УДК 004.912
МРНТИ 81.93.29
https://doi.org/10.53364/24138614_2025_36_1_10

М. Мехдиев[1], А.К. Шайханова[1*], Г.Б. Бекешова[1], И.Е. Икласова[2], К.С. Бакенова[1]

[1]Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан
[2]Северо-Казахстанский университет им. Манаша Козыбаева, Петропавловск, Казахстан

*E-mail: shaikhanova_ak@enu.kz**

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ТЕКСТОВ (TEXT MINING)

Аннотация. *В статье рассматриваются методы интеллектуальной обработки текстов (Text Mining), которые позволяют преобразовывать слабоструктурированные текстовые данные в структурированную и легко анализируемую информацию. С ростом объемов данных в цифровую эпоху Text Mining становится незаменимым инструментом анализа текстов в различных сферах. Эти технологии находят широкое применение в информационной безопасности, где анализ текстов помогает выявлять угрозы и аномалии, в здравоохранении — для обработки медицинских записей и извлечения диагностической информации, в маркетинге — для анализа потребительских предпочтений, а также в юридической практике, где автоматизация анализа документов повышает точность и снижает временные затраты.*

В статье подробно рассматриваются как традиционные статистические методы, такие как TF-IDF, Word2Vec, Latent Dirichlet Allocation (LDA), так и современные подходы, включая модели глубокого обучения на основе архитектуры трансформеров, например BERT, GPT и их производные. Современные методы демонстрируют значительные успехи в учёте контекста, анализе семантики и извлечении скрытых смыслов из текстов, что делает их незаменимыми для решения сложных задач.

Особое внимание уделено сравнению эффективности различных методов и их применимости в задачах автоматизации. Описаны возможности интеграции Text Mining для анализа больших объемов данных, выявления закономерностей и автоматизации процессов извлечения знаний. Представленные результаты исследования подчеркивают актуальность использования этих технологий для повышения эффективности работы специалистов, ускорения процессов анализа информации и решения задач в ключевых отраслях, что открывает новые перспективы для внедрения интеллектуальных систем обработки данных.

Ключевые слова: *Text Mining, интеллектуальная обработка текстов, машинное обучение, обработка естественного языка, TF-IDF, Word2Vec, BERT, GPT, автоматизация анализа текстов.*

Введение.

Современная эпоха цифровых технологий характеризуется быстрым ростом объема текстовой информации, создаваемой и распространяемой в различных областях деятельности, начиная от научных исследований и заканчивая социальными сетями и корпоративными документами. Этот огромный объем данных содержит ценную информацию, которую можно извлекать, анализировать и применять для решения разнообразных задач. Однако традиционные методы обработки данных оказываются недостаточно эффективными для работы с текстами, так как они не способны адекватно учитывать лингвистическую структуру и семантические особенности текста.

Интеллектуальная обработка текстов, также известная как Text Mining, представляет собой междисциплинарное направление, включающее методы обработки естественного языка, машинного обучения и статистического анализа. Основная цель Text Mining — преобразование слабоструктурированных текстовых данных в структурированную информацию, которая может быть использована для анализа и принятия решений. Применение технологий Text Mining позволяет решать такие задачи, как автоматическая классификация и аннотирование текстов, выделение ключевых тематических направлений, поиск скрытых закономерностей и построение предсказательных моделей на основе текстовых данных.

Актуальность исследования методов Text Mining обусловлена необходимостью автоматизации анализа текстовой информации, что особенно важно в таких областях, как информационная безопасность, медицина, маркетинг и право. Автоматизация позволяет существенно снизить трудозатраты и повысить точность анализа, что в условиях роста объемов информации становится ключевым фактором успеха. Настоящая статья направлена на обзор и анализ современных методов Text Mining, включая традиционные статистические методы, такие как TF-IDF и Word2Vec, и новейшие контекстно-зависимые модели, такие как BERT и GPT, для глубокого анализа текстов.

Материалы и методы исследования.

Для выполнения исследования использовались различные методы интеллектуальной обработки текстов. На этапе подготовки данных проводился сбор слабоструктурированных текстов из открытых источников, таких как базы данных угроз, классификаторы и реестры, с последующей их предобработкой: удаление стоп-слов, лемматизация и нормализация текстов. Для представления текстовых данных применялись как традиционные подходы, такие как TF-IDF для выделения значимых слов и Word2Vec для создания плотных векторных представлений слов, так и современные модели глубокого обучения, включая трансформеры BERT и GPT, способные учитывать контекст и семантическую многозначность текста.

В процессе исследования использовались задачи тематической классификации текстов, автоматического реферирования, кластеризации и создания текстов с заданным содержанием, а также поиск информации по ключевым словам. Для моделирования применялись искусственные нейронные сети, обученные на количественных показателях актуальности угроз, что позволило интегрировать их с большим объемом текстовых данных.

Материалами исследования выступали текстовые корпуса, содержащие описания тактик и техник злоумышленников, взятые из специализированных баз данных, таких как MITRE ATT&CK или CVE. Для работы с текстами применялись программные библиотеки для обработки естественного языка, включая NLTK, spaCy, TensorFlow и PyTorch. Обработка больших массивов данных осуществлялась с использованием вычислительных мощностей, таких как GPU и облачные платформы, обеспечивающие эффективное выполнение задач глубокого обучения. Экспериментальная часть исследования опиралась на данные, представленные в ряде научных и прикладных работ по теме Text Mining.

Результаты и их обсуждение.

Традиционно применяемые базы данных (классификаторы, реестры) содержат обширные текстовые описания угроз, уязвимостей и тактик (техник) злоумышленников. Несмотря на явные преимущества таких баз данных, работа по оценке и анализу угроз и уязвимостей на конкретном предприятии остаётся экспертно ориентированной, что требует значительных временных затрат, высокой когнитивной нагрузки и специализированных знаний. В исследовании [1] предлагается использовать количественные показатели

актуальности угроз информационной безопасности для создания искусственной нейронной сети (НС) и подачи этих факторов в качестве входных данных. В то же время современные технологии искусственного интеллекта позволяют обрабатывать большие массивы текстовых данных на естественном языке (ЕЯ), решая следующие задачи:

- поиск информации, по ключевым словам;
- тематическая классификация текстов;
- автоматическое реферирование;
- кластеризация текстов по содержанию;
- автоматическое создание текстов с заданным содержанием и т.д.

Этим вопросам посвящено множество исследований, таких как [2, 3, 4]. Направление анализа слабоструктурированных текстовых данных на ЕЯ, известное как интеллектуальный анализ текстов (Text Mining), продемонстрировало значимые результаты в работах [5, 6]. Переход к обработке текстов на естественном языке требует выполнения ряда этапов: подготовки корпуса текстовых данных, их векторного представления в многомерном семантическом пространстве и, наконец, решения задач семантического анализа.

Использование методов для создания векторных представлений слов и оценки их семантической близости является ключевым в анализе слабоструктурированных текстов на ЕЯ [7]. В современных подходах машинного обучения для таких задач применяются как классические методы (линейные методы классификации, гауссовские модели, деревья решений), так и методы обработки последовательностей (скрытые марковские модели, модели условных случайных полей). В последние годы большую популярность получили нейронные сети: многослойные перцептроны, сверточные и рекуррентные НС, для которых векторные представления слов создаются с помощью таких инструментов, как TF-IDF, Word2vec, Doc2Vec, GloVe, FastText и других.

Прежде чем перейти к обсуждению методов машинного обучения, важно рассмотреть основные этапы предварительной обработки текстовых данных. Эти этапы включают:

1. Нормализация — упрощение текста, удаление пунктуации, аббревиатур, служебных слов (например, союзов и предлогов).
2. Токенизация — разбиение текста на слова и предложения.
3. Стеммизация — приведение слов к их корню, удаление суффиксов и окончаний.
4. Лемматизация — приведение слов к их канонической форме (например, инфинитиву или именительному падежу).
5. Фильтрация — удаление нерелевантных символов и слов.

При этом следует учитывать, что данный алгоритм не сохраняет порядок слов в предложении.

Метрика TF-IDF часто применяется для анализа текстов и информационного поиска, например, для определения релевантности документа запросу, а также при кластеризации документов.

Word2Vec [8] — инструмент, разработанный группой исследователей Google под руководством Т. Миколова в 2013 году, представляет собой набор алгоритмов для создания векторных представлений слов. На вход подаётся текстовый корпус, а на выходе формируются вектора слов. Инструмент обучается на большом объеме текстов, запоминая контексты, в которых встречается каждое слово. По завершении обучения каждому слову соответствует вектор в 300-мерном пространстве признаков (семантическом пространстве), где слова, близкие по смыслу, располагаются рядом. В Word2Vec реализованы две основные архитектуры: Continuous Bag of Words (CBOW) и Skip-gram.

CBOW и Skip-gram — нейросетевые модели Word2Vec, описывающие процесс обучения и формирования векторных представлений слов. В CBOW модель предсказывает слово по контексту, в то время как Skip-gram предсказывает контекст по заданному слову.

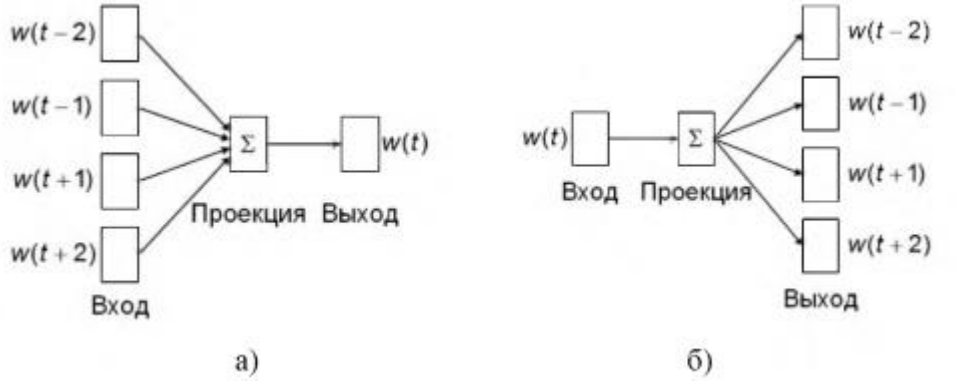

Рисунок 1 – Архитектуры алгоритмов обучения Continuous Bag of Words (а) и Skip-gram (б)

Хотя Word2Vec эффективно создает векторные представления для слов, у него имеются существенные ограничения: 1) алгоритм не учитывает порядок слов, что влияет на смысл фраз; 2) не принимает во внимание семантические значения; 3) не поддерживает формирование вектора для всего документа. Для решения этих проблем был создан инструмент Doc2Vec.

Doc2Vec (paragraph2vec) [9] — это инструмент, разработанный в 2014 году на основе Word2Vec, который позволяет формировать пространство признаков фиксированной длины для документов. Такой вектор можно применять для анализа семантической близости текстов (абзацев, документов) и использования в задачах кластеризации, классификации и прогнозирования.

Перечисленные модели относятся к первой волне обработки естественного языка (NLP). Вторая революция в NLP связана с развитием нейросетевых языковых моделей, таких как GPT и BERT, основанных на механизме внимания (Self-Attention), которые обозначаются как «трансформеры».

Первая версия GPT (Generative Pre-trained Transformer) [10], выпущенная OpenAI в 2018 году (создатели — Илон Маск и Сэм Альтман), стала прорывом в генеративных языковых моделях на архитектуре трансформера. Третья версия, GPT-3, выпущенная в 2020 году, была обучена на более чем 500 ГБ текстовых данных. Для русского языка существует аналогичная модель ruGPT-3 13B, разработанная компаниями SberDevices и SberCloud, обученная на 600 ГБ данных. Однако полное использование GPT требует мощных вычислительных ресурсов, что недоступно большинству исследователей.

Модель BERT (Bidirectional Encoder Representations from Transformers) [11] — двунаправленная трансформерная модель, предназначенная для предобучения на крупных текстовых корпусах для последующего применения в широком спектре NLP-задач. Мультиязычные версии BERT предобучены на больших наборах данных и могут быть легко интегрированы в проекты, позволяя избегать трудоёмкого обучения модели с нуля. BERT также может быть запущен как на локальном компьютере, так и на бесплатных облачных платформах, таких как Google Colab. В отличие от GPT, BERT имеет более лёгкую структуру: например, версия BERT-Large включает 24 слоя и 240 млн параметров. Модель BERT поддерживает тонкую настройку для конкретных задач, что делает её подходящей и для анализа данных в области ИБ. В отличие от Word2Vec и других традиционных языковых моделей, BERT генерирует контекстозависимые представления, что позволяет учитывать контекст и генерировать уникальные векторы для омонимов в зависимости от контекста.

Архитектура трансформера, представленного на рисунках, состоит из кодировщика (encoder) и декодировщика (decoder), каждый из которых включает слои с механизмом внутреннего внимания и НС прямого распространения (Feed-Forward NN).

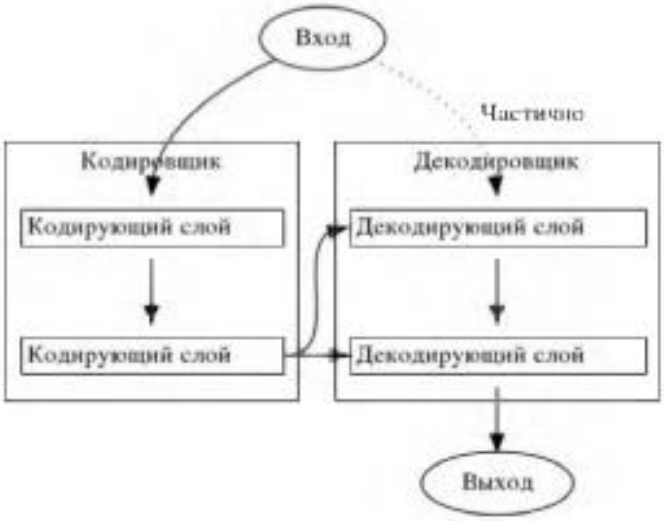

Рисунок 2 – Общая схема архитектуры трансформера

Кодировщик преобразует входные слова в векторы (эмбеддинги) в семантическом пространстве, а декодировщик генерирует последовательность выходных слов для выполнения конкретных задач, таких как классификация, поиск или перевод. Механизмы внимания выделяют ключевые слова в тексте, формируя содержательное представление текста и оптимизируя обработку.

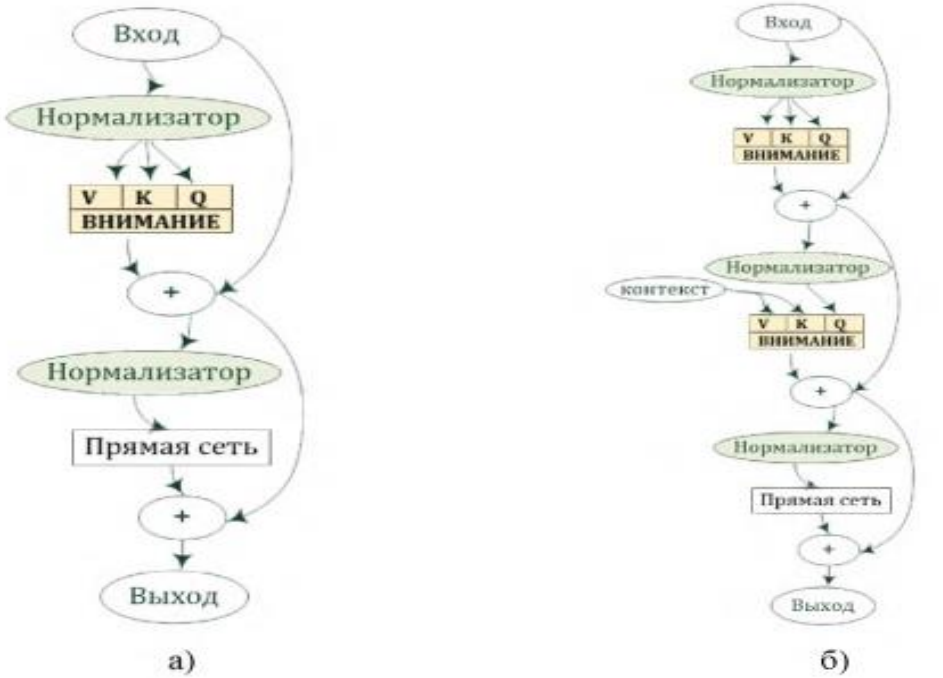

Рисунок 3 – Схема архитектуры трансформера: а) кодирующий слой;
б) декодирующий слой

Методы векторизации, такие как TF-IDF, Word2Vec и Doc2Vec, относятся к статистическим. Эти методы часто используются в задачах NLP, однако они имеют определённые недостатки:

– не учитывают многозначность и контекст использования слов, что приводит к созданию одного усреднённого вектора для омонимов и может искажать результаты анализа;

– недостаточно эффективны при обработке редких или новых слов.

Контекстуализированные модели, такие как GPT и BERT, избавлены от этих недостатков: они вычисляют векторное представление слова с учётом его контекста. В исследовании [12] проводился сравнительный анализ векторизации текстов с использованием статистических моделей и одной из трансформерных моделей для анализа диалоговых чат-ботов. Результаты показали высокую эффективность как комбинации TF-IDF + Word2Vec, так и трансформерных моделей.

Алгоритмы кластеризации широко применяются для анализа текстовых описаний и определения сходства между документами. Основная их задача заключается в выявлении скрытой структуры текстовых корпусов на основе признаков, позволяющих группировать текстовые данные по смыслу. Центральное место в решении этой задачи занимает оценка семантической близости между векторными представлениями текстов, визуализируемыми как точки в пространстве признаков. Для этого используются метрики семантической близости, или метрики расстояний [13]. Суть этих метрик заключается в вычислении расстояния между точками (векторами) в многомерном пространстве, причём более близкие точки обозначают более схожие текстовые описания.

Кластеризация представляет собой ключевой этап анализа текстовой информации, когда текстовые векторы, полученные после предварительной обработки, делятся на группы по общим признакам [14].

Существует несколько подходов к реализации алгоритмов кластеризации:

1. Иерархическая кластеризация. Эти алгоритмы создают дендрограмму вложенных кластеров, где каждый новый кластер строится из существующих, образуя древовидную структуру. На каждом шаге вычисляется расстояние между кластерами и обновляется для вновь образованных кластеров. Расстояние R (U, V) между кластерами U и V, содержащими несколько элементов, может определяться различными функциями, выбор которых зависит от задачи.

2. Плоская кластеризация предполагает разбиение объектов сразу на несколько кластеров, к одному из которых будет принадлежать каждый объект. Самый распространённый алгоритм данного типа — k-means (k-средних). Он случайным образом определяет центры кластеров и находит ближайшие к ним векторы данных. Затем центры кластеров смещаются, и процедура повторяется, пока кластеры не стабилизируются. Недостатки k-means: 1) количество кластеров задаётся вручную; 2) алгоритм чувствителен к выбору начальных центров; 3) не позволяет элементу принадлежать к нескольким кластерам.

3. Чёткая кластеризация предполагает, что каждый объект принадлежит только одному кластеру, тогда как нечёткая кластеризация допускает принадлежность объекта к нескольким кластерам с определённой вероятностью. К нечетким алгоритмам относится c-means (c-средних). Однако этот метод не всегда корректен при разном разбросе значений по осям элементов в кластерах.

При кластеризации текстов одной из ключевых проблем является высокая размерность данных [15], что усложняет процесс машинного обучения. Чтобы справиться с этой проблемой, используются методы снижения размерности, описанные в ряде отечественных и зарубежных исследований [16, 17]. Процедура основывается на выборе наиболее значимых признаков, что позволяет:

– улучшить визуализацию результатов;

– повысить точность классификации;

– сократить ресурсы на вычисления;

– ускорить обучение моделей.

Снижение размерности векторного пространства является одним из эффективных методов для экономии вычислительных ресурсов при сохранении информативности результатов [18].

Популярные алгоритмы для снижения размерности:

– t-SNE (t-distributed stochastic neighbor embedding) [19] — стохастическое вложение соседей с использованием t-распределения Стьюдента. Этот итерационный процесс визуализирует данные в низкоразмерных пространствах, сохраняя информацию об относительном расположении точек, что способствует улучшению поиска глобальных минимумов и визуализации данных. Несмотря на низкую скорость работы, t-SNE хорошо сохраняет базовую структуру данных.

– UMAP (Uniform Approximation and Projection) [20] — высокоэффективный метод для снижения размерности, который строит взвешенный граф, соединяя ближайших соседей (объекты) ребрами. Граф в низкоразмерном пространстве оптимизируется так, чтобы приблизиться к исходному графу путем минимизации дивергенции Кульбака-Лейблера. UMAP отличается высокой скоростью и отсутствием ограничений на исходную размерность пространства, превосходя t-SNE по быстродействию.

– PCA (Principal Component Analysis) [21] — метод анализа главных компонент, выполняющий линейное преобразование данных в новую систему координат, где можно описать вариации данных меньшим числом измерений. PCA характеризуется высокой скоростью вычислений, однако это достигается за счёт некоторой потери информации.

Методы, применяемые в тематической классификации текстов:

– LSA (Latent Semantic Analysis) [22] — латентный семантический анализ, который позволяет выявлять скрытые взаимосвязи между документами и терминами. Он предполагает, что слова с близким значением чаще встречаются в схожих текстах. Для анализа создается матрица, где строки представляют слова, а столбцы — текстовые описания. Путем сингулярного разложения (SVD) уменьшается количество строк, сохраняя при этом структурные связи между столбцами.

– LDA (Latent Dirichlet Allocation) [23] — метод латентного распределения Дирихле, также применяемый в тематическом моделировании. В отличие от LSA, LDA предполагает, что каждый документ представляет собой вероятностный набор тем, а каждое слово в документе ассоциировано с одной из этих тем.

На основании анализа представленных исследований можно заключить, что актуальность автоматизации обработки текстовых данных обусловлена потребностью в снижении трудоемкости и повышении эффективности работы специалистов по ИБ. Это особенно важно для задач анализа актуальных угроз ИБ и уязвимостей программного обеспечения объектов критической информационной инфраструктуры.

Технологии интеллектуального анализа текстов (Text Mining) представляют собой набор методов и алгоритмов для анализа и обработки текстовых данных. Text Mining охватывает широкий спектр задач, таких как классификация, кластеризация, аннотирование, выявление скрытых связей и прогнозирование на основе текстовой информации. Данные технологии становятся особенно востребованными в условиях растущих объемов текстовых данных, где требуется автоматизация анализа для эффективного извлечения знаний.

Технологии Text Mining открывают новые возможности для обработки текстовых данных и извлечения значимой информации, что позволяет автоматизировать анализ и значительно повысить эффективность работы специалистов в различных областях.

Заключение.

Интеллектуальная обработка текстов (Text Mining) занимает важное место в современных подходах к анализу текстовой информации, представляя собой набор

инструментов и методов, позволяющих автоматизировать процессы извлечения знаний из слабоструктурированных текстовых данных.

В условиях роста объемов информации технологии Text Mining становятся незаменимыми, обеспечивая эффективный анализ и структурирование данных в различных областях — от информационной безопасности и здравоохранения до права и маркетинга.

В данной статье были рассмотрены ключевые методы Text Mining, начиная с традиционных статистических моделей, таких как TF-IDF и Word2Vec, и заканчивая современными трансформерными моделями, такими как BERT и GPT. Каждый из методов имеет свои особенности и ограничения, которые важно учитывать при выборе подхода к решению конкретной задачи. Традиционные методы эффективны при обработке больших объемов данных с минимальными требованиями к вычислительным ресурсам, в то время как более современные модели обеспечивают высокую точность анализа за счет учета контекста, но требуют значительных вычислительных мощностей.

Преимущества использования Text Mining очевидны: автоматизация обработки текстов снижает когнитивную нагрузку на специалистов, повышает точность анализа и открывает возможности для глубокого понимания и прогнозирования на основе текстовой информации. Развитие технологий Text Mining будет способствовать дальнейшему росту эффективности и расширению применения этих методов в различных отраслях.

Список литературы

1. Жук, Р. В., Дзьобан, П. И., & Власенко, А. В. (2020). Определение актуальности угроз информационной безопасности в информационных системах обработки персональных данных с использованием математического аппарата нейронных сетей. *Прикаспийский журнал: управление и высокие технологии*, 1(49), 169-178. https://doi.org/10.21672/2074-1707.2020.49.4.169-178

2. Гузаиров, М. Б., & Машкина, И. В. (2013). Управление защитой информации на основе интеллектуальных технологий. Москва: Машиностроение.

3. Коноваленко, С. А., & Королев, И. Д. (2016). Выявление уязвимостей информационных систем. *Инновации в науке*, 9(58), 12-20.

4. Аникин, И.В. (2015). Нечеткая оценка уязвимостей, основанная на метриках CVSS V.2.0. Проблемы информационной безопасности. *Компьютерные системы*, (3), 111-117.

5. Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. *2015 IEEE international conference on intelligence and security informatics* (ISI), 85-90. IEEE.

6. Datta, P., Lodinger, N., Namin, A. S., & Jones, K. S. (2020). Cyber-attack consequence prediction. arXiv preprint arXiv:2012.00648.

7. Бондарчук, Д. В. (2017). Векторная модель представления знаний на основе семантической близости термов. *Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика*, 6(3), 73-83. https://doi.org/10.14529/cmse170305.

8. Python-School. (2024). Как работает Word2Vec: нейросети для NLP. [Электронный ресурс]. - Режим доступа: https://python-school.ru/blog/what-is-word2vec/ (дата обращения: 26.10.2024).

9. Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. International Conference on Machine Learning, 1188-1196.

10. Habr. (2024). GPT-2 в картинках (визуализация языковых моделей Трансформера). [Электронный ресурс]. - URL: https://habr.com/ru/post/490842/ (дата обращения: 26.10.2024).

11. Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. arXiv preprint arXiv:1807.03121.

12. Жеребцова, Ю. А., & Чижик, А. В. (2020). Сравнение моделей векторного представления текстов в задаче создания чат-бота. *Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация*, 18(3), 16-34. https://doi.org/10.25205/1818-7935-2020-18-3-16-34

13. Бенгфорт, Б., Билбро, Р., & Океда, Т. (2019). Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. СПб.: Питер.

14. Тюрин, А. Г., & Зуев, И. О. (2014). Кластерный анализ, методы и алгоритмы кластеризации. *Вестник МГТУ МИРЭА*, 2(3), 86-97.

15. Habr. (2024). Об одной задаче Data Science [Электронный ресурс]. - Режим доступа: https://habr.com/ru/company/mlclass/blog/266727/ свободный (дата обращения: 26.10.2024).

16. Pramoditha, R. (2021). Dimensionality reduction techniques you should know in 2021. Towards Data Science. [Электронный ресурс]. - Режим доступа: https://towardsdatascience.com/11-dimensionalityreduction-techniques-you-shouldknow-in-2021-dcb9500d388b. - 11 (дата обращения: 26.10.2024).

17. Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: A review. Complex & Intelligent Systems, 8, 2663-2693.

18. Maitra, S. (2024). Feature reduction strategy to make better generalization models. Medium. [Электронный ресурс]. - Режим доступа: https://medium.com/swlh/feature-selection-techniques-to-make-bettergeneralization-models-6a19dd6dc9b1 свободный (дата обращения: 25.10.2024).

19. Habr. (2024). Препарируем t-SNE [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/267041/ (дата обращения: 25.10.2024).

20. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

21. Золотых, Н.Ю (2013). Машинное обучение и анализ данных. [Электронный ресурс]. - Режим доступа: http: //www. uic. unn.ru: 8103/~zny/ml/Course/04. Principal%20Component%20analysis.pdf (дата обращения: 25.10.2024).

22. Habr. (2024). Латентно-семантический анализ: реализация [Электронный ресурс]. - Режим доступа: https://habr.com/ru/post/240209/ (дата обращения: 25.10.2024).

23. Kanakogi, K., Washizaki, H., Fukazawa, Y., Ogata, S., Okubo, T., Kato, T., & Yoshioka, N. (2021). Tracing CVE vulnerability information to CAPEC attack patterns using natural language processing techniques. Information, 12(8), 298.

References

1. Juk, R. V., Dz'oban, P. I., & Vlasenko, A. V. (2020). Opredelenie aktual'nosti ugroz informatsionnoi bezopasnosti v informatsionnykh sistemakh obrabotki personal'nykh dannykh s ispol'zovaniem matematicheskogo apparata neironnykh setei. Prikaspiskii zhurnal: upravlenie i vysokie tekhnologii, 1(49), 169-178. https://doi.org/10.21672/2074-1707.2020.49.4.169-178

2. Guzairov, M. B., & Mashkina, I. V. (2013). Upravlenie zashchitoi informatsii na osnove intellektual'nykh tekhnologii. Moscow: Mashinostroenie.

3. Konovalenko, S. A., & Korolev, I. D. (2016). Vyiavlenie uiazvimostei informatsionnykh sistem. Innovatsii v nauke, 9(58), 12-20.

4. Anikin, I. V. (2015). Nechetkaia otsenka uiazvimostei, osnovannaia na metrikakh CVSS V.2.0. Problemy informatsionnoi bezopasnosti. *Kompiuternye sistemy*, (3), 111-117

5. Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. 2015 *IEEE International Conference on Intelligence and Security Informatics* (ISI), 85-90. IEEE. https://doi.org/10.1109/ISI.2015.7165945.

6. Datta, P., Lodinger, N., Namin, A. S., & Jones, K. S. (2020). Cyber-attack consequence prediction. arXiv preprint arXiv:2012.00648.

7. Bondarchuk, D. V. (2017). Vektornaia model predstavleniia znanii na osnove semanticheskoi blizosti termov. *Vestnik Iuzhno-Ural'skogo gosudarstvennogo universiteta. Seriia: Vychislitel'naia matematika i informatika*, 6(3), 73-83. https://doi.org/10.14529/cmse170305

8. Python-School. (2024). Kak rabotaet Word2Vec: neiroseti dlia NLP. Retrieved October 26, 2024, from https://python-school.ru/blog/what-is-word2vec/

9. Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning*, 1188-1196.

10. Habr. (2024). GPT-2 v kartinkakh (vizualizatsiia iazykovykh modelei Transformera). Retrieved October 26, 2024, from https://habr.com/ru/post/490842/

11. Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. arXiv preprint arXiv:1807.03121.

12. Zherebtsova, Y. A., & Chizhik, A. V. (2020). Sravnenie modelei vektornogo predstavleniia tekstov v zadache sozdaniia chat-bota. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriia: Lingvistika i mezhkul'turnaia kommunikatsiia*, 18(3), 16-34. https://doi.org/10.25205/1818-7935-2020-18-3-16-34.

13. Bengfort, B., Bilbro, R., & Okeda, T. (2019). Applied text analysis with Python: Machine learning for the natural language processing. St. Petersburg: Piter.

14. Tiuirin, A. G., & Zuev, I. O. (2014). Klasternyi analiz, metody i algoritmy klasterizatsii. *Vestnik MGTU MIREA*, 2(3), 86-97.

15. Habr. (2024). Ob odnoi zadache Data Science. Retrieved October 26, 2024, from https://habr.com/ru/company/mlclass/blog/266727/

16. Pramoditha, R. (2021). Dimensionality reduction techniques you should know in 2021. Towards Data Science. Retrieved October 26, 2024, from https://towardsdatascience.com/11-dimensionalityreduction-techniques-you-shouldknow-in-2021-dcb9500d388b

17. Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: A review. Complex & Intelligent Systems, 8, 2663-2693. https://doi.org/10.1007/s40747-022-00745-5

18. Maitra, S. (2024). Feature reduction strategy to make better generalization models. Medium. Retrieved October 25, 2024, from https://medium.com/swlh/feature-selection-techniques-to-make-bettergeneralization-models-6a19dd6dc9b1.

20. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

21. Zolotykh, N. Y. (2013). Mashinnoe obuchenie i analiz dannykh. Retrieved October 25, 2024, from http://www.uic.unn.ru:8103/~zny/ml/Course/04.Principal%20Component%20analysis.pdf

22. Habr. (2024). Latentno-semanticheskii analiz: realizatsiia. Retrieved October 25, 2024, from https://habr.com/ru/post/240209/

23. Kanakogi, K., Washizaki, H., Fukazawa, Y., Ogata, S., Okubo, T., Kato, T., & Yoshioka, N. (2021). Tracing CVE vulnerability information to CAPEC attack patterns using natural language processing techniques. Information, 12(8), 298. https://doi.org/10.3390/info12080298

МӘТІНДЕРДІ ЗИЯТКЕРЛІК ӨҢДЕУ ӘДІСТЕРІ (TEXT MINING)

Аңдатпа. *Мақалада мәтінді интеллектуалды өңдеу әдістері (Text Mining) қарастырылады, олар әлсіз құрылымдалған мәтіндік деректерді құрылымдалған және*

оңай талданатын ақпаратқа айналдыруға мүмкіндік береді. Цифрлық дәуірде деректер көлемінің өсуімен Text Mining әртүрлі салаларда мәтіндерді талдау үшін таптырмас құралға айналуда. Бұл технологиялар ақпараттық қауіпсіздік саласында кең қолданысқа ие, мұнда мәтіндерді талдау қауіптер мен ауытқуларды анықтауға көмектеседі, денсаулық сақтау саласында – медициналық жазбаларды өңдеу және диагностикалық ақпаратты алу үшін, маркетингте – тұтынушылардың қалауларын талдау үшін, сондай-ақ құқықтық тәжірибеде – құжаттарды талдауды автоматтандыру арқылы дәлдікті арттыруға және уақыт шығындарын азайтуға мүмкіндік береді.

Мақалада дәстүрлі статистикалық әдістер, мысалы, TF-IDF, Word2Vec, Latent Dirichlet Allocation (LDA), сондай-ақ трансформер архитектурасына негізделген терең оқыту модельдерін, мысалы, BERT, GPT және олардың туындыларын қоса алғанда, заманауи тәсілдер егжей-тегжейлі қарастырылады. Қазіргі әдістер мәтіндерден жасырын мағыналарды алу, семантиканы талдау және контекстті есепке алуда айтарлықтай жетістіктерге жетуде, бұл оларды күрделі міндеттерді шешуде таптырмас құрал етеді.

Әр түрлі әдістердің тиімділігін салыстыруға және олардың автоматтандыру міндеттеріндегі қолданылуына ерекше назар аударылған. Text Mining технологияларын үлкен деректер көлемін талдау, заңдылықтарды анықтау және білім алу процестерін автоматтандыру үшін интеграциялау мүмкіндіктері сипатталған. Зерттеу нәтижелері осы технологияларды мамандардың жұмыс тиімділігін арттыру, ақпаратты талдау процестерін жылдамдату және негізгі салалардағы міндеттерді шешу үшін пайдаланудың өзектілігін көрсетеді, бұл интеллектуалды деректерді өңдеу жүйелерін енгізудің жаңа перспективаларын ашады.

***Түйін сөздер**: Text Mining, мәтінді интеллектуалды өңдеу, машиналық оқыту, табиғи тілді өңдеу, TF-IDF, Word2Vec, BERT, GPT, мәтінді талдауды автоматтандыру.*

METHODS OF INTELLIGENT TEXT PROCESSING (TEXT MINING)

***Abstract.** The article discusses methods of intelligent text processing (Text Mining), which allow to transform poorly structured text data into structured and easily analysed information. With the growth of data volumes in the digital age, Text Mining is becoming an indispensable tool for analysing texts in various fields. These technologies find wide application in information security where text analysis helps to identify threats and anomalies, in healthcare to process medical records and extract diagnostic information, in marketing to analyse consumer preferences, and in legal practice where automation of document analysis improves accuracy and reduces time costs.*

The paper details both traditional statistical methods such as TF-IDF, Word2Vec, Latent Dirichlet Allocation (LDA) and state-of-the-art approaches including deep learning models based on transformer architecture such as BERT, GPT and their derivatives. The state-of-the-art methods show significant advances in context-awareness, semantics analysis, and extraction of hidden meanings from texts, which makes them indispensable for solving complex problems.

Particular attention is paid to comparing the effectiveness of different methods and their applicability in automation tasks. The possibilities of Text Mining integration for analysing large amounts of data, identifying patterns and automating knowledge extraction processes are described. The presented research results emphasise the relevance of using these technologies to improve the efficiency of specialists' work, accelerate the processes of information analysis and problem solving in key industries, which opens new perspectives for the implementation of intelligent data processing systems.

***Keywords**: Text Mining, intelligent text processing, machine learning, natural language processing, TF-IDF, Word2Vec, BERT, GPT, text analysis automation.*

Сведение об авторах

Турадж Мехманоглы Мехдиев	Магистрант 2-го курса по специальности «Информационной безопасности», Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан. ORCID: https://orcid.org/ 0009-0004-6771-1584 E-mail: mehdiev.t@gmail.com
Шайханова Айгуль Кайрулаевна	PhD, профессор кафедры информационной безопасности, Евразийский национальный университет имени Л.Н.Гумилева, ORCID: https://orcid.org /0000-0001-6006-4813 E-mail: shaikhanova_ak@enu.kz
Бекешова Гульвира Бауржановна	Старший преподаватель кафедры информационной безопасности, Евразийский национальный университет имени Л.Н.Гумилева, ORCID: https://orcid.org/0000-0002-1635-4693 E-mail: bekeshova_gb@enu.kz
Икласова Кайнижамал Есимсеитовна	PhD, доцент кафедры «Информационно-коммуникационные технологии», Северо-Казахстанский университет им. Манаша Козыбаева, Петропавловск, Казахстан, e-mail: keiklasova@ku.edu.kz, ORCID: https://orcid.org/0000-0002-8330-4282.
Бакенова Камила Сериковна	Докторант 2-го курса по специальности «Информационной безопасности», Евразийский национальный университет имени Л.Н. Гумилева, Республика Казахстан. ORCID: https://orcid.org/ 0009-0004-2567-173X E-mail: bakenova_ks@enu.kz

Авторлар туралы мәлімет

Турадж Мехманоглы Мехдиев	«Ақпараттық қауіпсіздік» мамандығының 2 курс магистранты, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Қазақстан Республикасы. ORCID: https://orcid.org/ 0009-0004-6771-1584 E-mail: mehdiev.t@gmail.com
Шайханова Айгуль Кайрулаевна	PhD, ақпараттық қауіпсіздік кафедрасының профессоры, Л.Н. Гумилев атындағы Еуразия ұлттық университеті; ORCID: https://orcid.org/0000-0001-6006-4813. E-mail: shaikhanova_ak@enu.kz
Бекешова Гульвира Бауыржановна	Аға оқытушы, «Ақпараттық технологиялар» факультеті, «Ақпараттық қауіпсіздік» кафедрасы, Л.Н. Гумилев атындағы Еуразия ұлттық университеті. ORCID: https://orcid.org/0000-0002-1635-4693 E-mail: bekeshova_gb@enu.kz
Икласова Кайнижамал Есимсеитовна	PhD, «Ақпараттық-коммуникациялық технологиялар» кафедрасының доценті, Манаш Қозыбаев атындағы Солтүстік Қазақстан университеті, Петропавл, Қазақстан, e-mail: keiklasova@ku.edu.kz, ORCID: https://orcid.org/0000-0002-8330-4282.
Бакенова Камила Сериковна	«Ақпараттық қауіпсіздік» мамандығының 2 курс докторанты, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті, Қазақстан Республикасы. ORCID: https://orcid.org/ 0009-0004-2567-173X E-mail: bakenova_ks@enu.kz

Information about the authors

Mekhdiev Turaj	2st year master's degree; specialty of «Information security», Eurasian National University named after L.N. Gumilyov, The Republic of Kazakhstan. ORCID: https://orcid.org/ 0009-0004-6771-1584 E-mail: mehdiev.t@gmail.com
Shaikhanova Aigul Kairulaevna	PhD, Professor, Department of Information Security, L.N. Gumilyov Eurasian National University; e-mail: shaikhanova_ak@enu.kz. ORCID: https://orcid.org/0000-0001-6006-4813.

	E-mail: shaikhanova_ak@enu.kz
Bekeshova Gulvira Baurzhanovna	Senior Lecturer at the Department of Information Security, L.N. Gumilyov Eurasian National University, ORCID: https://orcid.org/0000-0002-1635-4693 E-mail: bekeshova_gb@enu.kz
Iklasova Kajnizhamal Esimseitovna	PhD, Acting Professor of the Department of Information and Communication Technologies, Manash Kozybayev North Kazakhstan University, e-mail: keiklasova@ku.edu.kz, ORCID: https://orcid.org/0000-0002-8330-4282.
Bakenova Kamila Serikovna	2nd year doctoral student; specialty of «Information security», Eurasian National University named after L.N. Gumilyov, The Republic of Kazakhstan. ORCID: https://orcid.org/ 0009-0004-2567-173X E-mail: bakenova_ks@enu.kz