Ю.Ю. Михайлов, Д.С. Топкасов, Т.И. Паюсова

Тюменский государственный университет, г. Тюмень

УДК 004.912

ОСНОВНЫЕ АСПЕКТЫ ПРОЕКТИРОВАНИЯ КАТЕГОРИЗАТОРА ЭКСПЕРИМЕНТАЛЬНОГО ПРОТОТИПА DLP-СИСТЕМЫ

Аннотация. В статье рассмотрен процесс проектирования текстового категоризатора DLP-системы, основанном на машинном обучении.

Ключевые слова: машинное обучение, анализ текстов на естественных языках, DLP-системы, утечки данных, логистическая регрессия.

Введение

В современном постоянно подвергающемся влиянию цифровизации мире ежегодно растет количество утечек информации, в том числе конфиденциальной. Большинство этих утечек являются внутренними, что говорит о переносе акцента в области информационной безопасности с защиты от внешних угроз к контролю собственных сотрудников, которые имеют прямой доступ к информационным системам и хранилищам данных. Также, как можно увидеть на рисунке 1, наблюдается значительное увеличение объема скомпрометированных персональных данных и финансовой информации в последние годы [1].

Рис. 1. Объем данных, скомпрометированных в результате внутренних утечек (по данным InfoWatch)

Одним из методов, предупреждающим такие неправомерные действия, с использованием санкционированного доступа, являются Data Loss Prevention-системы, или коротко DLP. Суть их работы состоит в анализе внутреннего трафика компании, отслеживании активности работников и их действий на рабочем месте, чтобы в случае утечки, можно было оперативно реагировать и осуществлять работу по противодействию распространению данных.

Исследовательский интерес заключается как раз в отслеживании трафика, а именно, сообщений, которые отправляют работники, с использованием корпоративной электронной почты. В информации, полученной из этих материалов, могут содержаться данные, позволяющие прямо или косвенно определить, имеют ли действия сотрудника корыстный умысел. Дальнейшая реакция по противодействию, действительно, может помочь избежать негативных последствий.

В рамках данной статьи будет рассматриваться проектирование классификатора, работающего с сообщениями корпоративной электронной почты. Основная задача классификатора – предоставление данных о потенциальных утечках информации, своего рода попытка предсказать такие утечки. Полученные результаты будут использоваться в других модулях DLP-системы. Например, если сообщения сотрудника были опознаны системой как потенциально опасные, то системе следует отметить такого пользователя, чтобы в дальнейшем следить за отклонениями в модели его поведения.

Категоризация сообщений корпоративной почты

Категоризатор работает с семантическим анализом информации, содержащейся в текстах сообщений. Например, если сообщения носят агрессивный характер в адрес компании или начальства, то это говорит о сильном недовольстве этого работника, что, в свою очередь, может сигнализировать о его желании в дальнейшем уволиться или нанести вред компании. В этом случае необходимо внимательно относиться к файлам и документам, которые данный сотрудник может отправлять. Аналогичные действия стоит предпринимать, если по семантике текстов можно понять, что

сотрудник уже ищет новую работу (может быть, даже направляет резюме) или обсуждает в переписке финансовые вопросы, что может говорить о возможных сделках по продаже разработок или персональных данных клиентов и сотрудников.

То есть, в данной части модуля будет классифицироваться информация по угрозе, которую несет то или иное сообщение. Мы условно выделили их следующим образом:

– «Агрессивные высказывания» – тексты, семантика которых имеет ярко выраженный негативный эмоциональный окрас;

– «Поиск другой работы» – сообщения, которые могут указывать на планы перехода сотрудником на другое место работы;

– «Обсуждение финансовых вопросов» – письма, содержащие в себе решение финансовых вопросов, которые могут относиться к продаже данных (но нужно обязательно учитывать специфику работы сотрудника, отдела или даже организации, так как функционирование такой категории проверки текстов сообщений, поступающих, например, из бухгалтерии будут приводить к ошибкам второго рода);

– «Стандартные сообщения» – сообщения, написанные в официально-деловом стиле, не несущих угроз, описанных выше;

– «Бессмысленные сообщения» – которые могут оказаться некоторой шифровкой, используемой для осуществления корыстных целей.

Несмотря на то, что с точки зрения функционирования DLP-системы необходимости в настолько подробной категоризации нет (важно лишь то, представляет ли сообщение угрозу или нет), было решено оставить категории как есть, не объединяя их, так как такое решение позволит исключать определенные категории с целью снижения количества ошибок второго рода и для осуществления более эффективного расследования инцидентов информационной безопасности. Таким образом, важным аспектом при разделении на данные категории является условие их функциональности (возможность убрать определенные категории для определенных отделов).

Авторами статьи также была проведена работа по категоризации семантики текстов комментариев в социальных сетях с применением модели логистической регрессии с использованием алгоритма стохастического градиентного спуска. Результаты работы легли в основу данного исследования: был разработан категоризатор, указывающий насколько тот или иной текст комментария относится к определенной категории. Категории, по которым классифицировались тексты, а также процесс работы категоризатора продемонстрированы на рисунке 2.

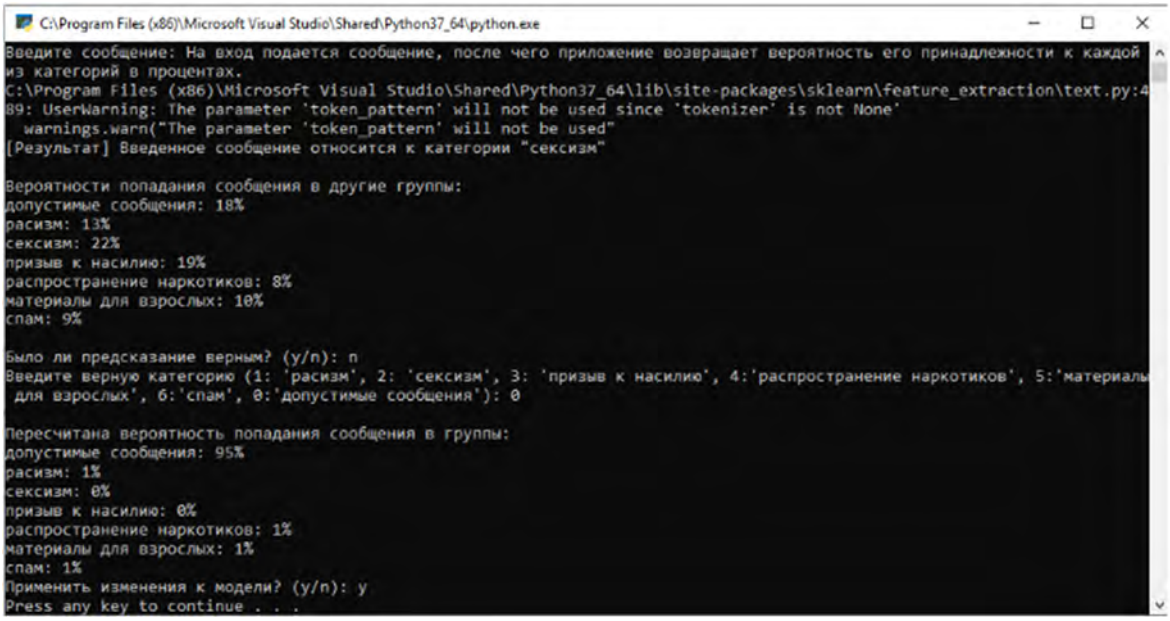

Рис. 2. Пример работы категоризатора

Эта же схема будет применяться и в рамках разработки текущего прототипа DLP-системы, как наиболее качественный из испробованных и проанализированных нами методов.

Для правильного распределения текстов, поступающих в категоризатор, по категориям, были выделены следующие методы предобработки:

- очистка текста от стоп-слов;
- приведение текста к нормальной форме;
- векторизация текста с использованием хеширования.

Выбор модели машинного обучения

Модели машинного обучения можно разделить на два вида [2]: модели обучения с учителем и обучения без учителя. Для категоризации было решено применить первый вид моделей, так как категории заранее определены.

К моделям обучения с учителем относятся:

1. Скрытая Марковская модель (HMM)
2. Условные случайные поля (CRF)
3. Maximum Entropy (MaxEnt)
4. Модель опорных векторов (SVM)
5. Деревья решений (DT)
6. Наивный байесовский алгоритм
7. Deep learning
8. Логистическая регрессия

Для реализации поставленных задач из всех вышеназванных моделей была выбрана модель логистической регрессии с использованием алгоритма стохастического градиентного спуска. Данная модель характерна тем, что позволяет прогнозировать вероятность возникновения события путем его сравнения с логистической кривой. Пример графика логистической кривой и процесс ее корректировки новыми данными можно увидеть на рисунке 3. Эта особенность полезна при категоризации текстов, поскольку позволяет определить принадлежность сообщения к тому или иному классу в процентном соотношении. Модель логистической регрессии наиболее популярна при классификации данных из-за своей универсальности и легкости трактования [3].

Алгоритм стохастического градиентного спуска, в свою очередь, позволяет выполнять корректировку модели, используя новые данные. Это позволяет осуществлять более гибкую настройку категоризатора при возникновении ошибок первого и второго порядка.

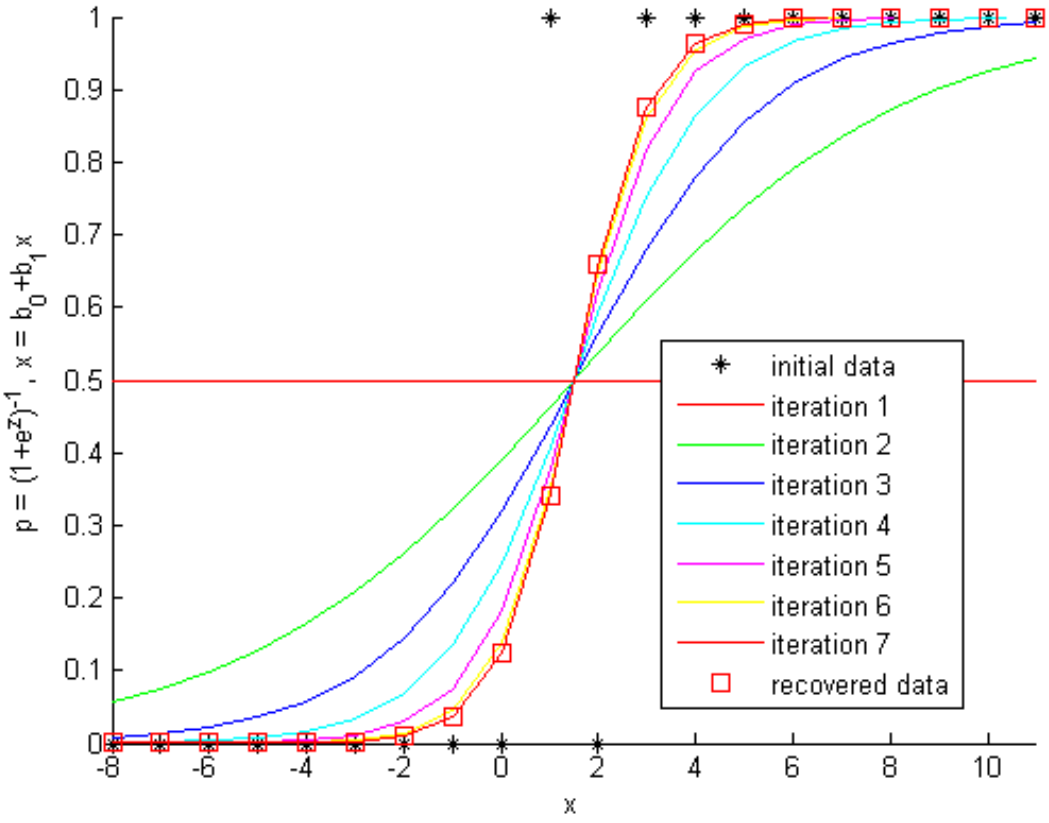

Рис. 3. График логистической функции

Заключение

Таким образом, в данной статье приведены основные аспекты разработки категоризатора для экспериментального прототипа DLP-системы. Был рассмотрен поэтапный процесс для проектирования такого продукта: выделены необходимые шаги по предобработке анализируемых текстов; обозначены категории, по которым они будут делиться по результату семантического анализа; выбрана модель машинного обучения и оптимальный алгоритм для корректировки модели.

Система, реализованная с учетом вышеуказанных аспектов, позволит не просто выявлять факт утечки, а предсказывать возможность ее совершения пользователями, которые пишут сообщения, относящиеся к потенциально опасным категориям.

СПИСОК ЛИТЕРАТУРЫ

1. Утечки данных организаций по вине внутреннего нарушителя. Сравнительное исследование. 2013-2019 гг., – 2020 – URL: https://www.infowatch.ru/sites/default/files/analytics/files/InfoWatch_Analytical_Report.pdf (дата обращения: 30.05.2021).

2. Khan W. et al. A survey on the state-of-the-art machine learning models in the context of NLP // Kuwait journal of Science. – 2016. – Т. 43. – № 4.

3. Лукинов В.Л. Численно устойчивый вероятностный классификатор логистической регрессии // Труды Международной конференции «АПВПМ». 2019. URL: https://cyberleninka.ru/article/n/chislenno-ustoychivyy-veroyatnostnyy-klassifikator-logisticheskoy-regressii-1 (дата обращения: 30.05.2021).