Сергей АФАНАСЬЕВ и др.

Три года назад в банке «Ренессанс Кредит» было выделено направление моделирования, связанное с обработкой естественного языка (Natural Language Processing, NLP). Пока ИТ-подразделение занималось разработкой платформы для внедрения нейронных сетей, NLP-моделистов нужно было чем-то занять. Расскажем о трех кейсах, которые банк реализовал без существенных ИТ-доработок и инвестиций в ИТ-инфраструктуру, и о том, как будет выглядеть новая NLP-платформа.

Сергей АФАНАСЬЕВ, *КБ «Ренессанс Кредит», вице-президент, начальник управления статистического анализа*
Диана КОТЕРЕВА, *КБ «Ренессанс Кредит», руководитель направления R&D*
Константин СТАРОДУБ, *КБ «Ренессанс Кредит», консультант направления R&D*

Обработка естественного языка: три практических кейса банка «Ренессанс Кредит»

Сейчас «Ренессанс Кредит» — это 75 миллионов кредитных заявок, 9 миллионов кредитов, выданных физическим лицам за последние 5 лет, 15 миллионов клиентов. То есть у нас достаточно данных, чтобы разрабатывать модели различной степени сложности, развивать Data Science проекты.

Команда Data Science в банке небольшая — всего 14 человек. Направление Data Science централизованное: та часть задач, которая касается разработки моделей, аккумулирована в одном подразделении. Если же говорить о полном цикле моделирования — от постановки задачи до эксплуатации, — то многие функции распределены по разным направлениям банка. Заказчики — CRM, подразделения рисков, сбора задолженности, противодействия мошенничеству, маркетинга, финансов и т.д. (на данный момент 14 направлений). То есть источниками доменной экспертизы выступают большей частью бизнес-заказчики. ИТ-подразделения помогают нам внедрять модели в промышленную эксплуатацию, разрабатывают инфраструктуру для Data Science проектов. В части регуляторных моделей есть отдельное подразделение валидации. Задачи моделирования

Обработка естественного языка: три практических кейса банка «Ренессанс Кредит»

разделены на три группы: классические модели, регуляторные модели и задачи исследований и разработки (R&D).

Большинство моделей мы строим с помощью внутренней библиотеки AutoML, которая позволяет нам автоматизировать большую часть процесса моделирования. Направления моделирования представлены на рис. 1.

Рисунок 1

Направления моделирования

Три года назад было решено выделить в рамках R&D направление NLP-моделирования. Это связано с тем, что банк ежемесячно собирает сотни тысяч клиентских звонков, диалогов в чатах, письменных обращений клиентов — данные, на которых можно строить модели и эмбеддинги для обогащения классических банковских моделей.

В итоге мы разделили задачи NLP-направления на две группы:

1. Обогащение текстовыми эмбеддингами уже существующих банковских моделей: моделей взыскания, CRM-моделей, скоринговых карт. Это требует минимума ресурсов Data Science, поскольку можно обучить одну семантическую модель и обогатить этими эмбеддингами много классических банковских моделей.

Первые результаты оказались многообещающими. Модели сбора задолженности «приросли» на 3% Gini за счет текстовых эмбеддингов, в CRM прирост составил 2% Gini, причем для роботизированных звонков. Для звонков операторов эффект может быть еще лучше за счет более глубоких и осмысленных диалогов с клиентами.

2. Сервисные NLP-задачи — автоматизация ручной обработки звонков, текстов и т.д. Здесь есть крупный блок задач от контактного центра: например, маршрутизация запросов в программе Mail Stream, выявление дополнительных тематик для чат-бота, анализ телефонных разговоров с клиентами. Также с NLP-задачами к нам обращаются другие подразделения.

Для внедрения всех этих NLP-моделей нужна промышленная среда, платформа, куда будут внедряться нейросетевые технологии. Классические банковские среды, такие как кредитный конвейер или CRM-система, не предназначены для внедрения нейросетей, поэтому мы разрабатываем такую платформу самостоятельно ресурсами банка. Сейчас проект находится на финальной стадии разработки, платформа будет запущена в феврале 2023 г. Общая концепция архитектуры представлена на рис. 2.

В верхнем ряду показаны источники данных — это звонки клиентов, чаты, письменные обращения клиентов и информация другого типа, включая готовые табличные агрегаты. Из этих данных мы получаем текстовые эмбеддинги, которые хранятся в единых витринах, и с помощью этих эмбеддингов обогащаем существующие банковские модели либо строим отдельные NLP-модели для автоматизации сервисных задач.

На разработку NLP-платформы требуются инвестиции и время. И пока ИТ-подразделение разрабатывает такую платформу, NLP-моделисты могут заниматься другими полезными для банка задачами, для которых не требуются дорогие инфраструктурные доработки. Далее мы расскажем о трех таких задачах, в которых с помощью NLP-моделирования была успешно автоматизирована работа других подразделений.

Кейс 1. Поиск клиентов, работавших в покинувших российский рынок компаниях

Первый кейс связан с событиями весны 2022 г., когда Россию начали покидать западные бренды. Перед риск-менеджерами встал вопрос: что делать с клиентами, которые работают в этих компаниях? Скорее всего, они окажутся безработными и по ним начнет расти просрочка. Задача заключалась в том, чтобы найти таких клиентов в портфеле и поставить их на контроль. Нужно было сопоставить два списка:

1) список компаний, ушедших с российского рынка (публиковался на сайте sravni.ru);

2) данные о работодателях из кредитных заявок.

Обработка естественного языка: три практических кейса банка «Ренессанс Кредит»

Рисунок 2

Архитектура платформы DL/NLP

При этом список компаний на сайте sravni.ru публиковался в английской транслитерации, а в клиентских анкетах работодатели были указаны в русской транслитерации в свободном формате, то есть один и тот же работодатель мог иметь разные названия в разных кредитных заявках. Портфельные менеджеры попросили нас автоматизировать процесс сопоставления списков с помощью NLP-инструментов, чтобы повысить точность.

Мы построили фонетическую модель. На ее реализацию ушло две недели с учетом сжатых сроков. Общая концепция показана на рис. 3.

На входе было два списка: список компаний с сайта sravni.ru и клиентские данные о работодателе. Эти тексты подавались на NLP-препроцессинг. Мы их обрабатывали с помощью готового фонетического модуля transliterate и собственных разработанных правил для фонетики. На выходе мы сравнивали обработанный текст с помощью расстояния Левенштейна, получая клиентов, у которых названия работодателей похожи на те, что были в списке на sravni.ru. Финальный список проверяли портфельные менеджеры.

Интересный факт: когда мы проанализировали рисковые показатели по таким клиентам, риски по ним оказались не выше, чем в среднем по портфелю, а в некоторых сегментах даже ниже, то есть мы получили так называемый эффект adverse selection.

Кейс 2. Выявление событий операционного риска на основе текстов бухгалтерских проводок

Со следующей задачей к нам обратились коллеги из подразделения операционных рисков. В рамках Положения Банка России от 08.04.2020 № 716-П банк должен выявлять события операционного риска и регистрировать их в базе СУОР. Одним из источников таких событий являются тексты бухгалтерских проводок. Риск-менеджер ежедневно вычитывает эти проводки, классифицирует их на рисковые и не рисковые и заносит первые в базу событий. Коллеги обратились к нам с запросом: можно ли автоматизировать эту процедуру с помощью NLP-инструментов? В результате мы построили NLP-классификатор (рис. 4). На полную автоматизацию процесса ушло примерно полтора месяца.

Общая схема выглядит так. На входе у нас есть размеченная выборка с текстами бухгалтерских проводок. Каждая проводка имеет флаг 0 (не рисковая) или 1 (рисковая): ручная разметка, которую раньше делал риск-менеджер. Затем тексты проходят через стандартный NLP-препроцессинг:

Рисунок 3

Модель для поиска ушедших брендов среди работодателей клиентов

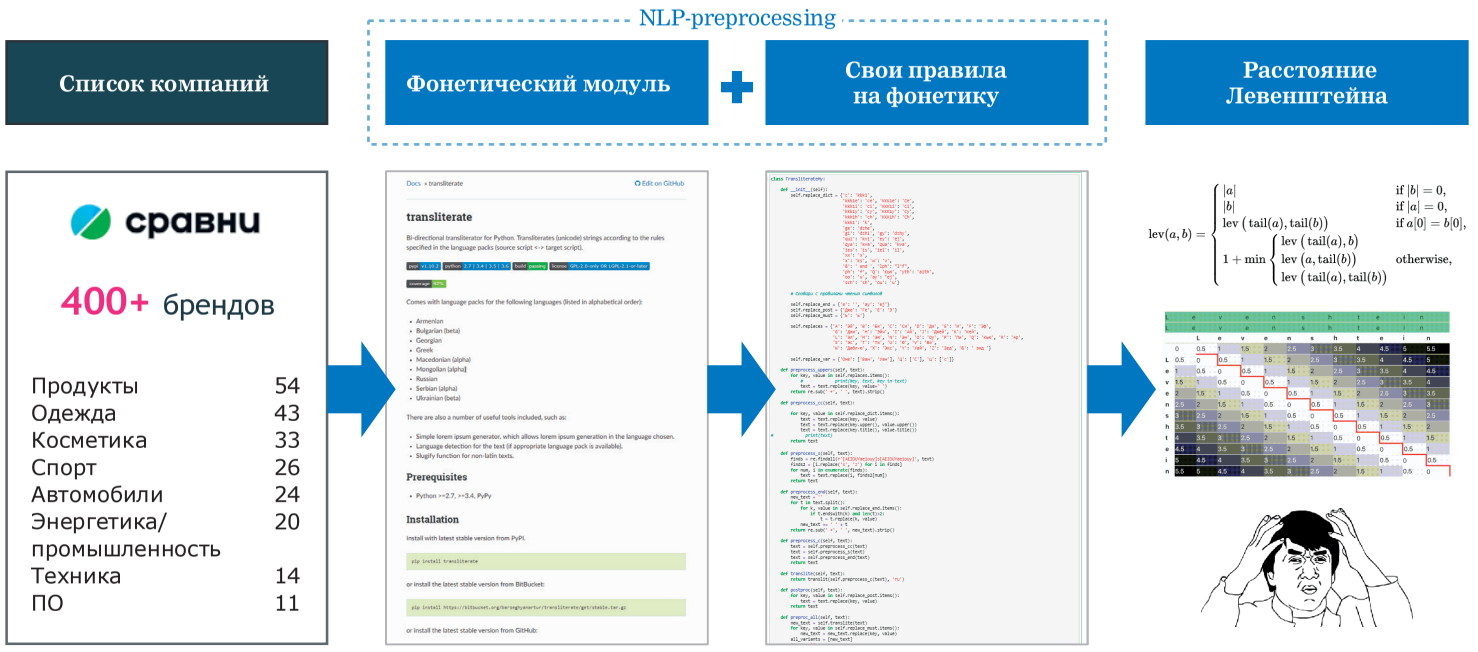
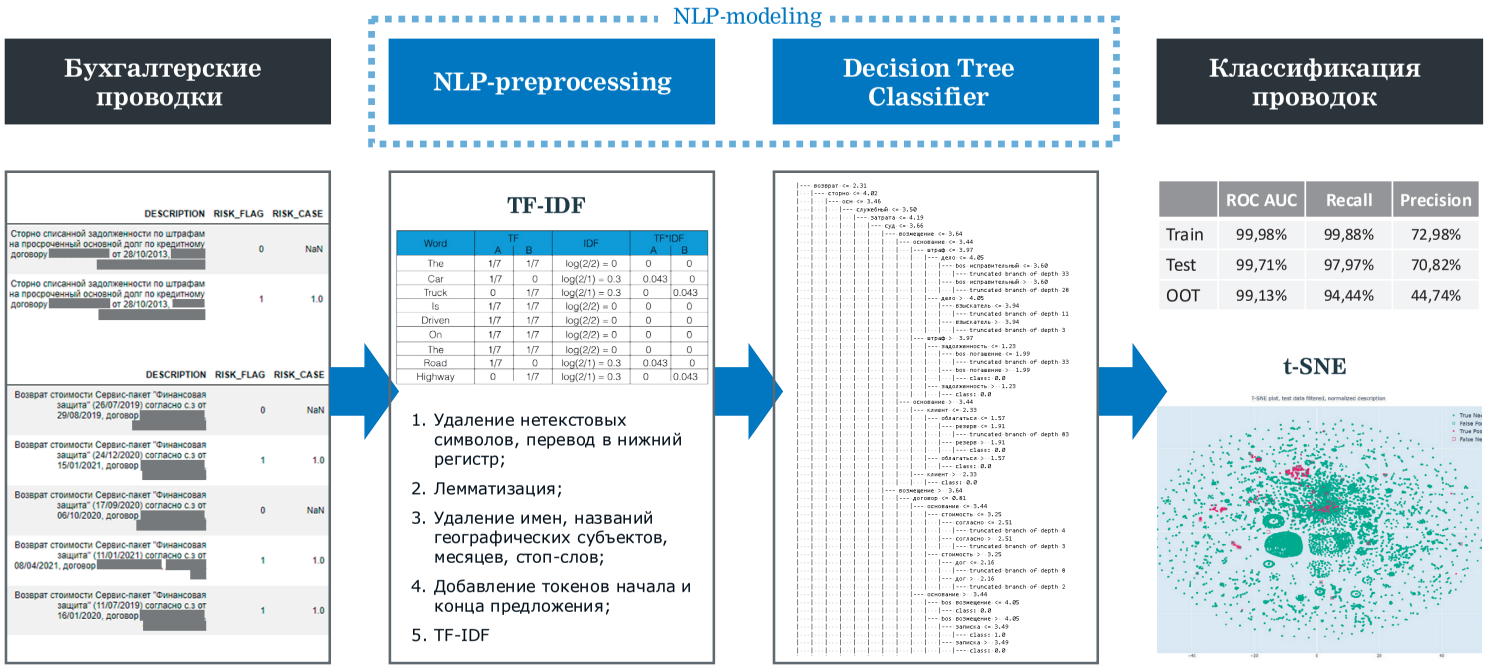


Рисунок 4

Выявление событий операционного риска на основе текстов бухгалтерских проводок



— удаление нетекстовых символов, перевод в нижний регистр;
— лемматизация;
— удаление имен, названий географических субъектов, месяцев, стоп-слов и т.д.;
— добавление токенов начало и конец предложения.

Далее на предобработанных данных строится TF-IDF представление. На TF-IDF представлении обучаются бинарные классификаторы. Мы обучали логистическую регрессию и деревья решений, поскольку выборка небольшая.

Качество финального классификатора оказалось не хуже, чем при ручной обработке. Таким образом, удалось полностью автоматизировать работу сотрудников подразделения операционных рисков по анализу бухгалтерских проводок, не потеряв в точности классификации.

Кейс 3. Автоматизация работы с мошенническими фирмами по списанию долгов

Последний кейс наиболее сложный. Это автоматизация обработки писем от фирм-«раздолжнителей» — мошеннических «юридических» фирм, которые завлекают банковских клиентов предложениями списать долги, провести реструктуризацию, оформить банкротство и т.д. Как правило, реальной пользы такие компании клиентам не приносят. Они готовят стандартные письма от имени клиентов и направляют их в банк, а в банке операторы контактного центра пытаются выявить такие письма среди общего потока и готовят шаблонные ответы.

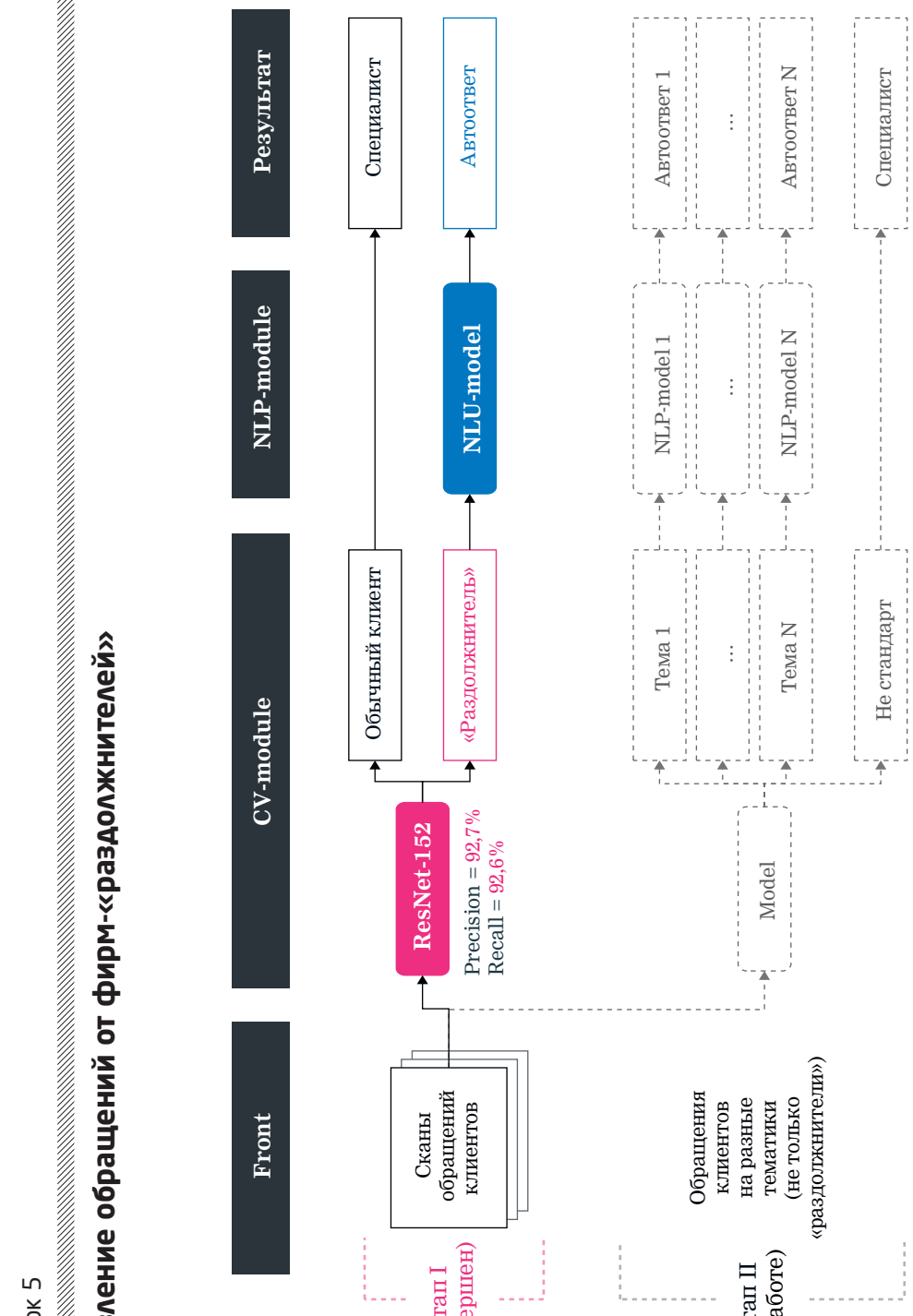
Коллеги из контактного центра обратились к нам с запросом об автоматизации этого процесса. Здесь понадобился более широкий стек технологий: не только NLP, но и компьютерное зрение (CV), и оптическое распознавание символов (OCR). На автоматизацию ушло около полугода. Общая схема приведена на рис. 5.

На входе мы имеем PDF-сканы. С помощью сверточной нейронной сети ResNet они классифицируются на два типа: письма от обычных клиентов и письма от фирм-«раздолжнителей». Письма от обычных клиентов идут по стандартному банковскому процессу на анализ специалисту контактного центра, а письма от «раздолжнителей» попадают в автоматический NLP-модуль, где делается препроцессинг, определяется тематика письма и готовится автоматический ответ.

Техническая реализация NLP-модуля оказалась наиболее сложной. Выглядит он так (рис. 6). Когда ResNet — сверточная нейронная

Рисунок 5

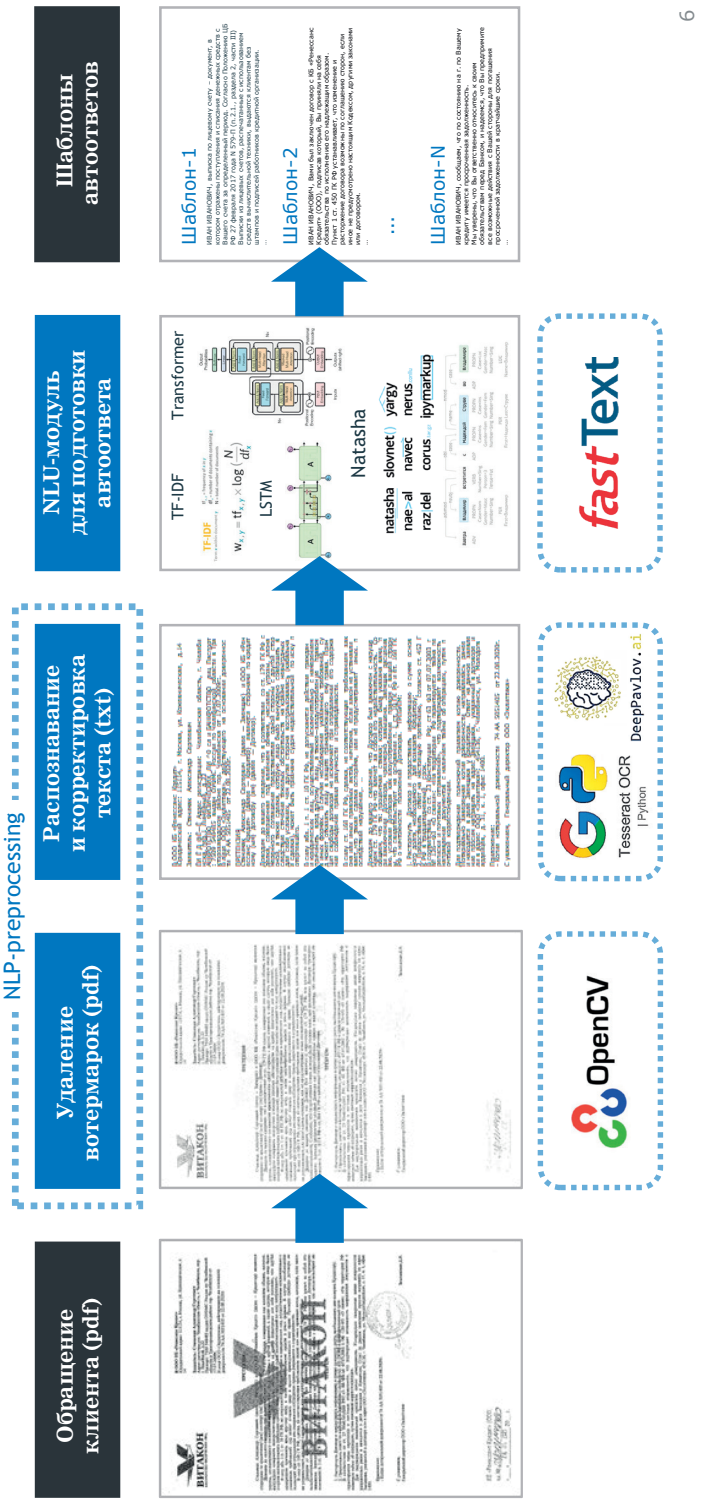
Выявление обращений от фирм-«раздолжнителей»



Обработка естественного языка: три практических кейса банка «Ренессанс Кредит»

Рисунок 6

NLP-модуль для обработки писем от фирм-«раздолжнителей»



сеть — понимает, что письмо от «раздолжнителя», PDF-скан подается на NLP-препроцессинг, где с помощью открытых библиотек OpenCV, Tesseract и DeepPavlov производятся чтение PDF, перевод в черно-белую гамму, увеличение контрастности, бинаризация, сглаживание, ротация, удаление водяных знаков, OCR, корректировка текста, очистка, удаление именованных сущностей. Далее текст подается в модуль понимания естественного языка (NLU), который определяет тематику письма. На этом этапе мы с помощью библиотеки fastText получали эмбеддинги, на которых обучали несколько мультиклассовых моделей (Multinomial Classification) с помощью нейронных сетей LSTM и Transformer. Также пробовали классический подход — обучение градиентного бустинга на TF-IDF представлении. Лучший результат получился на рекуррентной нейронной сети LSTM. Для поиска клиентов использовалась готовая реализация NER-модели (распознавание именованных сущностей) с помощью открытой библиотеки Natasha. После классификации текстов, получения интентов (тематик письма) и поиска клиентов автоматически создается шаблон ответа.

Таким образом, получилось автоматизировать весь процесс обработки писем от «раздолжнителей», причем MVP мы смогли запустить на локальном сервере без серьезных ИТ-доработок.

Однако автоматизация данного процесса подвержена проблеме «противодействия брони и снаряда» — о таком подходе становится известно самим «раздолжнителям» и они пытаются подстроиться под процесс. Так, если раньше можно было определить (например, по наличию печати), что это письмо от юридического лица, то сейчас «раздолжнители» становятся умнее и меняют шаблоны своих писем. Для решения этой проблемы нужна регулярная поддержка со стороны операторов контактного центра, разбирающих часть сообщений вручную (контрольная группа). В кейсах с «раздолжнителями» и выявлением событий операционного риска у нас реализован процесс с контрольными группами, позволяющий нам перерабатывать и обновлять модели.