УДК 004.896

ИНСТРУМЕНТ СБОРА И КЛАССИФИКАЦИИ ИНФОРМАЦИИ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ

Стрельников Н.Р.[1]
Научный руководитель – Михайлова С.А.[1]
1 – Военно-космическая академия имени А.Ф. Можайского
vka@mil.ru

Введение

Мировая политическая обстановка позволяет фиксировать глобальное изменение принципов ведения боевых действий между суверенными государствами. Данной тенденции в первую очередь способствует широкомасштабное развитие современных технологий как в сфере разработки и проектирования вооружения и военной техники, так и в задачах непосредственного воздействия на противника невоенными силами. В исследованиях ряда отечественных авторов [1] отмечается сходство стратегий ведения войны различными государствами. В одном из таких примеров автор [2] вводит концепцию бесконтактной войны, определяя её как «военные действия, в которых решающая роль отводится массированному применению высокоточных систем вооружений в неядерном оснащении с межконтинентальной дальностью действия, оружию на новых физических принципах и средствам информационного противоборства с целью поражения, в первую очередь, объектов экономического потенциала».

Основными инструментами воздействия на население страны-противника является манипуляция и дезинформация с помощью информации из открытых источников [3, 4].

Актуальной задачей защиты от такого рода воздействия является сбор и классификация информации из открытых источников, с целью своевременного принятия мер по цензурированию и противодействию. Решение задачи требует эффективных инструментов обработки больших данных. Классический поиск по ключевым словам не учитывает морфологию и семантику формулировок, что снижает качество выполнения задач. Возникает задача сочетания лингвистической обработки текста (лемматизации) с семантическим поиском по эмбеддингам.

На практике всемирно используют поисковые движки (например, ElasticSearch, Yandex Search API) и предобученные языковые модели (например, sentence-transformers), которые позволяют находить результат на основе запросов по аналогии с поисковыми запросами браузеров. Однако подход сочетания семантического поиска, лингвистического анализа и фильтрации по метаданным в едином инструменте для мониторинга описан недостаточно.

Основная часть

В проекте реализована веб-система мониторинга открытых каналов из мессенджера Telegram с помощью библиотеки Telethon, выполняющая задачи поиска каналов по фильтрам, поиска по семантическому сходству и ключевым словам, а также формирования формализованного отчёта и экспорта его в удобный формат для дальнейшего анализа специалистами различных сфер.

В рамках инструмента реализована фильтрация каналов по ключевым словам и метаданным: количество подписчиков, категория, наличие регистрации канала в базе Роскомнадзора и т. д. Поиск информации в отобранных каналах осуществляется с применением сочетания семантического поиска по индексу векторной базы данных

FAISS и традиционного поиска по предобработанным текстам. Результат обрабатывается с помощью дополнительных критерием в базе данных PostgreSQL с расширением pgvector.

Особенностью системы является организация предобработки и хранения данных. Из текстов постов удаляются пиктограммы (эмодзи) и стоп-слова из расширенного русскоязычного списка, текст приводится к нижнему регистру и проходит процесс лемматизации. Обработанный текст фиксирует в отдельном поле в базе данных для дальнейшего анализа. Эмбеддинги с размерностью 384, сформированные с помощью предварительно обученной модели paraphrase-multilingual-MiniLM-L12-v2, индексируются в векторную базу данных FAISS с возможностью инкрементального обновления.

Эффективность решения обеспечивается за счёт использования комбинированного подхода к использованию готовых языковых моделей и инструментов обработки русскоязычных текстов в сочетании с индексацией в векторную базу данных FAISS. Инновационность подхода заключается в комбинации семантического поиска, фильтров, ключевого поиска и предобученных языковых моделей для сбора и классификации различного рода неструктурированной информации.

Выводы

Разработанная система может применяться в отделах мониторинга СМИ, медианалитических центрах и подразделениях, обрабатывающих большие данные из открытых источников. В дальнейшем целесообразно провести внедрение системы в работу на ограниченном наборе каналов с целью анализ узконаправленной тематики. Замерить показатели точности, проанализировать полученные результаты на практике и выработать стратегию развития системы.

Литература

1. Копичев О. А., Николаев А. Е. Современные войны: анализ тенденций развития межгосударственного противоборства, классификация форм и способов борьбы, формирование признаков и критериев военного конфликта // Системы управления, связи и безопасности. 2021. № 1. С. 1-32. DOI: 10.24411/2410-9916-2021-10101.
2. Слипченко В. И. Войны шестого поколения. Оружие и военное искусство будущего. – М.: Вече, 2002. – 381 с.
3. Аскерова Л.Ф. Информационная война как вид манипуляции // Гуманитарные научные исследования. 2017. № 6 [Электронный ресурс]. URL: https://human.snauka.ru/2017/06/24211 (дата обращения: 16.02.2026).
4. Фролов Д. Б. «Информационная война: эволюция форм, средств и методов» // «Социология власти». — 2005. — №5. — С. 121–146.