**УДК 004.912**

# ИНСТРУМЕНТЫ PYTHON ДЛЯ СЕМАНТИЧЕСКОГО СРАВНЕНИЯ ТЕКСТОВ В КОНТЕКСТЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ.

**Стрельников Н.Р.** (ВКА)

**Научный руководитель – Михайлова С.А.** (ВКА)

**Введение.** Развитие современных технологий способствовало созданию новой для мирового сообщества угрозы. Обусловлена она в первую очередь тем, что вооружённые силы любой страны активно используют их как оружие массового поражения сознания людей.

В 1990 году Министерство обороны США разработало военную доктрину, которая ввела новый термин – «сетецентрическая война», подразумевающая достижение победы за счёт инфокоммуникационного преимущества. Это значит, что согласно современным правилам ведения войны в период боевых действий должны активно вестись информационно-психологические операции, направленные на дезорганизацию личного состава противника, формирование положительного имиджа активной стороны среди населения страны-противника, подкупа политиков [1].

Данный тезис подтверждается содержанием статьи 12 «Характерные черты современных военных конфликтов» Военной доктрины Российской Федерации , а также дополняется пунктом «г» статьи 13 «Особенности современных военных конфликтов» устанавливающим тот факт, что важной ролью в современных военных конфликтах является заблаговременное проведение мероприятий информационной безопасности для достижения политических целей без применения военной силы, а в последующем – в интересах формирования благоприятной реакции мирового сообщества на применение военной силы [2].

«Цель современной войны – завоевание или установление контроля над мировыми ресурсами жизнедеятельности человечества, установление лояльности власти в государствах, на территории которых эти ресурсы находятся, управление массовым сознанием народов и больших групп людей» – обобщают авторы статьи «Сущность войны и современность» [3].

Важной задачей при противодействии информационному воздействию является семантическое сравнение текстов для выявления вредоносного содержания. На данный момент с этой задачей успешно справляются многие проекты и компании как в нашей стране, так и за рубежом. Так, например, на популярной платформе для разработки и распространения ИИ-инструментов Hugging Face представлена подборка обученных моделей из раздела «sentence similarity» (сходство предложений), которые позволяют сравнивать между собой тексты на разных языках. Среди них модели популярного фреймворка SentenceTransformers, Пекинской академии искусственного интеллекта (BAAI), компании Alibaba-NLP и многие другие. Наибольших успехов в данной сфере достигают компании имеющие собственные проекта, обрабатывающие каждый день естественный язык. Так, например, корпорация, занимающаяся системой поисковиков Google, использует собственную Сеть знаний, а также алгоритмы Колибри, RankBrain и BERT. Российский же поисковик Яндекс – алгоритмы Королёв, Спектр, YATI и Y1 [4]. Также важно обратить внимание на компании, которые занимаются исследованиями в области искусственного интеллекта, такие как AIRI.

Необходимость разработки эффективных инструментов семантического сравнения текстов на Python обусловлена существенной ролью этого языка в научных исследованиях и развитой экосистемой библиотек для NLP. Применение таких инструментов в контексте информационной безопасности позволяет повысить качество анализа текстовой информации и обеспечить более надежную защиту от информационных угроз.

**Основная часть**. Предлагаемое решение базируется на использовании Python и его возможностей для обработки естественного языка. Основу составляют современные методы машинного обучения, включая глубокие нейронные сети и модели трансформеров. Одним из ключевых инструментов является библиотека NLTK (Natural Language Toolkit), которая предоставляет широкий спектр функций для обработки и анализа текстовых данных.

Для семантического сравнения текстов используются предобученные модели трансформеров, такие как BERT (Bidirectional Encoder Representations from Transformers). Эти

модели способны учитывать контекст слов в предложении, что значительно повышает точность семантического анализа. Использование русскоязычных версий моделей, например, DeepPavlov или rubert, позволяет эффективно работать с текстами на русском языке.

Процесс семантического сравнения включает несколько этапов: Предобработка текста: включает очистку от лишних символов, нормализацию, токенизацию и стемминг или лемматизацию. Это необходимо для приведения текста к единому виду и уменьшения размерности данных. Преобразование текста в векторное представление: с помощью моделей эмбеддинга слов, таких как Word2Vec, GloVe или более современных трансформеров. Это позволяет представить слова и фразы в виде числовых векторов, отражающих их семантическое значение. Вычисление степени сходства: применяется косинусное расстояние или другие метрические функции для определения степени схожести между векторными представлениями текстов. Анализ результатов: на основе полученных данных можно выявлять семантически близкие тексты, кластеры тем, а также обнаруживать аномалии или дезинформацию.

Предлагаемый метод является экономичным, так как основывается на бесплатных и открытых инструментах. Новизна подхода заключается в интеграции современных моделей трансформеров в процесс семантического сравнения для задач информационной безопасности. Это обеспечивает более глубокое понимание контекста и позволяет обнаруживать скрытые связи между текстами. Кроме того, Python предоставляет возможности для автоматизации и масштабирования решения. С помощью библиотек scikit-learn и TensorFlow можно создавать корректные и эффективные модели машинного обучения, адаптированные под конкретные задачи и объемы данных.

**Выводы**. Разработанные инструменты семантического сравнения текстов на Python обладают широким потенциалом практического применения в сфере информационной безопасности. Они позволяют автоматизировать процесс анализа больших объемов текстовой информации, повышают эффективность обнаружения дезинформации и информационных угроз. Предлагается внедрение данных инструментов в системы мониторинга информационного пространства, что позволит оперативно реагировать на появление негативного или опасного контента. Также они могут быть использованы в аналитических отделах организаций для глубокого анализа информационных потоков и принятия обоснованных решений. Для успешного внедрения необходимо провести испытания инструментов в реальных условиях, адаптировать модели под специфические задачи и данные. Обучение персонала и интеграция с существующими системами также являются важными шагами в процессе внедрения.

Таким образом, использование средств семантического сравнения текстов на основе Python и современных моделей машинного обучения открывает новые возможности в сфере информационной безопасности, способствует повышению информационной безопасности и эффективности аналитической работы.

**Список использованных источников**:

1. Военное обозрение: Современная сетецентрическая война и военная операция на Украине. - URL: https://topwar.ru/211525-sovremennaja-setecentricheskaja-vojna-i-voennaja-operacija-na-ukraine.html (дата обращения 05.02.2025).

2. Президент России: Военная доктрина Российской Федерации. - URL: http://www.kremlin.ru/supplement/461#sel=22:2:yA5,22:8:GCF (дата обращения 05.02.2025).

3. Армейский сборник: Сущность войны и современность. - URL: https://army.ric.mil.ru/Stati/item/334161/ (дата обращения 05.02.2025).

4. Дзен. Digital-агентство Idea Promotion. Семантический поиск в Яндекс и Google. - URL: https://dzen.ru/a/YNBUU6kOXDaNSOuJ (дата обращения 05.02.2025).