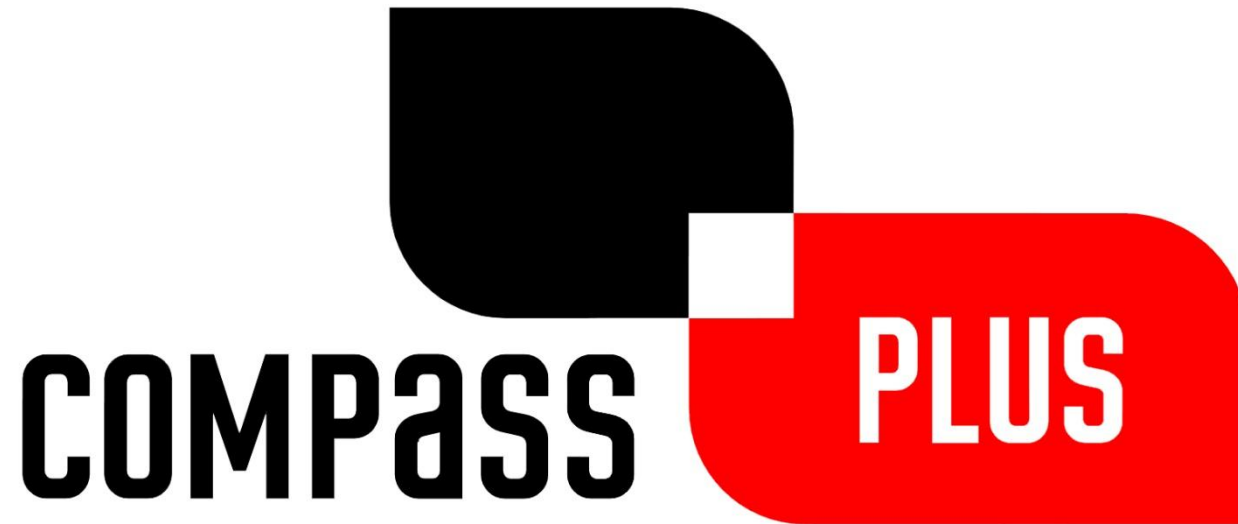


Разработка системы классификации обращений клиентов в техподдержку приложения



Над проектом
работали:
Кононов Н.А.
Пчелин А.В.
Шевченко А.Е.

Постановка задачи

- Разработать интеллектуальную автоматизированную систему для классификации пользовательских обращений в техническую поддержку приложения, включающую в себя подсистему предварительной обработки исходных сообщений, а также модель для классификации сообщений.

Цели и задачи

- Провести исследование предметной области – задачи обработки данных на естественном языке
- Проанализировать исходный набор данных
- Разработать концептуальную схему работы системы
- Исследовать способы выполнения предварительной обработки исходных текстов и реализовать их
- Исследовать существующие линейные модели классификации и провести отбор моделей с наилучшим результатом классификации
- Для отобранных наилучших моделей провести полное исследование доступных для изменения параметров, подобрать наилучшие параметры и произвести анализ результатов обучения

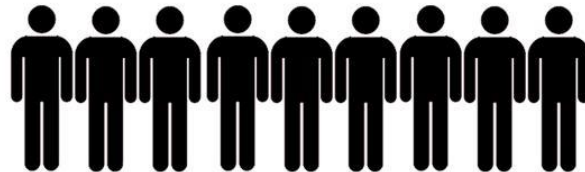
План разработки

- Создание программного модуля предобработки обращений
- Тестирование и доработка модуля предобработки
- Обучение модели-классификатора
- Тестирование классификатора и анализ результатов, поиск методов их улучшения
- Тестирование всего программного комплекса (предобработка в тандеме с классификатором)
- Доработка модулей по результатам тестирования

Концепция и предполагаемые

результаты

Клиенты



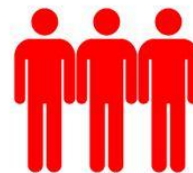
Классификатор



Запрос классифицирован
Клиенту отправляется
заготовленный ответ
по теме

Если обращение не удалось
классифицировать,
оно переводится на оператора

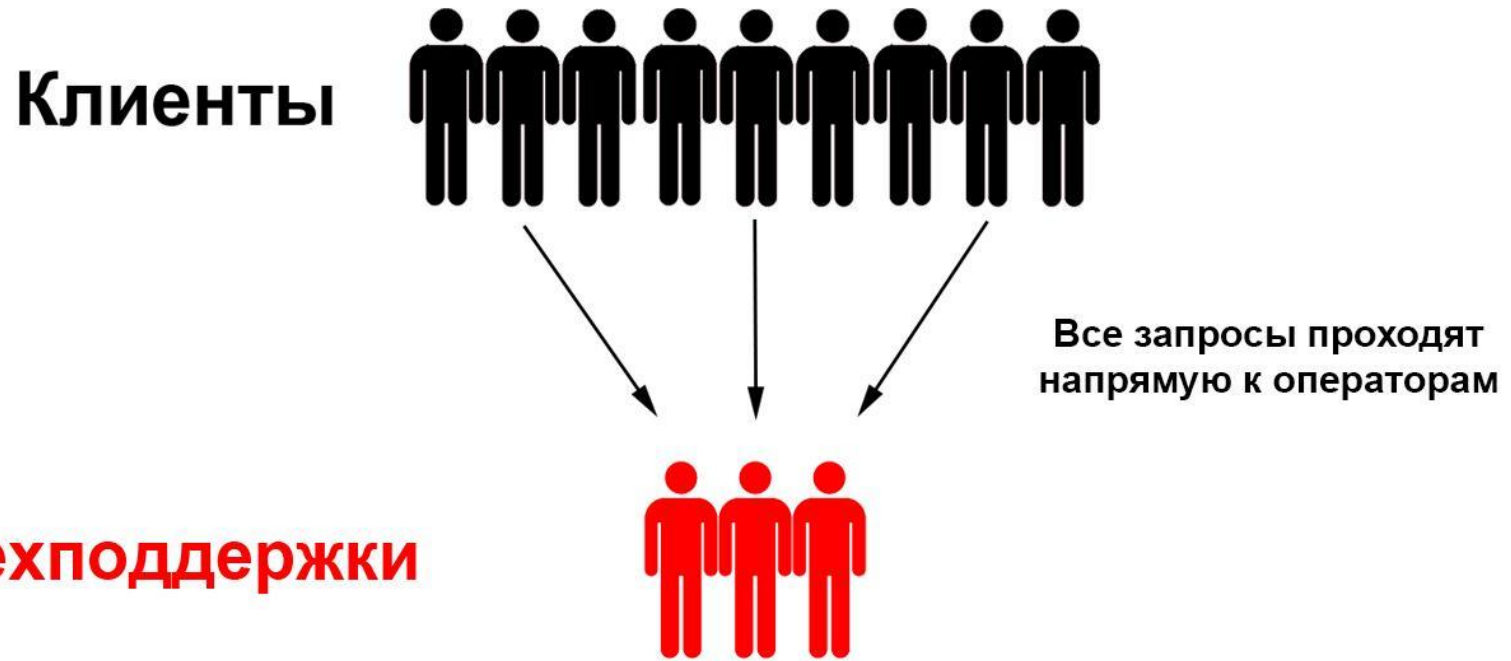
Операторы техподдержки



Если ответ не устроил клиента,
обращение переводится
на оператора



Концепция и предполагаемые результаты



Исходные данные

- Формат:
 - Таблица Excel с размеченными данными
- Поля:
 - Header – заголовок
 - Body – само обращение
 - Class – класс обращения
 - Group – группа обращения
- Объем:
 - 2504 записи, 2315 уникальных
- Самое частое обращение встретилось 15 раз

Проблемы датасета

- Малый объем данных
- Дисбаланс классов
- Множество слов с одной и более опечатками
 - рекаищиты, тноефона, прогоаммоц
- По ошибке склеенные и разделенные слова
 - оплаченданные, удо бнее
- Обилие словоформ и синонимов
 - моби кеш, мобикэш, mobicash, moby cash
- Даты и время, суммы, и др. числа которые не несут смысла
 - 21.09.2019, 4576 рублей, в 14:15, версия ОС 7.18.2020
- Служебные части речи

Методы решения проблем

- Для исправления опечаток в словах был использован модуль PyEnchant, а также ряд составленных своими силами словарей замен
- С помощью регулярных выражений даты, время и суммы были заменены словами DATE, TIME, SUM, остальные числа были удалены
- С помощью модуля PyMorphy был произведен морфологический разбор – удалены слова служебных частей речи, остальные слова приведены к начальной форме

Полученные результаты

- Изначальное количество словоформ:
 - Число
- Количество словоформ после предобработки:
 - число
- Число уникальных обращений после предобработки:
 - 2176
- Ключевые термины были приведены к одному образцу (н-р tobicash, cashback)
- Даты, время, суммы и др. числовые данные были либо заменены словами, либо исключены из обращений

Анализ разбиения данных на классы

