Методы выявления конфиденциальной информации в неструктурированных данных

Г.В. Гарбузов, С.В. Дворянкин

***Аннотация*— Статья посвящена методам выявления (классификации, распознавания) конфиденциальной информации в неструктурированных данных, представленных в виде файлов, передаваемых по каналам связи или хранимых в файловых ресурсах. Акцент на неструктурированные данные обусловлен тем, что, с одной стороны, именно неструктурированные данные представляют наибольший интерес для злоумышленника с точки зрения содержания конфиденциальной информации коммерческого предприятия (коммерческой тайны, ноу-хау), и, в то же время, именно неструктурированные данные плохо поддаются анализу используемыми сегодня сигнатурными алгоритмами и правилами на основе регулярных выражений, которые используются в современных средствах защиты от утечек информации.**

***Ключевые слова*— неструктурированные данные, конфиденциальная информация, коммерческая тайна, утечки информации, системы защиты от утечек информации.**

I. Введение

Вопрос защиты нематериальных активов коммерческих предприятий в настоящее время крайне актуален. Как показано в [1], именно нематериальные активы составляют более 90% активов современной компании, а в наукоемких отраслях, например, таких, как фармацевтика, дизайн, проектирование, консалтинг и подобные, эта цифра может быть еще выше. Необходимо также учитывать, что в подавляющем большинстве случаев, такие активы представлены в виде неструктурированных данных, что привносит свои сложности в процесс их идентификации [2], при этом объемы неструктурированной информации растут втрое быстрее, чем объемы структурированных данных [3]. Неструктурированная информация представлена в виде сообщений электронной почты или отдельных текстовых, графических, аудио- и видео файлах различных форматов. Их невозможно обработать методами, применяемыми для анализа структурированной информации, поскольку ценность такой информации определяется не позицией ячейки в базе данных или атрибутом модели данных, а её содержанием, предполагающим её осмысление. В таких условиях защита нематериальных активов должна основываться на механизмах её достоверной и эффективной классификации в целях предотвращения её утечек. Для решения этих задач в мире сегодня повсеместно используются системы защиты от утечек (Data Loss Prevention, DLP), использующие различные технологии анализа информации, передаваемой по каналам связи и сохраняемой в файловых ресурсах предприятия [4]. Анализу существующих технологий и способам их совершенствования посвящена данная статья.

II. Существующие методы анализа данных

Задача классификации конфиденциальной информации представляет собой решение задачи построения функции $fR(Dt)$, которая среди множества всех данных $Dt=\{KD,O\}$ находит конфиденциальную информацию PD, принадлежащую множеству атрибутов $A=\{a1,\ldots,ak\}$, где k – количество распознаваемых атрибутов конфиденциальных данных, при этом в настоящей работе мы сделаем особый упор на атрибуты неструктурированной информации, как представляющей наибольшую сложность при классификации существующими методами.

Метод распознавания конфиденциальной информации, реализующий функцию fR, может быть построен на основе:

- ручного труда, при которой описание модели данных и их классификацию выполняет эксперт;
- лингвистических машинных методов анализа информации, в т.ч. классификация на основе правил (rule-based) [5];
- статистических машинных методов анализа информации;
- методов с использованием нейросетей [6].

Рассмотрим их подробнее.

A. Метод ручной классификации

Анализ выполняется человеком (экспертом, владеющим предметной областью) без применения средств автоматизации, а значит метод обладает следующими недостатками:

- недостаточная эффективность. Ручная классификация требует значительного времени и усилий, особенно при обработке больших объемов данных. Особенно актуальной эта проблема является для

Статья получена 13 ноября 2025.

Г.В. Гарбузов – аспирант кафедры информационной безопасности факультета информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации; ORCID: http://orcid.org/0009-0008-7717-1488 (e-mail: g.garbuzov@mail.ru)

С.В. Дворянкин – д.т.н., профессор, профессор кафедры информационной безопасности факультета информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации; Московский гуманитарный лингвистический университет; ORCID: http://orcid.org/0000-0001-6908-0676 (e-mail: s_dvm@mail.ru)

организаций, обрабатывающих и хранящих большие объемы данных, а также ведущих активную переписку;

• субъективизм и ошибки. Одни и те же операции могут выполняться разными исполнителями, обладающими различной квалификацией. Человек может не заметить некоторые важные детали ввиду своей неосведомлённости или усталости, что может привести к неполному или неточному анализу. Точность классификации и количество ошибок прямо зависят от квалификации исполнителя, которая, в свою очередь, определяет величину затрат на выполнение классификации;

• низкая масштабируемость. Ручная классификация может быть сложной для масштабирования на большие объемы данных. Кроме того, большие объемы или потоки данных при ручной классификации становятся проблемой, поскольку эти данные должны быть доставлены и сохранены в месте проведения анализа;

• высокая стоимость. Ручная классификация предполагает интеллектуальный ручной труд квалифицированных специалистов, поэтому требует значительных затрат на оплату их труда. В случае классификации большого объема данных и высоких требований к точности классификации, затраты на оплату труда специалистов, выполняющих ручные операции, могут быть очень значительными.

B. Лингвистические методы анализа информации

К лингвистическим методам относятся морфологический анализ на ключевых словах, метод регулярных выражений и использование шаблонов. Суть всех перечисленных способов сводится к анализу текстового слоя на предмет выявления в нем либо заранее определенных слов или словосочетаний, последовательностей, строковых шаблонов, присущих тем или иным конфиденциальным данным (например, номер телефона, номер паспорта, водительского удостоверения, ИНН, СНИЛС и т.д.) или цифровых меток специальной формы или содержания.

Основным преимуществом данной группы технологий является простота её применения и универсальность. Она легко автоматизируется и может использоваться как для контроля каналов утечки информации, передаваемой по каналам связи, отправляемой на печать или копируемой на съемные носители, так и для поиска конфиденциальных документов, хранящихся в файловых ресурсах и на компьютерах пользователей. Простота применения обусловлена тем, что конфиденциальные документы не требуется как-либо специально обрабатывать и технология начинает работать сразу после включения правила в DLP системе и сразу по всем контролируемым каналам.

Главным недостатком морфологического анализа является чрезвычайно высокая её зависимость от качества настройки. Например, для использования метода ключевых слов требуется подчас экспертное владение терминами бизнеса, в противном случае использование ключевых слова из общих словарей («конфиденциально», «совет директоров», «портфель продуктов» и т.д.) приведет к росту уровня ложноположительных срабатываний на крайне высокий уровень, поскольку почти каждый проанализированный документ или сообщение будут распознаваться как содержащие конфиденциальную информацию. Методам классификации информации на основе регулярных выражений и шаблонов свойственны общие недостатки:

• высокий уровень ложноположительных срабатываний, особенно в случаях, если данные имеют необычные форматы шаблоны или регулярные выражения не были правильно определены. Например, номер паспорта состоит из шести цифр, но сформированный таким образом шаблон будет срабатывать каждый раз, когда в документе встретится 6 цифр подряд;

• быстрая деградация. Поскольку состав, атрибуты и цели обработки конфиденциальной информации в современной коммерческой организации могут меняться, правила также должны регулярно пересматриваться, причем чем динамичнее бизнес, тем чаще должен производиться пересмотр;

• ограниченность применения. В методах на основе регулярных выражений и шаблонов используются классификация конфиденциальной информации осуществляется на основе определенных правил. Однако, эти правила могут быть ограниченными и не учитывать все возможные вариации конфиденциальной информации в конкретной организации. Это может привести к неправильной классификации конфиденциальной информации определенных видов;

• ограниченная гибкость. Алгоритмы на основе регулярных выражений и шаблонов могут быть менее гибкими, чем другие подходы, такие как использование ключевых слов или машинное обучение. Они могут иметь трудности с классификацией новых типов конфиденциальной информации или адаптацией к изменяющимся требованиям

C. Статистические методы анализа информации

Ядром метода является массив заготовленных цифровых отпечатков (хэшей) определенных заранее цифровых объектов (файлов), с которым впоследствии сравнивается хэш, рассчитанный для анализируемого объекта. Поскольку поиск осуществляется по совпадению отпечатка, технология не является вероятностной и позволяет с исключительной точностью выявлять заранее определенные фрагменты цифровых объектов (например, только подпись в документе) в анализируемом документе, таблице или схеме, и даже аудио- или видеозаписи. Изменяя размеры и количество фрагментов, появляется возможность управлять степенью соответствия анализируемого цифрового объекта отпечатку имеющегося образца, что, в свою очередь, предоставляет возможность создавать дифференцированные правила, в которых описаны разные действия для разной степени совпадения анализируемого объекта с образцом (например, при 100% совпадении блокировать передачу объекта, при 50% информировать офицера безопасности и т.д.).

Главными недостатками статистического метода являются значительная ресурсоемкость (для каждого анализируемого объекта нужно рассчитать его хэш и сравнить его с эталонным) и низкая универсальность,

поскольку для надежной защиты нематериальных активов придется заготовить цифровые отпечатки (хэши) для каждого файла, при этом при каждом изменении файла хэш придется пересчитывать. Это делает данный метод хорошим инструментом для точечной защиты или, например, проведения расследований инцидентов резонансных утечек, но непригодным для промышленного применения на потоке данных.

D. Метод анализа информации с использованием нейросетей

Вопросы использования нейросетей для защиты конфиденциальной информации сегодня активно исследуется во всем мире [7], [8], [9]. Суть метода заключается в применении предварительно обученной нейросети для решения задачи классификации конфиденциальной информации [10], которая сводится к задаче машинного обучения по распознаванию именованных сущностей, при этом нейронные сети показывают хорошие результаты при решении задачи распознавания именованных сущностей, выраженных атрибутами конфиденциальной информации.

Для обучения нейросети необходимо подготовить обучающую выборку с размеченными сущностями конфиденциальной информации. Подготовка данных заключается в их извлечении из документов, в отношении которых заведомо подтверждено содержание конфиденциальной информации (в основном офисного формата – текстов и таблиц, реже презентаций), очистке от лишних служебных символов и разбиении на токены[1] по пробельным символам и знакам пунктуации.

Отбор документов, содержащих конфиденциальную информацию, осуществляется экспертом по принципу их ценности для компании (влияние на рыночную позицию, недополучение прибыли и пр.) [11]. Далее из отобранных документов выделяется текстовый слой – удаляются графические и прочие объекты, не имеющие структуры текста, после чего выполняется разметка документа, представляющая собой разбиение его на функциональные элементы (предложения, фразы или слова) с последующей их разметкой тегами. Разметка может осуществляться одним из нескольких методов, выбор которого определяется минимальным объёмом информации, необходимым для определения наличия конфиденциальной информации, а также и сложностью/трудозатратами разработки той или иной модели. В зависимости от выбранной методики разметки может решаться одна из задач машинного обучения в области NLP (обработки естественного языка): binary text classification (бинарная классификация текстов) или NER (распознавание именованных сущностей). Задача NER в большинстве случаев решается с помощью нейросетей, например, может использоваться формирование эмбеддингов и распознавание именованных сущностей с помощью архитектуры трансформеров, таких как BERT.

В данной работе при подготовке данных для обучения нейросети использовалась методика разбиения текста по предложениям и была решена задача бинарной классификации текстов. Выбор данной методики обусловлен, прежде всего, её простотой, а также тем, что она не требует привлечения экспертов из конкретной области, что удобно поскольку разметка осуществляется вручную и является одним из самых трудозатратных этапов в обучении нейросети.

III. НЕЙРОСЕТЬ ДЛЯ КЛАССИФИКАЦИИ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ

В общем случае процесс обучения нейросети состоит из следующих основных этапов:

- предварительная очистка данных от лишних служебных символов;
- токенизация данных;
- разделение данных на несколько выборок (обучающую, валидационную) в определённых пропорциях;
- выбор архитектуры нейросети и обучение нейросети на базе архитектуры, показавшей наилучшие результаты;
- расчёт метрик, характеризующих качество нейросети.

В целях обучения нейросети была сформирована выборка реальных документов, содержащих информацию, составляющую коммерческую тайну в соответствии с требованиями действующего режима коммерческой тайны реального предприятия и извлеченных из систем внутреннего документооборота, кадровой системы, центрального архива и системы обеспечения заседаний коллегиальных органов.

В выборку включены 15960 оригинальных документов формата *.doc по семи пунктам Перечня сведений, составляющих коммерческую тайну по направлениям маркетинга, стратегического развития и планирования, решений коллегиальных органов управления, внутреннего аудита, математических моделей скоринга, а также результатов психофизиологических исследований (таблица 1).

Таблица 1. Состав документов по пунктам Перечня сведений, составляющих коммерческую тайну

Тип	**Количество оригинальных документов**
Конфиденциальная информация (Пункт 1)	133
Конфиденциальная информация (Пункт 2)	56
Конфиденциальная информация (Пункт 3)	29
Конфиденциальная информация (Пункт 4)	1
Конфиденциальная информация (Пункт 5)	36
Конфиденциальная информация (Пункт 6)	22
Конфиденциальная информация (Пункт 7)	171
Не конфиденциальная информация	15512

[1] Токен – часть текста, в большинстве случаев представленная словом или его частью

Необходимо отметить, что различные пункты могут быть сформулированы с использованием различных уровней абстракции, например, документы по пункту «Отчеты внутренних проверок системы менеджмента качества» будут однородны (даже, вероятно, шаблонированы), многочисленны и не потребуют интерпретации исполнителем, в то время как документы по пункту «Стратегические планы развития потребительского рынка» будут разнородны (представлены в различных формах, в т.ч. графических) и потребуют анализа и интерпретации о возможности отнесения конкретного документа или сообщения в данную категорию. В силу этих причин количество имеющихся в распоряжении документов по различным пунктам Перечня может сильно разниться (например, по пункту 7 имеется 171 документ, в то время, как по пункту 4 только один).

Все документы таблицы 1 разметки разделены на 128378 фраз, размеченных согласно выбранной методике тэгами конфиденциальности (рисунок 1).

	A	B	C	D
922	ABC-стратегия международного бизнеса	0	Стратегия - в.95.pdf	
923	Адаптация *** к международному кризису поставок, платеже	1	Стратегия - в.95.pdf	
924	1. Пересмотр стратегии в***	1	Стратегия - в.95.pdf	
925	Выбор опции в ***	0	Стратегия - в.95.pdf	
926	3. Перезапуск трансграничного кредитования	1	Стратегия - в.95.pdf	
927	Защита прав и возврат капитала	0	Стратегия - в.95.pdf	
928	Чистая прибыль 2026 составит *** млн (*** БП)	1	Стратегия - в.95.pdf	
929	Из-за санкций *** потеряет ***% клиентов ФЛ в 2027	1	Стратегия - в.95.pdf	
930	Стратегия ** в Беларуси до 2025 года будет представлена на	1	Стратегия - в.95.pdf	
931	Бизнес-модель *** нежизнеспособна	1	Стратегия - в.95.pdf	
932	3 стратегических опции: ******* (обслуживание оборотов РФ	1	Стратегия - в.95.pdf	
933	**** уменьшение капитала на $2,9 млн продажа за оставший	1	Стратегия - в.95.pdf	
934	Пилот 2021-2022 в Казахстане и Европе успешен: КП $1 млрд,	1	Стратегия - в.95.pdf	
935	Расширение проекта в ****: цель 2026 по КП ***, х** за 4 год	1	Стратегия - в.95.pdf	
936	Сохранение средств акционера ***** и поиск способа их воз	1	Стратегия - в.95.pdf	
937	Работа с регуляторами и юристами с целью защиты интересо	0	Стратегия - в.95.pdf	
938	Adapt	0	Стратегия - в.95.pdf	
939	Цель ЧП **** 2025 = $80 млн	1	Стратегия - в.95.pdf	
940	Цель ЧП **** 2025 = $75-100 млн	1	Стратегия - в.95.pdf	
941	Новые направления бизнеса и новые рынки для замещения в	0	Стратегия - в.95.pdf	

Рис. 1. Результаты разметки документов

IV. Оценка обученной нейросети

С точки зрения обеспечения эффективности и качества процесса защиты от утечек конфиденциальной информации показательны две ключевые метрики:

1. точность (*precision*) – характеризует количество ложных срабатываний и сохранение доверия к системе;
2. полнота (*recall*) – характеризует охват и способность системы выявить и классифицировать всю конфиденциальную информацию во всех потоках и массивах данных.

Поскольку идеальные (100%) значения метрик недостижимы и эффективность системы в целом в реальных условиях будет зависеть от соотношения показателей *Precision* и *Recall*, используется среднегармоническое значение полноты и точности (F_1-мера), которое вычисляется по формуле

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}.$$

Расчёт метрик осуществляется по тестовой части размеченной выборки, которая составляет 10% от её общего объема. Для расчета использовался метод матрицы ошибок, в которой истинные теги из размеченного набора размещаются по горизонтали, а теги, классифицированные нейросетью, по вертикали. С помощью этой матрицы вычислено количество истинно распознанных токенов конфиденциальной информации (*TP, true positive*), количество истинно нераспознанных токенов (*TN, true negative*), количество ложно распознанных токенов (*FP, false positive*) и количество ложно нераспознанных токенов (*FN, false negative*). После этого по формулам

$$recall = \frac{TP}{TP+FN}$$

и

$$precision = \frac{TP}{TP+FP}$$

определены полнота и точность, по которым, в свою очередь, вычисляется их среднее гармоническое – F_1-мера [12]. Средневзвешенное значение учитывает веса значений F_1-меры, полученное по каждому пункту Перечня, которые пропорциональны количеству документов *n*, представленных в обучающую выборку по каждому из пунктов. Средневзвешенное значение F_1-меры вычисляется по формуле:

$$\overline{F_1} = \frac{F1_1 n_1 + F1_2 n_2 + F1_3 n_3 + \ldots + F1_n n_n}{n_1 + n_2 + n_3 + \ldots + n_n}$$

Детальные результаты расчета метрик полученной нейросети классификации конфиденциальной информации представлены ниже (таблица 2).

Из таблицы 2 видно, что результаты классификации, демонстрируемые одной и той же нейросетью, разнятся в зависимости от пунктов Перечня сведений, составляющих коммерческую тайну. Такие результаты объясняются различными объемами и качеством обучающей выборки, зависящей, в свою очередь, от имеющихся реальных документов, покрываемых конкретными пунктами Перечня, количество которых, как отмечено ранее, сильно зависит от уровня абстракции конкретного пункта Перечня. Учитывая все вышесказанное, такие результаты следует считать нормальными.

Таблица 2. Метрики нейросети классификации конфиденциальной информации

Пункты Перечня информации, составляющей коммерческую тайну	**Precision** = tp / (tp + fp)	**Recall** = tp / (tp + fn)	**F_1** = 2·Precision·Recall /(Precision+Recall)
Не конфиденциальные документы	0,97	0,97	0,97
Пункт 1	0,86	1,00[2]	0,92
Пункт 2	1,00	0,80	0,89
Пункт 3	1,00	1,00	1,00
Пункт 4	1,00	1,00	1,00
Пункт 5	1,00	1,00	1,00
Пункт 6	1,00	1,00	1,00
Пункт 7	0,93	0,93	0,93
Среднее значение параметра	**0,97**	**0,97**	**0,97**

V. СОВЕРШЕНСТВОВАНИЕ ТЕХНОЛОГИЙ КЛАССИФИКАЦИИ ИНФОРМАЦИИ

Метод классификации конфиденциальной информации с использованием алгоритмов на основе правил плохо работает с неструктурированными данными, а метод классификации конфиденциальной информации с помощью нейросети демонстрирует недостаточное быстродействие при работе со структурированными данными [13]. Предлагаемый способ классификации конфиденциальной информации объединяет в себе два метода: одновременное использование алгоритмов на основе правил и нейросети. Использование комбинированного метода позволяет улучшить качество классификации конфиденциальной информации в смешанных (структурированных и неструктурированных) типах данных при одновременном снижении ресурсоемкости решения. Алгоритм классификации конфиденциальной информации, использующий одновременно алгоритмы на основе правил и нейросеть, представлен ниже (рисунок 2).

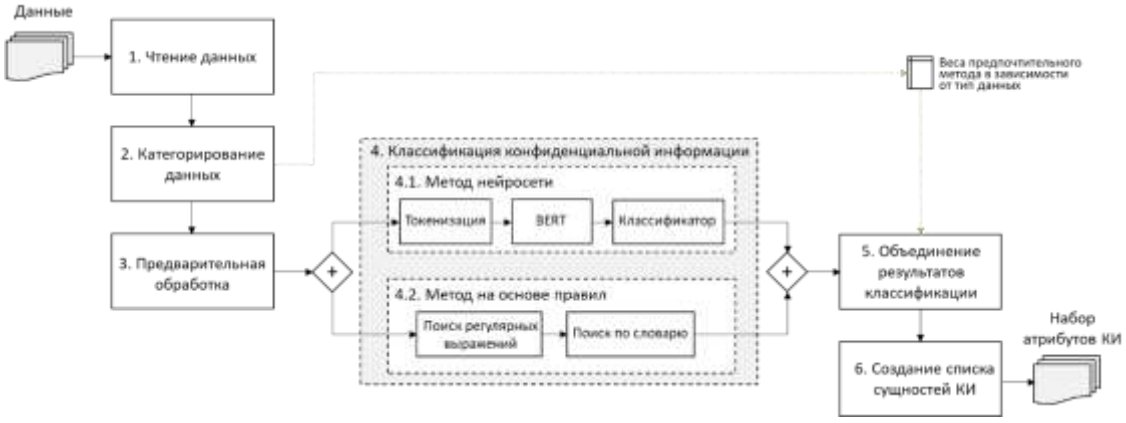

Рис. 2. Алгоритм классификации конфиденциальной информации

Описанный метод состоит из следующих этапов:

1. Чтение данных – На данном шаге извлекается содержимое из текстовых или табличных файлов. При выполнении данного шага извлекается

[2] Здесь и далее – получение параметров, близких или равным 1,00 обусловлено ограниченным набором выборки, связанным со сложностью формирования репрезентативного датасета документов, содержащих реальную коммерческую тайну по определенным пунктам Перечня

множество данных D={$d1$,…,dl}, где di – отдельный набор данных;

2. Категорирование данных – по расширению файла, его источнику, структуре и иным признакам определяется тип содержащейся в файле информации: структурированная или неструктурированная. В зависимости от определенного типа файлу присваивается весовой коэффициент, используемый в дальнейшем при назначении ему метода классификации. На данном шаге функция fT на основе данных D, расширения $Extension$, источника $Source$ и размера $Size$ формирует весовые коэффициенты W=fT(D,$Extension$,$Source$,$Size$), влияющие на приоритет методов распознавания в зависимости от типа информации;
3. Предварительная обработка данных – на данном шаге осуществляется очистка данных от служебных символов, разделение их на батчи таким образом, чтобы каждый батч содержал не более 512 токенов, но при этом не нарушался контекст исходных данных. После этого формируется множество батчей очищенных данных D'=fB(D)={D'1,…,$D'j$}, где $D'i$ – батч очищенных данных размером не более 512 токенов. Функция fB принимает на вход данные D, очищает их и разделяет на батчи;
4. Классификация конфиденциальной информации:

4.1. Классификация с использованием нейросети – входные данные разбиваются на токены, формируются эмбеддинги токенов, которые в свою очередь классифицируются на атрибуты.

4.2. Классификация алгоритмами на основе правил – выявление конфиденциальной информации с помощью регулярных выражений, проверки контрольных разрядов, поиска по словарю и нечёткого поиска.

5. Объединение результатов классификации нейросетью и алгоритмами на основе правил – результат работы нейросети, представленный списком токенов и соответствующих им тегов, преобразуется в массив атрибутов конфиденциальной информации и соответствующих им индексов начала и конца сущности по аналогии с алгоритмами на основе правил. Выбор атрибутов осуществляется на основе весовых коэффициентов, зависящих от типа данных, определённого на шаге 2, например, в структурированных данных многие алгоритмы на основе правил имеют большую значимость по сравнению с нейросетью и имеют больший весовой коэффициент. На данном шаге функция fU объединяет результаты двух методов классификации конфиденциальной информации AAI (4.1) и ARB (4.2), опираясь на весовые коэффициенты W, в общий вердикт и формирует финальный набор распознанных атрибутов конфиденциальности A=fU(AAI,ARB,W)={$a1$,…,ak};
6. Создание списка классифицированных сущностей конфиденциальной информации, состоящего из распознанных сущностей конфиденциальности, индексов начала и конца сущности, атрибутов конфиденциальной информации. Функция fEn на основе множества распознанных атрибутов конфиденциальной информации A и исходных данных D формирует множество распознанных сущностей En=fEn(A,D)={$e1$,…,ek}, содержащее атрибуты и индексы начала и конца сущностей конфиденциальной информации в исходных данных.

Комбинированный метод классификации конфиденциальной информации, использующий алгоритмы на основе правил в сочетании с нейросетью, при обработке структурированных данных отдаёт предпочтение распознаванию по правилам, так как контекст в таких данных практически отсутствует, а при обработке неструктурированных данных предпочтение отдаётся нейросети. Комбинация алгоритмов на основе правил с технологиями машинного обучения при классификации смешанных (структурированных и неструктурированных) улучшает метрики качества, демонстрируемые нативной нейросетью (таблица 3).

Таблица 3. Метрики алгоритмов классификации конфиденциальной информации

Пункты Перечня информации, составляющей коммерческую тайну	**Нейросеть**	**Нейросеть + Правила**
Пункт 1	0,92	0,96
Пункт 2	0,89	0,91
Пункт 3	1,00	1,00
Пункт 4	1,00	1,00
Пункт 5	1,00	1,00
Пункт 6	1,00	1,00
Пункт 7	0,93	0,98
Не конфиденциальные документы	0,97	0,98
Среднее значение параметра	**0,9690**	**0,9797**

Необходимо заметить, что приращение качества по различным пунктам при этом неоднородно по двум причинам:

• документы, представленные для обучения нейросети по различным пунктам Перечня, имеют различную степень структурированности. Алгоритмы на основе правил покажут более высокие результаты на документах, содержащих преимущественно структурированные данные;

• различные веса конкретной F_1-меры, значения с наибольшим весом внесут наибольший «вклад» в изменение средневзвешенного значения даже при минимальном изменении абсолютного значения F_1-меры.

VI. Заключение

На сегодняшний день ни одна технология не обеспечивает сама по себе одновременного соблюдения условий высокого быстродействия, универсальности и

точности классификации информации. Для построения эффективной системы защиты от утечек конфиденциальной информации необходимо выстраивать гибридную систему защиты, используя каждую технологию для решения тех задач, где она демонстрирует максимальную эффективность, сочетая использование лингвистических и статистических методов с технологиями машинного обучения. В данной работе представлен способ формирования датасета для обучения нейросети, алгоритм использования технологий искусственного интеллекта (нейросетей) совместно с методами классификации на основе правил, а также приведены оценки их эффективности. Применение данного алгоритма в реальной системе позволит повысить эффективность системы классификации конфиденциальной информации, представленной одновременно в структурированном и неструктурированном виде.

Данная гибридная система позволит распознавать даже ту конфиденциальную информацию, на которой не обучалась нейросеть, однако именно технологии машинного обучения являются её ядром, предоставляя гибкость и масштабируемость при управляемой точности распознавания конфиденциальной информации в неструктурированных данных, недоступную прочим технологиям.

Библиография

[1] Гарбузов Г. В. Технологии защиты нематериальных активов от атак на конфиденциальность // International Journal of Open Information Technologies. 2024. Т. 12, № 9. С. 142-149. EDN: CXXTYY

[2] Гарбузов Г. В. Проблемы выявления утечек конфиденциальной информации в неструктурированных данных // International Journal of Open Information Technologies. 2025. Т. 13, № 4. С. 26-32. EDN: HJLUXD

[3] Gartner, Consult the Board: Unstructured Data Management, URL: https://www.gartner.com/en/documents/4373899 (дата обращения 13.11.2025)

[4] Зарубин А. В., Смирнов М. Б., Харитонов С. В., Денисов Д. В. Основные драйверы и тенденции развития DLP-систем в Российской Федерации // Прикладная информатика. 2020. Т. 15. № 3(87). С. 75-90. doi: https://doi.org/10.37791/2687-0649-2020-15-3-75-90

[5] Tarmizi S. A. Named entity recognition for quranic text using rule based approaches // Asia-Pacific Journal of Information Technology & Multimedia. 2022. Vol. 11. No. 2. P. 112-122. doi: https://doi.org/10.17576/apjitm-2022-1102-09

[6] Раздьяконов Е. С. Поиск персональных данных в неструктурированных текстах с использованием нейронных сетей // Инженерный вестник Дона. 2023. № 7(103). С. 589-605. EDN: MXVMJW

[7] Donglan Liu, Xin Liu, Lei Ma, Yingxian Chang, Rui Wang, Hao Zhang, Hao Yu, Wenting Wang. Research on Leakage Prevention Technology of Sensitive Data based on Artificial Intelligence // 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC). Beijing, China: IEEE Computer Society, 2020. P. 142-145. doi: https://doi.org/10.1109/ICEIEC49280.2020.9152286

[8] Zhu T., Ye D., Wang W., Zhou W., Yu P.S. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence // IEEE Transactions on Knowledge and Data Engineering. 2022. Vol. 34, No. 6. P. 2824-2843. doi: https://doi.org/10.1109/TKDE.2020.3014246

[9] Guha A., Samanta D., Banerjee A., Agarwal D. Deep Learning Model for Information Loss Prevention From Multi-Page Digital Documents // IEEE Access. 2021. Vol. 9. P. 80451-80465. doi: https://doi.org/10.1109/ACCESS.2021.3084841

[10] Артюшкина Е. С., Скакун О. О., Гузь А. Р. Использование искусственного интеллекта в DLP-системах // Прикладные экономические исследования. 2023. № 2. С. 123-129. doi: https://doi.org/10.47576/2949-1908_2023_2_123

[11] Martinelli F., Marulli F., Mercaldo F., Marrone S., Santone A. Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence // 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE Computer Society, 2020. P. 1-8. doi: https://doi.org/10.1109/IJCNN48605.2020.9206801

[12] Williams C. K. I. The Effect of Class Imbalance on Precision-Recall Curves // Neural Computation. 2021. Vol. 33, No. 4. P. 853-857. doi: https://doi.org/10.1162/neco_a_01362

[13] Kim J., Lee C., Chang H. The Development of a Security Evaluation Model Focused on Information Leakage Protection for Sustainable Growth // Sustainability. 2020. Vol. 12, issue 24. Article number: 10639. doi: https://doi.org/10.3390/su122410639

Methods for Determining Confidential Information in Unstructured Data

G.V. Garbuzov, S.V. Dvoryankin

***Abstract*— This article addresses methods for identifying (classifying, recognizing) confidential information in unstructured data presented in the form of files transmitted over communication channels or stored in file resources. The emphasis on unstructured data is due to the fact that, on the one hand, it is unstructured data that is of the greatest interest to an attacker in terms of the content of confidential information of a commercial enterprise (trade secret, know-how), and, at the same time, it is unstructured data that is difficult to analyze by the signature algorithms and rules based on regular expressions used today, which are used in modern means of protecting against information leaks.**

***Keywords*— unstructured data, confidential information, trade secret, information leaks, data leakage protection systems.**

REFERENCES

[1] Garbuzov G. Technologies for Protecting Intangible Assets from Confidentiality Attacks. *International Journal of Open Information Technologies*, 2024, vol. 12, no. 9, pp. 142-149. (In Russ., abstract in Eng.) EDN: CXXTYY

[2] Garbuzov G. Issues in Detecting Confidential Information Leaks in Unstructured Data. *International Journal of Open Information Technologies*, 2025, vol. 13, no. 4, pp. 26-32. (In Russ., abstract in Eng.) EDN: HJLUXD

[3] Gartner, Consult the Board: Unstructured Data Management. [Online]. Available: https://www.gartner.com/en/documents/4373899

[4] Zarubin A., Smimov B., Kharitonov S., Denisov D., Main drivers and trends of DLP systems development in the Russian Federation *Prikladnaya informatika* = Journal of Applied Informatics, 2020, vol.15, no. 3, pp. 75-90. (In Russ., abstract in Eng.) doi: https://doi.org/10.37791/2687-0649-2020-15-3-75-90

[5] Tarmizi S. Named entity recognition for quranic text using rule based approaches. *Asia-Pacific Journal of Information Technology & Multimedia*. 2022, vol. 11, no. 2, pp. 112-122. doi: https://doi.org/10.17576/apjitm-2022-1102-09

[6] Razdyakonov E.S. Personal data recognition in unstructured texts using neural networks. *Engineering journal of Don*. 2023, no. 7, pp. 589-605. (In Russ., abstract in Eng.) EDN: MXVMJW

[7] Donglan Liu, Xin Liu, Lei Ma, Yingxian Chang, Rui Wang, Hao Zhang, Hao Yu, Wenting Wang. Research on Leakage Prevention Technology of Sensitive Data based on Artificial Intelligence. In: *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. Beijing, China: IEEE Computer Society; 2020. pp. 142-145. doi: https://doi.org/10.1109/ICEIEC49280.2020.9152286

[8] Zhu T., Ye D., Wang W., Zhou W., Yu P.S. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 2022, vol. 34, no. 6, pp. 2824-2843. doi: https://doi.org/10.1109/TKDE.2020.3014246

[9] Guha A., Samanta D., Banerjee A., Agarwal D. Deep Learning Model for Information Loss Prevention From Multi-Page Digital Documents. *IEEE Access*, 2021, vol. 9, pp. 80451-80465. doi: https://doi.org/10.1109/ACCESS.2021.3084841

[10] Artyushkina E.S., Skakun O.O., Guz A.R. Using artificial intelligence in DLP systems. *Applied economic research*, 2023, no. 2, pp. 123-129. doi: https://doi.org/10.47576/2949-1908_2023_2_123

[11] Martinelli F., Marulli F., Mercaldo F., Marrone S., Santone A. Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK: IEEE Computer Society, 2020. pp. 1-8. doi: https://doi.org/10.1109/IJCNN48605.2020.9206801

[12] Williams C. K. I. The Effect of Class Imbalance on Precision-Recall Curves. *Neural Computation*, 2021, vol. 33, no. 4, pp. 853-857. doi: https://doi.org/10.1162/neco_a_01362

[13] Kim J., Lee C., Chang H. The Development of a Security Evaluation Model Focused on Information Leakage Protection for Sustainable Growth. *Sustainability*, 2020, vol. 12, issue 24. Article number: 10639. doi: https://doi.org/10.3390/su122410639