НГТУ НЭТИ

Технологии,
которые работают.

Медведев Михаил Александрович

м.н.с., руководитель проектов, Представительство ЦК НТИ «Технологии доверенного взаимодействия» (НГТУ НЭТИ)

nstu.ru

Алгоритмы машинного обучения для семантической контент-фильтрации и защиты конфиденциальной информации

Семантическая контент-фильтрация с помощью ML?

- Семантический анализ не просто поиск ключевых слов, а понимание смысла и контекста текста.
- Машинное обучение (ML) – использование моделей, способных «учиться» отличать конфиденциальную/запрещенную информацию от нейтральной.
- Защита конфиденциальности – выявление и сокрытие персональных данных, коммерческих секретов и других чувствительных сведений.

Преимущества использования семантической контент-фильтрации с помощью ML?

- Адаптивность к новым формам выражения;
- Меньше ложных срабатываний по сравнению с простыми фильтрами на ключевых словах;
- Возможность учитывать контекст.

Основные проблемы

Сложность контекста – многообразие языковых форм, неоднозначных терминов, культурных нюансов.

Объемы данных – постоянный рост информации требует высокопроизводительных решений и больших наборов данных для обучения.

Правовые и этические аспекты – соблюдение законов о персональных данных и обеспечение прозрачности принятия решений ИИ.

Точность и ложные срабатывания – найти баланс между блокировкой необходимой информации (ложноположительные результаты) и пропуском опасной (ложноотрицательные результаты).

Глобальная проблема: Мультикультурный и многоязычный контент

Компаниям приходится фильтровать сообщения на разных языках, учитывая культурные нюансы и смешанную лексику.

Одна и та же фраза может иметь разные значения в разных странах или профессиональных сообществах.

Сложность распознавания скрытых смыслов (сарказм, жаргон, эвфемизмы) возрастает в условиях многоязычной среды.

Модель угроз. Внешние факторы:

Кибератаки и целенаправленные взломы

- Обход традиционных фильтров. Злоумышленники шифруют вредоносный код или используют малоизвестные уязвимости. Семантические фильтры на базе ML могут помочь выявлять аномалии в содержании, даже если подпись или сигнатура атаки еще не известна
- Фокус на конфиденциальной информации. Хакеры нацелены на конкретные документы или утечки персональных данных. ML-модели могут определять паттерны, характерные для попыток получить доступ к таким данным

Вредоносные вложения

- Динамическое определение угроз. ML-модели можно обучить выявлять новые тенденции и схемы, используемые злоумышленниками, без необходимости вручную обновлять правила

Фишинг

- Анализ текста веб-страниц. Семантические фильтры могут обнаруживать поддельные адреса, ссылки или нетипичные формулировки, указывающие на попытку мошенничества

Модель угроз. Внутренние факторы:

Намеренные действия сотрудников

- Кража или продажа чувствительных данных с целью личной выгоды
- Целенаправленная отправка конфиденциальной информации за пределы организации
- Саботаж при увольнении или конфликте с руководством

Ненамеренные утечки

- Неправильная настройка доступа к общим папкам и документам
- Слабое понимание внутренней политики безопасности

Человеческие уязвимости

- Подверженность фишингу и социальной инженерии (сотрудник может неосознанно открыть «дыру» в системе)
- Сложные или неоднозначные инструкции по работе с данными, приводящие к ошибкам

Основные нормативные акты РФ, регулирующие защиту данных и информационную безопасность:

Федеральный закон «О персональных данных» от 27.07.2006 № 152-ФЗ

Федеральный закон «О защите детей от информации, причиняющей вред их здоровью и развитию» от 29 декабря 2010 г. № 436-ФЗ

Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27 июля 2006 г. № 149-ФЗ

Федеральный закон «О безопасности критической информационной инфраструктуры Российской Федерации» от 26.07.2017 N 187-ФЗ

Алгоритмы машинного обучения для семантической контент-фильтрации. Традиционные методы:

Классификация на основе ключевых слов и регулярных выражений:

- Ключевые слова:
  - Создание списка «стоп-слов» (запрещенных выражений, терминов)
  - Автоматический поиск совпадений в тексте
  - Простота настройки – достаточно дополнить файл со словами, и фильтр заработает
- Регулярные выражения:
  - Использование паттернов поиска для шаблонных данных:
    - Форматы номеров банковских карт (16 цифр, разделенных пробелами или дефисами)
    - Адреса электронной почты (шаблоны вида *@mail.ru)

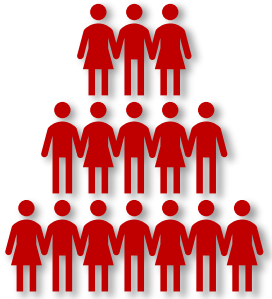
Недостатки:

1. Отсутствие учета контекста:
   - Система «видит» только точные совпадения, игнорируя смысл.
   - Возможны ложные срабатывания (например, слово «паспорт» в безобидном контексте).
   - Пропуск тонких смыслов, сарказма, эзопова языка, когда конфиденциальная информация зашифрована намеками.
2. Требуется постоянное ручное обновление правил:
   - Новые формулировки, сленг, иносказания, когда каждый раз нужно дополнять и корректировать списки.
   - При большом объеме данных усложняется поддержка и проверка на актуальность.
   - Безоперационная пауза: если ответственные лица забыли обновить «стоп-слова», фильтр перестает реагировать на свежие угрозы.
3. Невозможность обрабатывать неоднозначность:
   - Слова-омонимы, смешение языков, сокращения, когда фильтр часто не справляется.
   - Переход на другой язык или простой «звездочкой» в слове («п*спорт») обходят фильтр.

Алгоритмы машинного обучения для семантической контент-фильтрации. Современные методы:

1. Глубокое обучение: предобученные эмбеддинги (BERT, GPT)

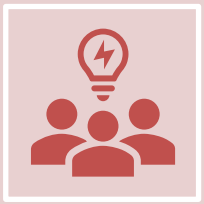
- Модель «учится» на огромных датасетах, формируя контекстные представления слов (эмбеддинги).
- При решении прикладной задачи (фильтрация, классификация) модель дообучается на конкретном датасете.



2. Методы полусупервизированного и обучение без учителя

- **Проблема разметки**: часто не хватает размеченных данных для всех сценариев (профессиональные термины, конфиденциальная лексика).
- **Полусупервизированный подход**:
  - Использует небольшое количество размеченных данных + большое количество неразмеченного материала.
  - Модель извлекает закономерности и постепенно улучшает свои прогнозы.
- **Обучение без учителя**:
  - Фокус на поиске скрытых шаблонов.
  - Полезно для предварительной сегментации данных, система группирует документы по сходной тематике, а эксперт уточняет категории.

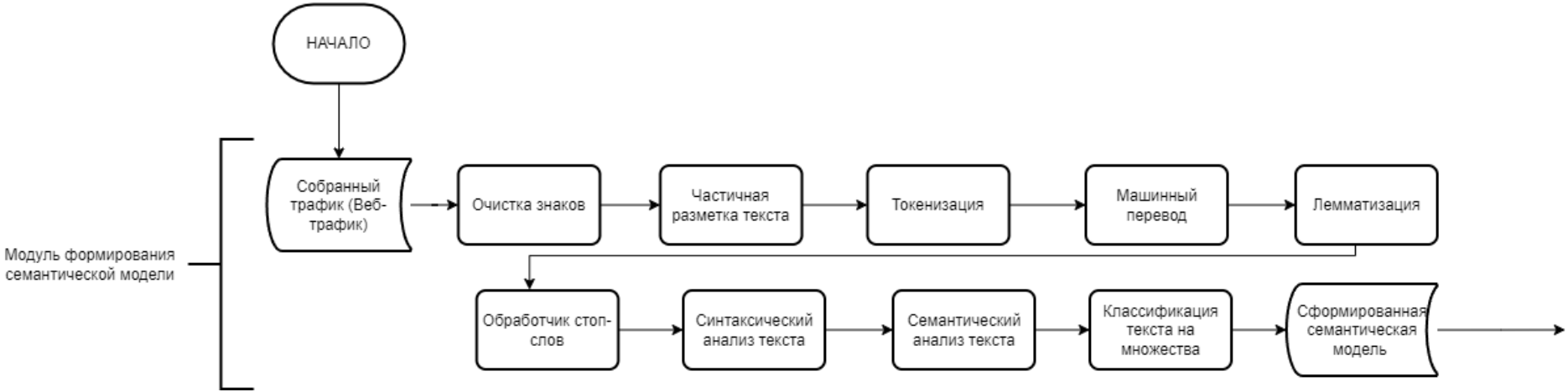
Наши разработки



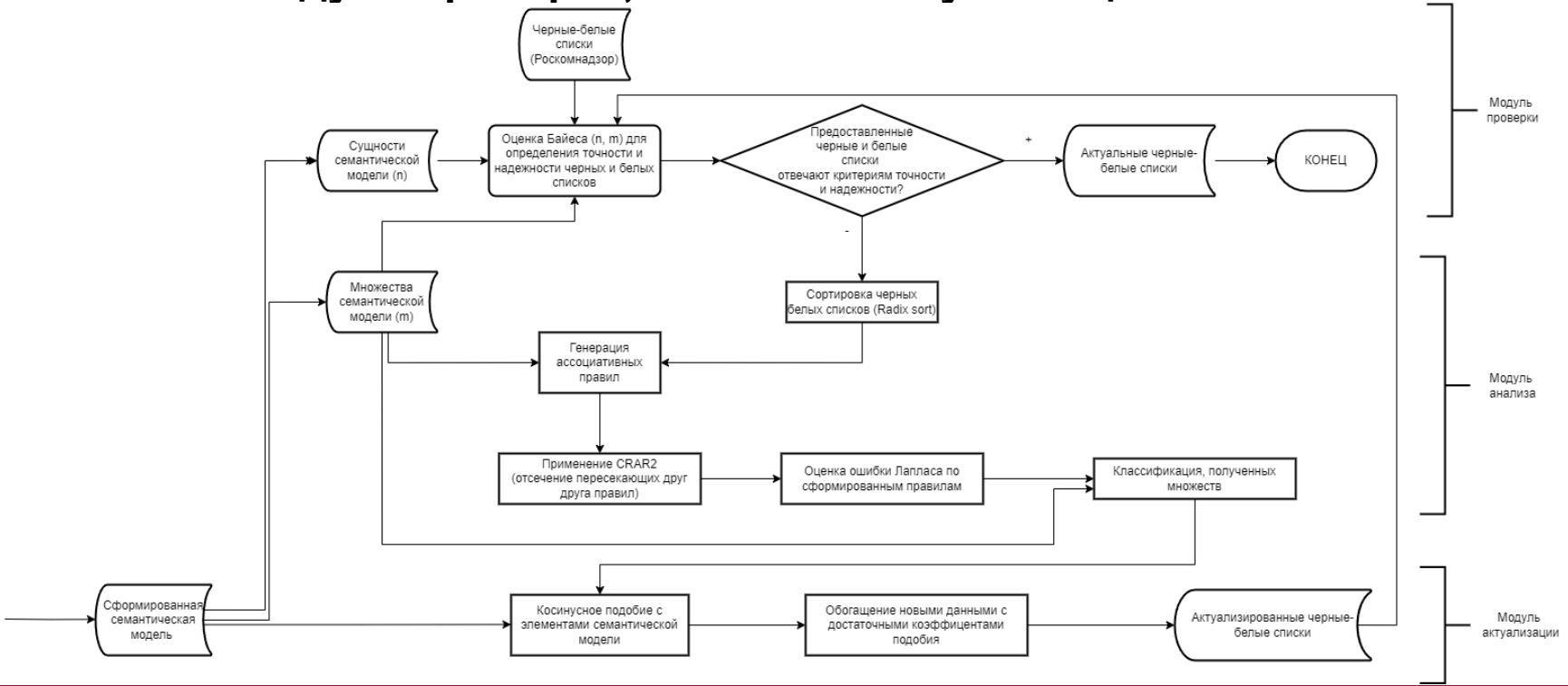
Реализован НИР «Разработка методики проверки, анализа и актуализации белых и черных списков фильтрации на основании семантической модели контента», в разработке находится ПО, реализующего модели и методики проверки, анализа и актуализации белых и черных списков фильтрации на основании семантической модели контента в рамках выполнения проектов программы развития ЦК НТИ «Технологии доверенного взаимодействия».

Основная цель разработки: повышение точности семантической контент-фильтрации входящего трафика с использованием машинного обучения.

Взаимодействие математических моделей на основании разработанных методик проверки, анализа и актуализации белых и черных списков фильтрации на основании семантической модели контента. Модуль формирования семантической модели:



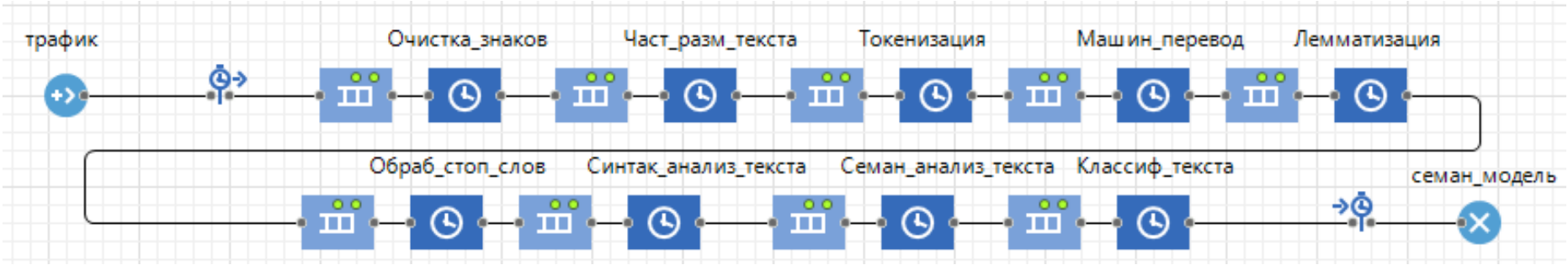
Взаимодействие математических моделей на основании разработанных методик проверки, анализа и актуализации белых и черных списков фильтрации на основании семантической модели контента. Модули проверки, анализа и актуализации:



Результаты моделирования

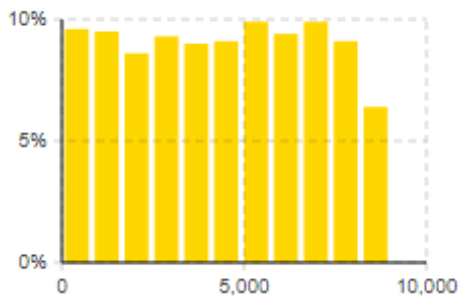
- Дискретно-событийная имитация (AnyLogic):
  - Среднее время обработки в разных модулях: ~96–4500 сек (зависит от сложности и объема трафика). Особенностью такого разброса во времени является неравномерное труднопредсказуемое количество поступающих заявок, так как заявки, неуспешно прошедшие проверку, отправляются на дополнительный анализ и актуализацию, который может проводиться многократно в отношении одной заявки.
  - Линейное распределение времени, высокое соответствие теоретическим моделям.
- Эффективность:
  - Стабильная работа, предсказуемые задержки, готовность к реальной эксплуатации.
  - Минимизация ложноположительных и ложноотрицательных результатов.

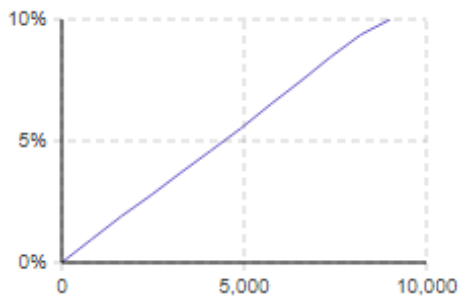
Результат моделирования процесса формирования семантической модели в программной системе AnyLogic



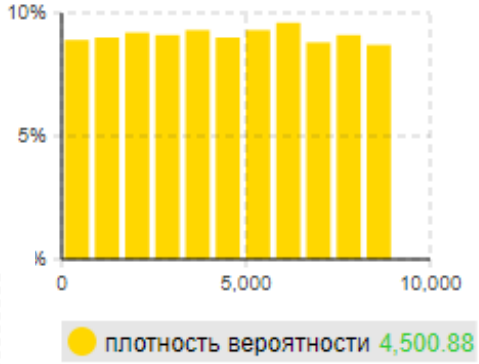
Распределение времени поступления трафика

Распределение времени поступления сущностей и множеств семантической модели и данных из модуля актуализации



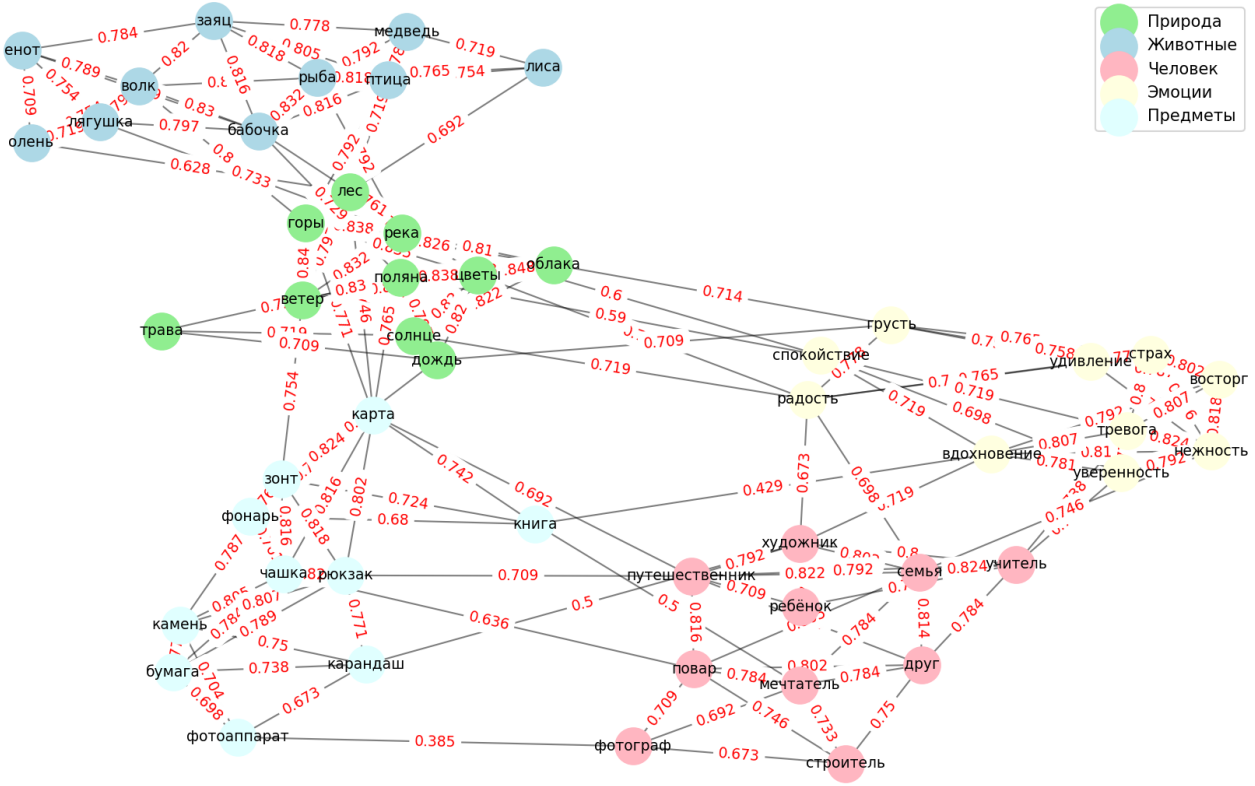


Распределение времени поступления множеств семантической модели и данных из модуля проверки



Распределение времени поступления множеств семантической модели и данных из модуля анализа

Пример построения семантической модели контента



Численные результаты работы модели

- Точность классификации ~90%
- Устойчивость к ошибкам ~85%

- Модуль формирования семантической модели: среднее время нахождения агента ~96,31с. (диапазон от 8,38-171,38с.)
- Модуль проверки: среднее время нахождения агента ~4392,17с. (диапазон от 1,6-8747,88с.)
- Модуль анализа: среднее время нахождения агента ~4500,88с. (диапазон от 3,8-9011,86с.)
- Модуль актуализации: среднее время нахождения агента ~4523,6с. (диапазон от 2,3-9034,5с.)

Планы и перспективы развития

1

Расширение языковых моделей – поддержка нескольких языков, учет культурных и региональных особенностей при формировании черных/белых списков.

2

Интеграция с корпоративными системами. Доработка API и обеспечение совместимости с имеющимися SIEM/CRM/ERP-платформами.

НГТУ НЭТИ

Технологии,
которые работают.

Медведев Михаил Александрович

м.н.с., руководитель проектов, Представительство ЦК НТИ «Технологии доверенного взаимодействия» (НГТУ НЭТИ)

nstu.ru

Алгоритмы машинного обучения для семантической контент-фильтрации и защиты конфиденциальной информации