Отчет за 2-й этап выполнения проекта Российского научного фонда № 18-71-10094 "Мониторинг и противодействие вредоносному влиянию в информационном пространстве социальных сетей"

Развитие информационных технологий и появление глобальной сети Интернет, социальных сетей, средств массовых коммуникаций подводит научное сообщество к кардинальному пересмотру имевшихся представлений о способах ведения войны и мире. Не только в научных кругах, но и в средствах массовой информации обсуждается аспект, касающийся национальной безопасности, а именно информационно-психологическая безопасность и информационные войны. В стратегиях ведущих стран мира в число сфер ведения боевых действий помимо земли, моря, воздуха и космоса вошло информационное пространство.
Наибольшая острота этой проблемы проявляется при попытках распознавания целевых информационных воздействий (атак) на субъектов (индивидуальных или коллективных, например, на человека, семью, группу или организацию) и оказания противодействия таким атакам.
Данный проект "Мониторинг и противодействие вредоносному влиянию в информационном пространстве социальных сетей" может быть разделен на 2 основные взаимосвязанные части – мониторинг и противодействие.

В рамках мониторинга осуществляется сбор, предварительная обработка и семантический анализ информационных объектов, а также выявление источников, каналов распространения и целевой аудитории вредоносного информационного воздействия. Для успешного противодействия такому влиянию необходимо учитывать возможные меры противодействия, способы выявления цели противодействия, оценку её результативности и осуществлять поддержку принятия решений при выборе наиболее перспективной меры противодействия из доступных.

Первый этап анализа социальной сети – это сбор данных. В ходе работы над проектом разработан программный прототип с графическим интерфейсом пользователя, осуществляющий сбор и предварительную обработку разнородных объектов социальной сети.

Для анализа собираемых данных были разработаны несколько прототипов: прототип для семантического анализа (анализирует текст методами машинного обучения), прототип для анализа атак (анализ источника атаки и на кого она направлена) и прототип для анализа каналов распространения (анализ того как именно атака распространяется).

Для анализа текстового содержимого социальных сетей был разработан прототип компонента семантического анализа вредоносных информационных объектов в социальных сетях, использующий методы интеллектуального анализа данных. В рамках компонента выполняется классификация текстов, которые содержатся в потенциально вредоносных постах в социальных сетях с помощью методов обучения на основе положительных и неразмеченных данных (англ. Positive-Unlabeled learning) и многоклассовой классификации. Разработанный прототип был экспериментально исследован на наборе данных, который содержит посты на шесть потенциально вредоносных тем и текстовую информацию из постов социальной сети на случайную тему. Сравнение результатов экспериментального исследования позволяет сделать выводы о достаточно высоком качестве и стабильной работе предложенного подхода к выявлению вредоносных информационных объектов.

Для анализа атак также были разработаны соответствующие методики и программный прототип. Информационная атака в качестве своих составляющих всегда содержит источник атаки и целевую аудиторию. Разница между ними заключается во времени возникновения вредоносных информационных объектов – объекты, размещаемые источником атаки, попадают в заданный экспертом временной интервал [t0, t0+e], остальные объекты, получившие доступ к вредоносным объектам или участвовавшие в их распространи после времени t0+e, относятся к целевой аудитории. Алгоритм анализа источника атаки, имеющей стихийную природу, сегментирует ее аудиторию по степени активности, позволяя тем самым приблизиться к аккаунтам-зачинщикам информационного воздействия. Алгоритм анализа целевой аудитории дает представление о половозрастном характере аудитории воздействия и о ее локации. Предложенные алгоритмы реализованы в программном прототипе, результат их работы в виде графиков и круговых диаграмм доступен в графическом интерфейсе пользователя.

Для анализа каналов распространения информации был разработан прототип их анализа. Каналом распространения является аудитория, которая хоть и не является инициатором атаки, но способствует ее распространению. Эту аудиторию можно представить в виде сети, где пользователи и группы, распространяющие информацию, связаны друг с другом. Для анализа были разработаны два модуля: параллельной обработки и визуализации. Модуль параллельной обработки используя GPU анализирует структуру сети и позволяет вычислить: ключевые узлы сети (лидеров мнений), ширину канала, каркас канала, социальные круги канала и оценить затухание информации в канале. Модуль визуализации позволяет визуализировать канал в виде 3D изображения, и оператор может визуально оценить его. Эти модули были реализованы как программные прототипы, а также протестированы на примере анализа 5ти различных каналов распространения.

Сбор данных, текстовый анализ, анализ атак и каналов распространения представляют собой звенья одной цепи, называемой мониторингом информационного пространства социальных сетей. Для него был разработан прототип, объединяющий вышеперечисленные прототипы в единую систему. Этап мониторинга состоит из фазы отслеживания, на которой происходит контроль социальных связей и поиск потенциально вредоносных информационных объектов, и фазы выявления опасного влияния, на которой происходит построение графа распространения информации и вычисление мер центральности, отделение источника атаки от целевой аудитории, сегментация и статистический анализ. Разработанный программный прототип реализует все этапы мониторинга и позволяет оператору контролировать весь процесс от обнаружения информационной атаки на этапе сбора данных до визуализации и выгрузки результатов их анализа.

Для того, чтобы противодействовать атакам в социальных сетях было необходимо определить:
1) как можно противодействовать атаке?
2) как определить цели противодействия?
3) как определить к какой цели какую меру нужно применять?
4) как оценить эффективность противодействия?

Для этого были разработаны соответствующие методы, модели и методики.

Способы противодействия многообразны и могут быть классифицированы по различным признакам. Поэтому мы разработали классификацию, которая позволяет разделить способы противодействия на основе объекта, типа, метода, широты и исполнителя воздействия, а также группы, к которой данный способ принадлежит. Разбиение способов противодействия на отдельные группы основано на ожидаемом результате, а именно: переключении внимания, размытии внимания, дискредитации, снижении интенсивности или блокировки объекта воздействия. На основе предложенной классификации, разработана модель способа противодействия. Сформированный список способов противодействия представляет собой единую таблицу, где строки отображают примеры способов противодействия для каждой из выделенных групп, а столбцы - их место в предложенной классификации. Данный список позволит в дальнейшем для каждого из способов определить цели, на которые он должен быть направлен, а также предложить оператору системы наиболее эффективные из них.

Для того чтобы выбирать цели противодействия был разработан соответствующий метод. В рамках проекта исследованы структуры социальных сетей и взаимодействия между элементами, способы распространения информации и меры, направленные на противодействия им. На основе исследованных данных построена модель выбора цели для реализации мер противодействия вредоносной информации в социальных сетях. Предложена пошаговая методика оценки сложности для реализации мер противодействия, которая позволяет ранжировать меры по степени сложности их выполнения. При этом настройка системы достаточно гибкая, оператором может быть, как исполнительный орган власти, так и родитель, защищающий ребенка от воздействия вредоносной информации. В зависимости от того, кто является оператором системы, он настраивает ее изначально по методике выбора коэффициентов сложности на основе экспертных оценок. Также в проекте для оптимального применения выбранных мер, предложена методика оценки приоритетов объектов, на которые будет оказываться воздействие в процессе противодействия вредоносной информации. Совокупность разрозненных методик объединена в единый метод, определяющий последовательность выполнения методик и формирование итоговой общей оценки противодействия.

После формирования списка мер противодействия и списка целей, оператору системы необходимо принять решение о том, какие меры противодействия и к кому их стоит применять. Для это был разработан метод, который позволяет предоставить меры противодействия и цели воздействия в виде изображения – матрицы. Метод включает в себя три методики:
1) Методика формирования модели визуализации – формирование модели данных необходимых для визуализации.
2) Методика отрисовки модели визуализации в виде матрицы. В данной матрице строки являются объектами, столбцы мерами, а ячейки показателями эффективности.
3) Методика визуального выбора мер противодействия оператором. С помощью данной методики оператор фильтрует и сортирует элементы матрицы, а потом выбирает к кому какие меры стоит применять.
Таким образом был разработан метод, который формирует изображение, с помощью которого можно выбрать какие именно меры противодействия следует применять.

Для того чтобы оценить эффективность работы противодействия были разработаны методы, модели и методики оценки результативности и эффективности применения мер противодействия вредоносному влиянию в социальных сетях.

Под эффективностью, в контексте данного исследования понимается соотношение между достигнутым результатом и использованными ресурсами. Первая разработанная методика оценки эффективности применения мер противодействия основана на использовании различных характеристик информационных объектов и пользователей, например, прироста количества просмотров в единицу времени, прирост количества лайков в единицу времени, прироста количества комментариев в единицу времени, прироста количества репостов. Вторая разработанная методика оценки эффективности применения мер противодействия основана на оценке состояния пользователя в некоторый момент времени, так как конечная цель противодействия вредоносному влиянию в социальных сетях состоит в минимизации заинтересованности пользователя во вредоносном контенте. Предложенные метрики и методики предлагается использовать в рамках интегрированного метода оценки эффективности мер противодействия вредоносному влиянию в социальных сетях. На первом шаге метрики оценки эффективности мер противодействия для каждого класса целей (отдельный информационный объект, пользователь, группа пользователей), могут быть применены для оценки ожидаемого эффекта от принятой меры. Для оценки состояния пользователя по отношению к вредоносному контенту и формирования окончательных выводов относительно эффективности применённой меры противодействия вредоносному влиянию предлагается использовать вторую методику и показатель состояния пользователя социальной сети.

Также в рамках выполнения проекта исполнители представили результаты на 5 конференциях в России и за рубежом. Опубликованы 4 публикации, входящие в индексы цитирования Scopus и/или Web of Science и 8 публикаций входящий в систему РИНЦ. Зарегистрировано 4 программы для ЭВМ. На следующий этап проекта (2020-2021 годы) запланированы две защиты диссертаций на соискание степени кандидата технических наук с использованием результатов данного проекта.

В Санкт-Петербургском государственном университете телекоммуникаций им проф. М.А. Бонч-Бруевича проведен курс практик для магистров 1-го семестра 2019/2020, в котором использовались результаты, полученные в рамках проекта.