УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2021 Том 21 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2021 Vol. 21 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2021-21-3-394-400
УДК 004.89

Алгоритм выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании

Алиса Андреевна Воробьева[1]✉, Владислав Владимирович Герасимов[2], Юлия Валерьевна Ли[3]

[1,2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] Alice_w@mail.ru✉, http://orcid.org/0000-0001-6691-6167
[2] gevol.69@gmail.com, http://orcid.org/0000-0001-8099-2414
[3] J_u_l_i_a_0908@mail.ru, http://orcid.org/0000-0003-3280-8197

Аннотация

Предмет исследования. Рассмотрена задача выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании. Создан оригинальный набор данных, представляющий собой записи разговоров между операторами и клиентами, представленный в виде диалогов в текстовом формате. Изучена применимость методов машинного обучения для автоматизации выявления утечек, возникающих в разговоре между оператором и клиентом. Приведены результаты использования данных методов для построения и обучения классификатора: вероятностные (наивный байесовский классификатор), метрические (метод *k*-ближайших соседей), логические (случайный лес), линейные (метод опорных векторов), методы на основе искусственных нейронных сетей. Рассмотрены различные подходы к построению модели текстов на естественном языке, такие как токенизация (bag of words, *n*-граммы слов: биграммы и триграммы) и векторизация (one hot encoding). **Метод.** Предлагаемый алгоритм выявления утечек базируется на применении метода опорных векторов (SVM) и токенизации по биграммам слов. **Основные результаты.** Полученные результаты демонстрируют, что использование SVM и токенизация по биграммам обеспечивают наиболее высокое качество выявления утечек. **Практическая значимость.** Результаты исследования могут найти применение при разработке программных систем и комплексов защиты информации, а также для дальнейшего развития методов обработки естественного языка применительно к задачам информационной безопасности.

Ключевые слова

обработка естественного языка, машинное обучение, нейронные сети, комплаенс-риски, инсайдерская информация

Благодарности

Работа выполнена в Университете ИТМО в рамках темы НИР № 50449 «Разработка алгоритмов защиты киберпространства для решения прикладных задач обеспечения кибербезопасности организаций банковской сферы».

Ссылка для цитирования: Воробьева А.А., Герасимов В.В., Ли Ю.В. Алгоритм выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 3. С. 394–400. doi: 10.17586/2226-1494-2021-21-3-394-400

An algorithm for detecting leaks of insider information of financial markets in investment consulting

Alisa A. Vorobeva[1]✉, Vladislav V. Gerasimov[2], Yulia V. Li[3]

[1,2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] Alice_w@mail.ru✉, http://orcid.org/0000-0001-6691-6167
[2] gevol.69@gmail.com, http://orcid.org/0000-0001-8099-2414
[3] J_u_l_i_a_0908@mail.ru, http://orcid.org/0000-0003-3280-8197

Abstract

The paper focuses on revealing insider information leaks of financial markets during investment consulting. An original dataset was created, containing the records of the conversations between consultants and clients, presented in the form

of dialogs in text format. The applicability of machine learning methods for automating the detection of leaks arising in a conversation between a consultant and a client has been studied. The authors examined the applicability of the following supervised machine learning methods for constructing and training a classifier: probabilistic (Naïve Bayes classifier), metric (*k*-nearest neighbors algorithm), logical (random forest), linear (support vector machine), and methods based on artificial neural networks. The paper considers various approaches to the construction of a natural language text model, such as tokenization (bag of words, word *n*-grams: bigrams and trigrams) and vectorization (one-hot encoding). The proposed algorithm for detecting financial markets insider information leaks is based on the use of support vector machine (SVM) and tokenization by bigrams. The obtained results demonstrate that SVM and bigram tokenization provide the highest leakage detection accuracy. The research results can be used in cybersecurity tools development, as well as for the further elaboration of natural language processing methods dealing with information security problems.

Keywords
natural language processing, machine learning, neural networks, compliance risks, insider information

Acknowledgments
The paper was prepared at ITMO University within the framework of the scientific project No. 50449 “Development of cyberspace protection algorithms for solving applied problems of ensuring cybersecurity of banking organizations”.

For citation: Vorobeva A.A., Gerasimov V.V., Li Yu.V. An algorithm for detecting leaks of insider information of financial markets in investment consulting. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 3, pp. 394–400 (in Russian). doi: 10.17586/2226-1494-2021-21-3-394-400

Введение

Комплаенс (compliance) — комплекс мер, принимаемых в соответствии с законодательством, требованиями, указаниями и рекомендациями регулирующих органов Российской Федерации и применимыми нормами международного права[1]. Комплаенс-контроль является частью политики управления конфликтом интересов [1]. Данная область регулируется в первую очередь Федеральным законом №224-ФЗ[2].

Перечень инсайдерской информации банка как профессионального участника рынка ценных бумаг (финансового рынка) достаточно велик и в общем виде включает сведения о совершении сделок с ценными бумагами в случае, когда исполнение таких поручений может оказать существенное влияние на стоимость соответствующих финансовых активов[3].

В качестве конфликтов интереса в области инвестиционного консультирования выделяют [2], например, предоставление клиентам информации по операциям с финансовыми инструментами банка в ущерб другим участникам финансового рынка. Также ярким примером конфликта интересов является ситуация, когда работник проводит консультацию по потенциальным сделкам с эмитентом, в отношении ценных бумаг которого работник имеет личный интерес.

В настоящее время существует проблема передачи инсайдерской информации рынка ценных бумаг сотрудниками колл-центра[4]. Участились случаи мошеннических соглашений между инвесторами и сотрудниками банков, когда последние конспиративно предоставляют инсайдерскую информацию путем заранее обдуманных кодовых слов, намеков и других словесных инструментов передачи имплицитной информации.

Метод экспертного анализа, используемый в настоящее время в решении проблемы выявления событий комплаенс-рисков, связанных с утечками инсайдерской информации финансовых рынков при инвестиционном консультировании, трудозатратен и малоэффективен. Применение методов машинного обучения (МО) — перспективный подход решения данной задачи, так как позволяет повысить степень автоматизации выявления фактов передачи инсайдерской информации в речи человека, а также увеличивает вероятность выявления таких случаев.

Анализ предметной области показал [3], что разговоры между сотрудником колл-центра и клиентом (инвестором) проходят по заранее определенным сценариям, в которых заданы правила формирования ответов на определенные запросы клиентов. Следовательно, задача выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании операторами, сводится к выявлению диалогов, отклоняющихся от установленных правил, которые являются подозрительными.

В настоящей работе сделано предположение, что методы обработки естественного языка (NLP), а также методы МО позволят автоматизировать выявление утечек инсайдерской информации рынка ценных бумаг. Основными вопросами при разработке подобных методов являются: поиск признакового пространства и метода МО, позволяющего получить максимальную точность на данных предметной области.

Анализ предыдущих работ по оценке текстовых сообщений позволил сделать выводы, что наиболее

[1] Комплаенс ПАО «Промсвязьбанк», 2021 [Электронный ресурс]. Режим доступа: https://www.psbank.ru/Bank/Compliance (дата обращения: 01.03.2021).

[2] Федеральный закон «О противодействии неправомерному использованию инсайдерской информации и манипулированию рынком и о внесении изменений в отдельные законодательные акты Российской Федерации» от 27.07.2010 №224-ФЗ.

[3] Перечень инсайдерской информации ПАО «Сбербанк». Приложение к приказу № 584-О от 27.12.2019 [Электронный ресурс]. URL: https://www.sberbank.com/common/img/uploaded/files/pdf/insider_perechen_140218.pdf (дата обращения: 04.03.2021).

[4] Политика группы ПАО «Сбербанк» по управлению конфликтом интересов. [Электронный ресурс]. URL: https://www.sberbank.com/common/img/uploaded/files/pdf/normative_docs/conflict_of_interest_management_policy_ru.pdf (дата обращения: 02.03.2021).

перспективными для исследования являются методы МО: вероятностные (наивный байесовский классификатор, NB) [4], метрические (метод *k*-ближайших соседей, *k*-NN) [5], логические (деревья принятия решений, DT, и метод случайный лес, RF) [6], линейные (метод опорных векторов, SVM [4, 5] и логистическая регрессия, LR) [7], методы на основе искусственных нейронных сетей (многослойный перцептрон, MLP) [8], которые рассматриваются в разделе «Исследование методов машинного обучения для выявления утечек инсайдерской информации финансовых рынков».

Формальная постановка задачи выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании

В банке приняты правила (R) совершения разговоров между операторами колл-центра банка (O) и клиентами-инвесторами (C), представленные в виде скриптов (S). Имеются записи телефонных разговоров (T) между O и C, преобразованные в текстовый формат, собранные за определенный промежуток времени. Для каждой новый записи разговора (k), преобразованной в текстовый формат (t_k), оператора (o_k) и клиента (c_k) необходимо определить вероятность (p), с которой k отклоняется от R, а также сравнить p с пороговым значением (τ). В случае превышения $\tau = 0{,}6$, k считается подозрительной.

Каждая запись t_k представлена как вектор признаков $t_k = (f_{k1}, \dots f_{km})$, где m — размер признакового пространства. При этом $T = (t_1, \dots t_z)$, где z — количество записей телефонных разговоров, преобразованных в текстовый формат. Для всех T известно, содержит ли запись утечку инсайдерской информации. T включает как записи T_a (подозрительные диалоги), содержащие утечку инсайдерской информации, так и легитимные записи (T_n), где T_a, $T_n \in T$.

Необходимо создать модель *Mcl* (классификатор), способную определить, с какими вероятностями запись t_k относится к классам «легитимный диалог» (*An*) и «подозрительный диалог» (*N*). Обучение данной модели производится методами МО с учителем по обучающей выборке $T_{tr} \in T$, а верификация по $T_{test} \in T$, где T_{tr} — обучающая выборка, T_{test} — тестовая выборка. Далее обученный алгоритм должен быть способен классифицировать новую запись t_k к *An* и *N*. Искомым классом является тот, для которого вероятность максимальна.

Формирование признакового пространства

Для решения поставленной задачи, связанной с обработкой текстов на естественном языке методами МО, необходимо преобразовать тексты в цифровой формат — произвести их токенизацию и векторизацию [9].

При токенизации текст разбивается на части, каждая из которых будет отдельно представляться в цифровом виде. Токенизация может производиться по отдельным символам (буквам, цифрам, знакам препинания), по словам, по предложениям. После разбивки текста на отдельные токены, каждый токен необходимо преобразовать в число, т. е. векторизировать текст.

В работе рассмотрены два подхода к токенизации текстов: bag of words (BOW) и *n*-граммы уровня слов. Модель представления текста BOW — это множество всех слов, встречающихся в тексте, где частота появления каждого слова используется в качестве признака для обучения классификатора [10]. *n*-граммы представляют собой последовательности слов длиной от 1 до *n*, где *n* — размерность *n*-граммы [11]. Например, биграммы являются последовательностью двух слов.

Проведенный анализ литературы позволил выявить наиболее перспективные методы векторизации для решения указанной задачи: числовое кодирование, векторизация в формате one hot encoding [12] и плотное векторное представление (embedding) [13]. В настоящей работе использована векторизация в формате one hot encoding, так как таким способом можно закодировать любое представление токенов. В этом случае каждому токену ставится в соответствие не одно число, а вектор.

Подготовка исходных данных для эксперимента

При изучении существующих наборов данных выявлено, что в открытом доступе не существует готовых наборов, которые могли бы быть использованы для проведения экспериментальных исследований.

Авторами создан набор данных, включающих тексты (диалоги), сгенерированные из скриптов для телефонных разговоров на русском языке[1] в автоматическом режиме.

Для автоматической генерации записей, являющихся легитимными (T_n), а также диалогов, не соответствующих скриптам — подозрительными (T_a), написан скрипт на языке программирования Python. Исходные данные для скрипта оформлены в формате JSON (JavaScript Object Notation).

Количество записей в T_a — 50, в T_n — 150. Набор данных является размеченным.

Минимальная длина диалога в символах — 2128, максимальная — 4828. При этом минимальное количество слов в диалоге — 302, а максимальное — 648. Все диалоги содержат 21 реплику. Распределение частот диалогов определенной длины в наборе данных представлено на рис. 1.

Исследование методов машинного обучения для выявления утечек инсайдерской информации финансовых рынков

Как было сказано в разделе «Введение», один из основных вопросов при разработке алгоритма выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании — выбор метода МО, который может быть использован для обучения модели, а также метода формирования признакового пространства: токенизации и векторизации. Обоснование выбора указанных методов для

[1] Скрипт исходящего звонка телефонных продаж [Электронный ресурс]. URL: https://best-business-info.ru/script/Q1.html (дата обращения: 15.02.2021).

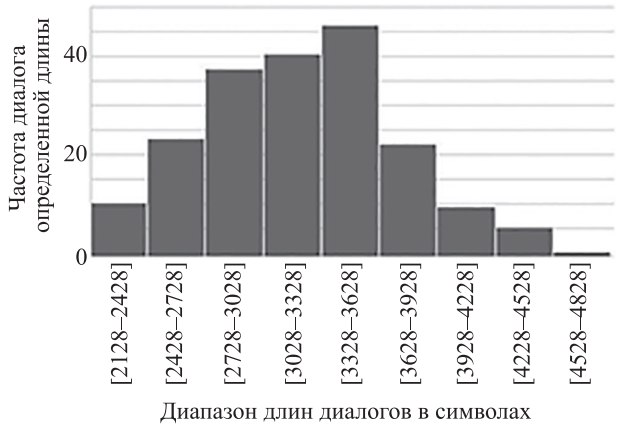

Рис. 1. Распределение частот диалогов определенной длины в наборе данных

Fig. 1. Frequency distribution of a certain length dialogues in the dataset

включения в разрабатываемой алгоритм основано на результатах экспериментальных исследований.

Рассмотрены следующие методы МО, которые могут применяться для решения поставленной задачи и обучения *Mcl*:

— нейронные сети (NN);
— наивный байесовский классификатор (NB);
— метод *k*-ближайших соседей (*k*-NN);
— метод случайного леса (RF);
— метод опорных векторов (SVM);
— многослойный перцептрон (MLP).

Произведена оценка достоверности (*A*), точности (*P*), полноты (*R*) и F-меры (*F*) выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании указанных методов МО, достигаемой с использованием методов токенизации (отобранных в результате предварительного анализа): BOW, биграммы и триграммы.

Под *F* выявления утечек инсайдерской информации понимается формальная метрика оценки качества классификатора, объединяющая *P* и *R* бинарной классификации [14]. Конечная формула для расчета *F* имеет следующий вид [15]:

$$F = \frac{2 \times P \times R}{P + R}.$$

Для проведения экспериментальных исследований по выбору наилучших методов токенизации текста и наилучшего классификатора для бинарной классификации текстов использованы язык программирования Python, библиотеки: Jupyter Notebook, Pandas, Keras, Sklearn.

Итоговая оценка качества бинарной классификации основана на значении *F*. Данный выбор обусловлен тем, что набор данных, в экспериментальных исследованиях, является сильно несбалансированным, класс *An* имеет высокую априорную вероятность. Данное условие ведет к возникновению так называемого «парадокса достоверности», когда значение *A* (доля правильно классифицированных объектов) имеет в результате довольно высокие значения для всех методов токенизации, при этом полнота методов токенизации *R* варьируется в пределах от 0,5 до 0,6 за исключением метода SVM. Вследствие этого было принято решение в качестве итоговой оценки качества бинарной классификации использовать результаты расчетов *F*.

В таблице представлены результаты эксперимента по оценке влияния метода токенизации и метода МО на достоверность, точность и полноту выявления утечек инсайдерской информации.

Наиболее высокие результаты получены с использованием SVM для всех методов токенизации и формирования признакового пространства. Для более наглядного отображения полученных результатов проведен ряд дополнительных экспериментов с построением матрицы ошибок для классификатора SVM и всех методов токенизации в графическом виде (рис. 2).

Для токенизатора, основанного на методе BOW:

— количество истинно отрицательных (TN) и ложноотрицательных (FN) решений классификатора на T_{test} равно 11 и 34 соответственно;
— количество истинно положительных (TP) и ложноположительных (FP) решений классификатора на T_{test} равно 134 и 1 соответственно (рис. 2, *a*).

Для токенизаторов, основанных на биграммах и триграммах, количество истинно положительных (TP) и ложноположительных (FP) решений классификатора на T_{test} равно 135 и 0 соответственно.

Для токенизатора, основанного на биграммах, количество истинно отрицательных (TN) и ложноотрицательных (FN) решений классификатора на T_{test} равно 32 и 13 соответственно (рис. 2, *b*).

Таблица. Результаты эксперимента по оценке влияния метода токенизации и метода машинного обучения на достоверность (*A*), точность (*P*), полноту выявления утечек инсайдерской информации (*R*), F-меру (*F*)

*Table. Re*sults of tokenization method and ML method effect on accuracy (*A*), precision (*P*), recall (*R*) and F-score (*F*) of insider information leaks detection

Метод машинного обучения	Методы токенизации											
	BOW				Биграммы				Триграммы			
	A	*P*	*R*	*F*	*A*	*P*	*R*	*F*	*A*	*P*	*R*	*F*
MLP	0,77	0,88	0,54	0,59	0,77	0,88	0,53	0,59	0,77	0,88	0,53	0,64
NB	0,78	0,89	0,57	0,64	0,78	0,89	0,57	0,59	0,79	0,89	0,59	0,62
k-NN	0,74	0,54	0,50	0,47	0,76	0,88	0,51	0,53	0,75	0,38	0,5	0,49
RF	0,77	0,88	0,53	0,51	0,76	0,88	0,52	0,51	0,76	0,88	0,52	0,51
SVM	0,81	0,86	0,62	0,68	0,93	0,96	0,86	0,82	0,82	0,90	0,64	0,69

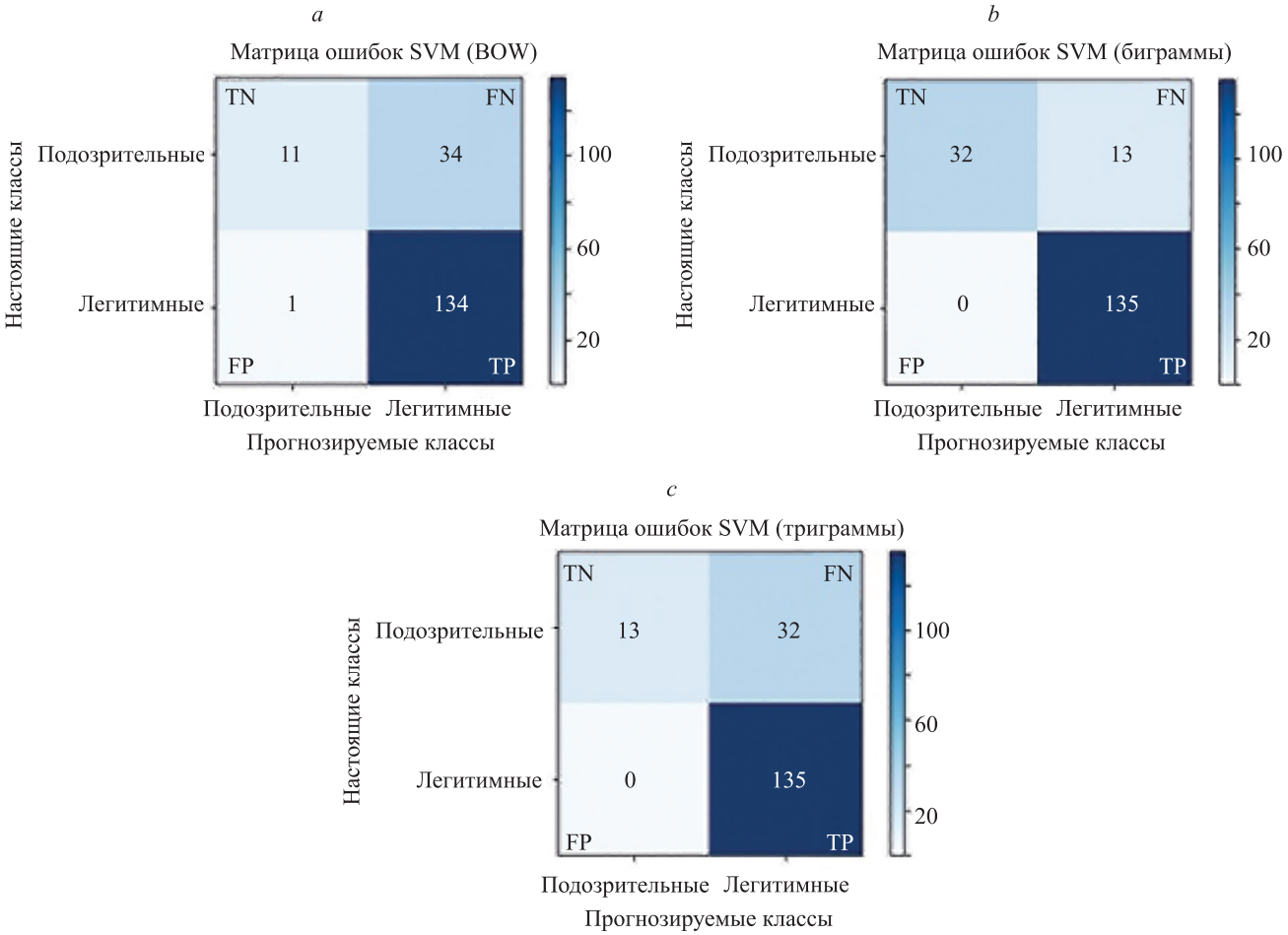

Рис. 2. Матрицы ошибок SVM с различными методами токенизации: с частотной токенизацией BOW (*a*), с токенизацией в виде: биграмм (*b*) и триграмм (*c*)

Fig. 2. SVM confusion matrix with different tokenization methods: (*a*) BOW frequency tokenization, (*b*) bigram tokenization, (*c*) trigram tokenization

Для токенизатора, основанного на триграммах, количество истинно отрицательных (TN) и ложноотрицательных (FN) решений классификатора на T_{test} равно 13 и 32 соответственно (рис. 2, *c*).

Использован классификатор SVM со следующими параметрами:

— ядро классификатора (kernel) линейное (linear);
— probability = True (включение оценок вероятности);
— class_weight = 'balanced' (сбалансированный режим распределения весов для классов использует значения классов для автоматической настройки весов, обратно пропорциональных частотам классов во входных данных)[1].

Экспериментальная оценка рассмотренных подходов обработки естественного языка и МО показала, что наилучший метод токенизации диалогов — разбиение текста на биграммы, а наилучший классификатор диалогов на подозрительные и легитимные — метод опорных векторов. Качество классификации *F* составляет 0,82.

Предложенный алгоритм выявления утечек инсайдерской информации

Разработанный алгоритм основан на анализе голосовых записей разговоров между оператором колл-центра и клиентом, представленных в текстовом формате с использованием методов обработки естественного языка и МО.

Алгоритм анализирует голосовые записи разговоров операторов колл-центра и клиентов, представленные в текстовом формате. Переговоры ведутся по заранее заданным правилам, оператор должен строго следовать предусмотренному сценарию, называемому скрипт[2]. Скрипт — это заранее продуманный текст разговора сотрудника с клиентом, в котором учтены все возможные ответы и возражения [16].

Разработанный алгоритм включает три этапа.

Этап 1. Производится обучение модели *Mcl* (с использованием SVM) выявления подозрительных диалогов по T_{tr} и верификация модели по T_{test}.

[1] «sklearn.svm.SVC» — документация sklearn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (дата обращения: 25.03.2021).

[2] «Сбербанк» — финансовым организациям. Подача поручений по телефону [Электронный ресурс]. URL: https://www.sberbank.ru/ru/credit_org/investments/ongr/broker_service/tradesystems/phone (дата обращения: 05.03.2021).

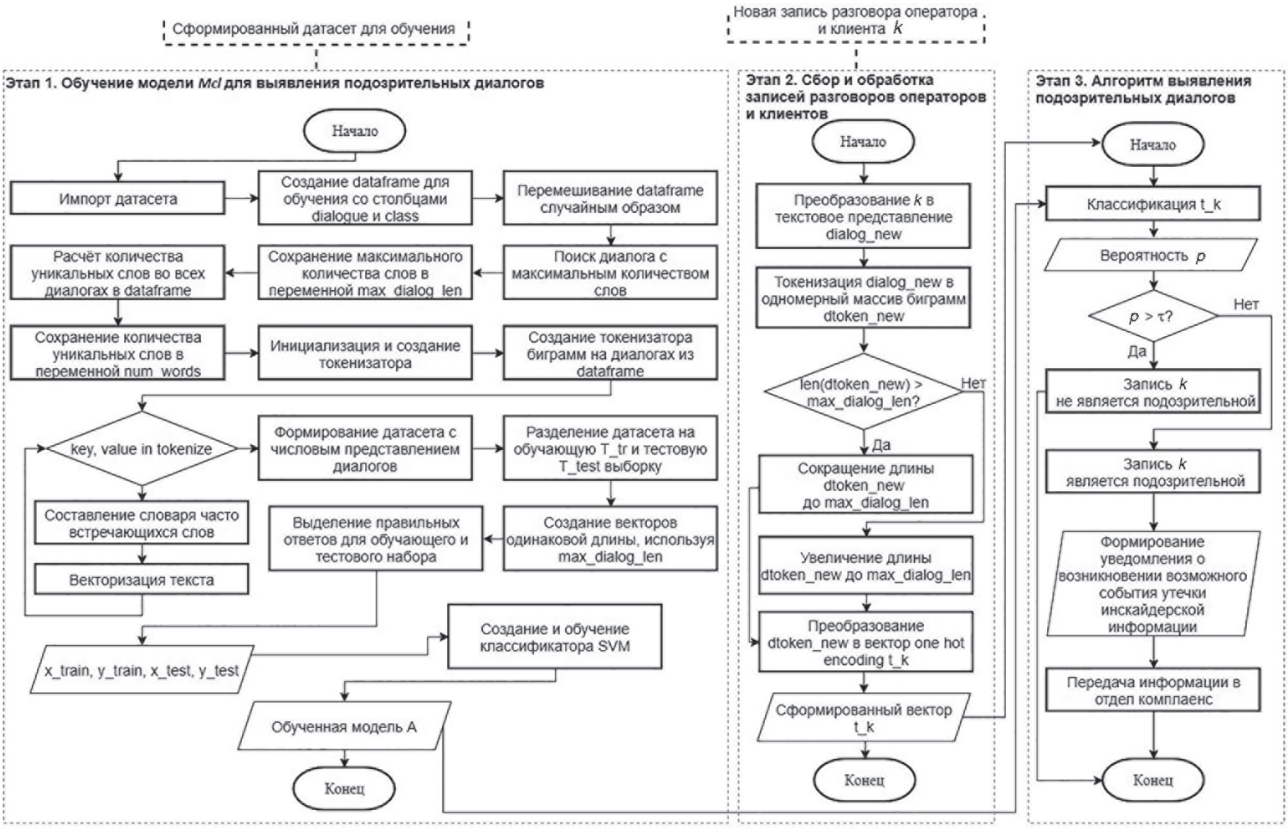

Рис. 3. Блок-схема алгоритма выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании

Fig. 3. Flowchart of the algorithm for detecting leaks of insider information of financial markets in investment consulting

Этап 2. При получении новой голосовой записи k выполняется ее преобразование в текстовый формат, токенизация (по биграммам слов) и векторизация в t_k (one hot encoding).

Этап 3. С использованием подготовленной модели *Mcl* определяется вероятность p, с которой t_k является подозрительным. В случае, если p превышает заданное пороговое значение τ, запись отмечается как подозрительная, и сотрудник отдела комплаенс получает уведомление о возникновении возможного события утечки инсайдерской информации.

Блок-схема алгоритма представлена на рис. 3.

Заключение

Рассмотрена важная для банковской сферы проблема — выявление утечек инсайдерской информации операторами колл-центров, консультирующих клиентов по вопросам инвестиций, влекущих за собой возникновения комплаенс-рисков. Данная проблема является новой и сложной для разработки автоматизированного решения.

В работе представлен разработанный алгоритм выявления утечек инсайдерской информации финансовых рынков при инвестиционном консультировании, анализирующий голосовые записи (представленные в текстовом формате) разговоров операторов и клиентов и основанный на методах обработки естественного языка и машинного обучения. Подготовлено теоретическое описание алгоритма и представлена его блок-схема.

Экспериментальная оценка рассмотренных подходов обработки естественного языка и машинного обучения показала, что наилучшим методом токенизации диалогов является разбиение текста на биграммы, а наилучшим классификатором диалогов на подозрительные и легитимные является метод опорных векторов. Качество классификации F составляет 0,82.

Литература

1. Nini G., Smith D.C., Sufi A. Creditor control rights and firm investment policy // Journal of Financial Economics. 2009. V. 92. N 3. P. 400–420. doi: 10.1016/j.jfineco.2008.04.008
2. Jaiswal S. Connections and conflicts of interest: investment consultants' recommendations, SSRN. 2018 [Электронный ресурс]. URL: https://ssrn.com/abstract=3106528 (дата обращения: 05.03.2021). doi: 10.2139/ssrn.3106528

References

1. Nini G., Smith D.C., Sufi A. Creditor control rights and firm investment policy. *Journal of Financial Economics*, 2009, vol. 92, no. 3, pp. 400–420. doi: 10.1016/j.jfineco.2008.04.008
2. Jaiswal S. *Connections and conflicts of interest: investment consultants' recommendations, SSRN*. 2018. Available at: https://ssrn.com/abstract=3106528 (accessed: 05.03.2021). doi: 10.2139/ssrn.3106528

3. Jenkinson T., Jones H., Martinez J.V. Picking winners? Investment consultants' recommendations of fund managers // Journal of Finance. 2016. V. 71. N 5. P. 2333–2370. doi: 10.1111/jofi.12289.
4. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: a survey // Ain Shams Engineering Journal. 2014. V. 5. N 4. P. 1093–1113. doi: 10.1016/j.asej.2014.04.011
5. Ghiassi M., Olschimke M., Moon B., Arnaudo P. Automated text classification using a dynamic artificial neural network model // Expert Systems with Applications. 2012. V. 39. N 12. P. 10967–10976. doi: 10.1016/j.eswa.2012.03.027
6. Fuller C.M., Biros D.P., Delen D. An investigation of data and text mining methods for real world deception detection // Expert Systems with Applications. 2011. V. 38. N 7. P. 8392–8398. doi: 10.1016/j.eswa.2011.01.032
7. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. № 1. С. 85–99. doi: 10.15827/0236-235X.030.1.085-099
8. Алексеева В.А. Использование методов интеллектуального анализа в задачах бинарной классификации // Известия Самарского научного центра РАН. 2014. Т. 16. № 6-2. Р. 354–356.
9. Бабаев А.М. Основные принципы обработки естественного языка // Дневник науки. 2019. № 12. С. 14.
10. Zhang Y., Jin R., Zhou Z.-H. Understanding bag-of-words model: a statistical framework // International Journal of Machine Learning and Cybernetics. 2010. V. 1. N 1-4. P. 43–52. doi: 10.1007/s13042-010-0001-0
11. Cappelle B., Depraetere I., Lesuisse M. The necessity modals have to, must, need to, and should: Using n-grams to help identify common and distinct semantic and pragmatic aspects // Constructions and Frames. 2019. V. 11. N 2. P. 220–243. doi: 10.1075/cf.00029.cap
12. Weiss S.M., Indurkhya N., Zhang T., Damerau F.F. Text Mining Predictive Methods for Analyzing Unstructured Information. Springer Science+Business Media, Inc., 2010. XII, 237 p. doi: 10.1007/978-0-387-34555-0
13. Kozhevnikov V.A., Pankratova E.S. Research of the text data vectorization and classification algorithms of machine learning // Theoretical & Applied Science. 2020. N 5. P. 574–585. doi: 10.15863/TAS.2020.05.85.106
14. Canbek G., Temizel T.T., Sagiroglu S., Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights // Proc. 2nd International Conference on Computer Science and Engineering (UBMK). 2017. P. 821–826. doi: 10.1109/UBMK.2017.8093539
15. Koyejo O., Natarajan N., Ravikumar P., Dhillon I.S. Consistent binary classification with generalized performance metrics // Advances in Neural Information Processing Systems. 2014. V. 27. P. 2744–2752.
16. Lee J. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls // Accounting Review. 2016. V. 91. N 1. P. 229–250. doi: 10.2308/accr-51135

3. Jenkinson T., Jones H., Martinez J.V. Picking winners? Investment consultants' recommendations of fund managers. *Journal of Finance*, 2016, vol. 71, no. 5, pp. 2333–2370. doi: 10.1111/jofi.12289.
4. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, 2014, vol. 5, no. 4, pp. 1093–1113. doi: 10.1016/j.asej.2014.04.011
5. Ghiassi M., Olschimke M., Moon B., Arnaudo P. Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications*, 2012, vol. 39, no. 12, pp. 10967–10976. doi: 10.1016/j.eswa.2012.03.027
6. Fuller C.M., Biros D.P., Delen D. An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*, 2011, vol. 38, no. 7, pp. 8392–8398. doi: 10.1016/j.eswa.2011.01.032
7. Batura T.V. Automatic text classification methods. *Software & Systems*, 2017, no. 1, pp. 85–99. (in Russian). doi: 10.15827/0236-235X.030.1.085-099
8. Alekseeva V. Using of mining techniques in problems of binary classification. *Izvestia of Samara Scientific Center of the Russian Academy of Sciences*, 2014, vol. 16, no. 6-2, pp. 354–356. (in Russian)
9. Babaev A.M. Basic principles of natural language processing. *Dnevnik Nauki*, 2019, no. 12, pp. 14. (in Russian)
10. Zhang Y., Jin R., Zhou Z.-H. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 2010, vol. 1, no. 1-4, pp. 43–52. doi: 10.1007/s13042-010-0001-0
11. Cappelle B., Depraetere I., Lesuisse M. The necessity modals have to, must, need to, and should: Using n-grams to help identify common and distinct semantic and pragmatic aspects. *Constructions and Frames*, 2019, vol. 11, no. 2, pp. 220–243. doi: 10.1075/cf.00029.cap
12. Weiss S.M., Indurkhya N., Zhang T., Damerau F.F. *Text Mining Predictive Methods for Analyzing Unstructured Information*. Springer Science+Business Media, Inc., 2010, XII, 237 p. doi: 10.1007/978-0-387-34555-0
13. Kozhevnikov V.A., Pankratova E.S. Research of the text data vectorization and classification algorithms of machine learning. *Theoretical & Applied Science*, 2020, no. 5, pp. 574–585. doi: 10.15863/TAS.2020.05.85.106
14. Canbek G., Temizel T.T., Sagiroglu S., Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. *Proc. 2nd International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 821–826. doi: 10.1109/UBMK.2017.8093539
15. Koyejo O., Natarajan N., Ravikumar P., Dhillon I.S. Consistent binary classification with generalized performance metrics. *Advances in Neural Information Processing Systems*, 2014, vol. 27, pp. 2744–2752.
16. Lee J. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls. *Accounting Review*, 2016, vol. 91, no. 1, pp. 229–250. doi: 10.2308/accr-51135

Авторы

Воробьева Алиса Андреевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191359167, http://orcid.org/0000-0001-6691-6167, Alice_w@mail.ru

Герасимов Владислав Владимирович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, http://orcid.org/0000-0001-8099-2414, gevol.69@gmail.com

Ли Юлия Валерьевна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, http://orcid.org/0000-0003-3280-8197, J_u_l_i_a_0908@mail.ru

Authors

Alisa A. Vorobeva — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191359167, http://orcid.org/0000-0001-6691-6167, Alice_w@mail.ru

Vladislav V. Gerasimov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, http://orcid.org/0000-0001-8099-2414, gevol.69@gmail.com

Yulia V. Li — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, http://orcid.org/0000-0003-3280-8197, J_u_l_i_a_0908@mail.ru

Статья поступила в редакцию 17.03.2021
Одобрена после рецензирования 13.04.2021
Принята к печати 11.05.2021

Received 17.03.2021
Approved after reviewing 13.04.2021
Accepted 11.05.2021