УДК 004.442.83

МОДУЛЬ ПРОГРАММЫ АНАЛИЗА КОРОТКИХ ТЕКСТОВЫХ СООБЩЕНИЙ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ НА ПРЕДМЕТ СБОЯ И МОШЕННИЧЕСТВА, ДЛЯ БАНКОВСКОЙ ОРГАНИЗАЦИИ, С ПОМОЩЬЮ МЕТОДОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Васильев М.В. (Университет ИТМО)
Научный руководитель – доцент, кандидат технических наук Воробьева А.А.
(Университет ИТМО)

В настоящее время существует потребность в извлечении и анализе полезной и скрытой информации из многочисленных онлайн-источников, которые хранятся в виде текста и написаны на естественном языке в социальных сетях. В связи с этим, большинство компаний, а также банковских организаций, которые занимаются мониторингом социальных сетей для выявления и предотвращения сбоев, инцидентов и репутационного ущерба. Для определения тематики коротких текстовых сообщений из социальных сетей, о сбое и мошенничестве, предлагается использовать методы тематического моделирования.

Введение

Своевременное реагирование на сбои, мошеннические операции является приоритетной задачей для банковской организации как для обеспечения безопасности, также для минимизации репутационного ущерба. В связи с актуальностью данной проблемы, большинство компаний занимаются мониторингом социальных сетей для выявления и предотвращения сбоев, инцидентов и репутационного ущерба. Потенциальными потребителями результатов данного исследования, а именно сотрудника линии поддержки SOC центра, обозначена потребность автоматизации процесса выявления коротких текстовых сообщений, размещенных на веб-сайтах социальных сетей и содержащих информацию о технических сбоях и банковском мошенничестве. В связи с этим отмечена необходимость выявления сообщений релевантных именно для заданной банковской организации. На сегодняшний день, отечественные и большинство зарубежных систем предлагают решения, которые позволяют автоматизировать поиск сообщений по заданному бренду и реализуют функционал по анализу тональности. Однако, они не имеют функциональной возможности решить задачи, обозначенные потенциальным потребителем. Отечественные решения определения тематики сообщения не предполагают автоматическую обработку текстовых сообщений с использованием тематического моделирования и не предлагают использовать данный подход для выявления потенциальных рисков безопасности и репутации для банковской организации.

Основная часть

Процесс определения тематики сообщений подразумевает первичную обработку приходящих сообщений, такую как: приведение текстов сообщений к нижнему регистру, удаление управляющих (непечатных) символов и спецсимволов, лемматизацию и токенизацию. Обработанные сообщения поступают на обученную модель, использующую метод Гиббсовской пробоотборной модели Дирихле (GSDMM), который является наиболее популярным и удобным способом для анализа коротких текстовых сообщений. Этот алгоритм группирует документы и извлекает структуры тем, которые присутствуют в пришедшем наборе данных. Если для количества тем задано большое значение, модель сможет автоматически узнать количество тем. После анализа сообщения происходит присваивание соответствующей ему тематики, в нашем случае типа объекта на котором произошел сбой – банкомат либо ДБО мобильный банк, либо вид мошенничества – мошенничество по телефону,

либо с банкоматами. Для обеспечения наиболее оптимального времени выявления тематики поступившего сообщения, предлагается отправлять их на блок системы выявления типа объекта, на котором произошел сбой или вида мошенничества в ограниченном объеме, от 1 до 100 сообщений. Данные значения были выбраны из-за особенностей реализации библиотеки метода GSDMM и удобства их качественной обработки сотрудниками поддержки SOC центра банковской организации. Обеспечение автоматизации процесса будет происходить с помощью реализации клиент-серверного приложения, также состоящего из других блоков, для наиболее качественного выявления тематики проблемы с типом объекта и его географическими координатами. Данное приложение будет включать в себя: модуль загрузки данных сервисов с комментариями и отзывами из различных социальных сетей с помощью API и сохранения их в базу данных. Далее модуль выявления категории проблемы будет получать сохраненные в эту базу данных сообщения и производить их дальнейшую обработку с определением типа объекта или типа мошенничества.

Выводы

На основе проведенного анализа и методик определения тематики сообщений с использованием тематического моделирования, может быть реализована система выявления сообщений о сбое или мошенничестве предназначенная для банковской организации, для более оперативного выявления проблем и принятия решений, снижающих риски для этой организации.

Васильев М.В. (автор)

Воробьева А.А. (научный руководитель)