УДК 004.8

РАЗРАБОТКА АЛГОРИТМА ЗАЩИТЫ ОТ СОСТЯЗАТЕЛЬНЫХ АТАК НА СИСТЕМЫ АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ

Веневцев И.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – к.т.н., ассистент Коржук В.М.

(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Данная работа описывает типы атак на системы обработки естественного языка. Также доклад содержит алгоритм противодействия данным типам атак.

Введение.

В настоящее время активно используются информационные технологии как в повседневной жизни людей, так и в работе на предприятиях. Согласно исследованиям, около 80% данных на предприятии генерируются в виде текстов. Именно для обработки такого большого объема данный существуют системы обработки естественного языка. При неверной интерпретации полученного текста данной системой предприятия могут понести убытки. Именно на это направлены состязательные атаки на системы обработки естественного языка.

Основная часть.

Для решения данной проблемы предлагается разделить состязательные атаки на два типа, а именно основанные на визуальном сходстве и основанные на семантическом сходстве. Это разделение зависит от того каким именно способом взаимодействуют с текстом, а остается ли цела визуальная или смысловая составляющая.

При получении теста система обработки естественного языка должна произвести меры по защите от атак, основанных на визуальном сходстве. А именно произвести проверку на орфографические ошибки и использовать seq2seq модель для их исправления.

После следует провести защиту от атак, основанных на семантическом сходстве. Для этого следует измерить семантическое сходство. За счет этого система не будет критично воспринимать семантически схожие слова.

Выводы.

Предложенный алгоритм направлен на увеличение защищенности систем обработки естественного языка. Различные этапы работы алгоритма могут быть внедрены в уже имеющиеся системы за счет использования дополнительных программных блоков для проверки. В перспективе исследования планируется программная реализация предложенного алгоритма, а также математическое усовершенствование его составных частей.

Веневцев И.В. (автор) Подпись

Коржук В.М. (научный руководитель) Подпись