электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

 # 06, июнь 2016

УДК 004.048

Выявление ботов в социальных сетях: обзор работ

Бушуев В.В., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерная безопасность»

Иванников П. В., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерная безопасность»

Сосенко А. С., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерная безопасность»

Научный руководитель: Ключарёв П.Г., к.т.н., доцент
кафедры «Информационная безопасность»,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
bauman@bmstu.ru

Социальные сети с каждым днем проникают в нашу жизнь все глубже и глубже. Появление носимых гаджетов, таких как смартфоны и планшеты лишь ускорило этот процесс. Развитие мобильного интернета, его доступности, распространение новых стандартов также влияет на этот процесс.

При помощи социальных сетей мы проецируем круг нашего общения в интернет, где существуют четкие разграничения людей по группам интересов. Сам процесс общения становится быстрым, удобным, а люди более коммуникабельными.

Ведущие СМИ и коммерческие организации тоже не могут больше обходить стороной социальные сети. Публикации в них информации, новостей или рекламы быстрее находят заинтересованных лиц, появляются возможности контекстной рекламы, SEO-менеджмента. Появляется возможность публиковать востребованную информацию, которая быстрее становится актуальной в определенных кругах.

С другой стороны, встает вопрос влияния на умы людей, манипулирование ими. Так, например, в проведенном психологом Робертом Эпштейном в 2013 году исследовании показывается, что компания Google может сама выбрать президента США, изменяя результаты ответов поисковика.

Возможность отказаться от дорогостоящей рекламы на телевидении притягивает в среду социальных сетей все больше коммерческих организаций. Действительно, зачем тратить круглую сумму на рекламу по телевидению, которую увидят все подряд, когда можно потратить гораздо меньше ресурсов и показать рекламу сразу целевой аудитории?

Становится очевидным, что пользователь, попадая в круг лиц с определенными интересами, рано или поздно начинает уподобляться общему поведению, «вливаться в коллектив» и следовать общепринятым в этом обществе правилам.

Все это подтолкнуло смекалистых разработчиков ПО к созданию нового объекта социальных сетей – ботов. В этой области их можно определить, как специальные программы, целью которых является подражание поведению типичного пользователя и удовлетворение политических или экономических интересов определенных лиц.

Боты могут заниматься абсолютно разными вещами: ставить определённым людям «лайки», распространять чьи-то посты, распространять ссылки на внешние ресурсы, рекламу и т.д.

Особенно стоит обратить внимание на класс ботов, целью которых является влияние на политическую ситуацию в отдельных частях света. С одной стороны, кажется безобидным, когда при помощи ботов какая-то поп-звезда приобретает большую популярность. С другой, когда при помощи ботов появляется возможность продвигать того или иного политического кандидата, необходимо суметь противодействовать такого рода деятельности. Особенно опасно это может быть для демократических основ общества.

В связи с этим, в первом квартале 2015 года DARPA провела «Соревнования по выявлению Twitter ботов»: 4 недельный конкурс, направленный на тестирование эффективности обнаружения ботов различными методами.

Целью участвующих команд было не просто выявление ботов среди обычных пользователей, но и идентификация тех, кто занят именно влиянием на массы. Более сложное задание заключалось в отделении ботов, участвующих в одной теме от других. И самым лучшим результатом считалось, если команда смогла в полуавтоматическом режиме определить группы настроений и разделить найденных ботов по ним.

В соревновании участвовало 6 команд, две из которых показали наиболее хороший результат. Особое внимание было уделено скорости нахождения всех ботов, ведь это важно, вовремя разоблачить попытки влияния на избирательную компанию или, что еще хуже, публикации постов, удовлетворяющих интересы террористов. В статье,

приуроченной к данному соревнованию описаны методы, которыми пользовались эти команды, на них мы и предлагаем взглянуть.

Все команды без исключения отметили, что использовать только нейронные сети и компьютерное ПО недостаточно. Для эффективного результата требуется применение полуавтоматического процесса, когда в нем участвует непосредственно человек.

Некоторые команды использовали накопленный ранее опыт по обнаружению ботов. Они старались обозначить типичный образ профиля для бота, чтобы начать двигаться в сторону распознавания более сложных комбинаций. К этому добавилось знание программ, способных генерировать ботов, и их особенностей.

Так, например, команды пытались определить профили у которых синтаксис публикаций совпадал с синтаксисом уже известных ботов. Уделялось внимание среднему числу хэштегов, упоминаний, ссылок, спец. символов, ретвитов. Если публикация заканчивалась пунктуационно правильно (точкой, восклицательным знаком) это тоже давало команде повод уделить данному профилю большее внимание.

Обозревалось семантическое поведение пользователей: как часто у него появляются публикации, связанные с данной темой, какое при этом он испытывает настроение, существуют ли противоречия в настроении пользователя с течением времени. Рассматривалось так же количество языков, на котором вел свой профиль пользователь – большое их разнообразие наталкивало на мысль, что аккаунт может быть ботом.

Временные характеристики, связанные с аккаунтами, также давали немало информации: предсказуемо ли время публикации, продолжительность самого длинного сеанса пользователя без перерыва, среднее количество публикаций в день – все это могло помочь акцентировать внимание на конкретных аккаунтах.

Рассматривались и типичные параметры профилей, например, такие как наличие фотографии и ее индивидуальность, наличие уникального имени, количество подписчиков и подписок, количество используемых устройств для выхода в интернет или GPS координаты, упоминающиеся в профиле.

Команды обращали внимание на взаимодействие пользователей с уже известными ботами, например, не является ли пользователь подписчиком бота или наоборот. Команды, использующие методы кластеризации, более внимательно относились к пользователям, попавшим в кластеры с большим количеством ботов.

Все вышеперечисленные особенности периодически обновлялись в процессе обнаружения новых ботов. Это требовало использования более универсальных методов

анализа и обновления изначальных баз. Некоторые команды использовали созданные самостоятельно панели анализа ботов.

На рис. 1 показан главный экран, предоставляя общую картину анализа всех пользователей.

Image	View	Active	Label	Tweeter Id	Screen Name	Name
					contains	contains
	View	false	bot	2895972942	CarrieWoolf	Carrie Woolf
	View	false	bot	2895957810	susan_east	Susan Eastwood
	View	false	bot	2892378024	rhondapgranger	Rhonda P Granger
	View	false	bot	2881282263	MildredMason19	Mildred G. Mason
	View	true		2874925778	NicoleSGeorge	Nicole George
	View	true		2873330546	Good_Aftenoon	Good Afternoon
	View	true		2870084108	RStationery	Ruby willow wedding
	View	true	profile imag...	2867441005	senpai156	TXGRIZ89
	View	true	human	2865325585	JanetteSohm	Janette Sohm
	View	true		2864158225	real_sex_story	Veil
	View	true		2863574275	DefendingBeef	Defending Beef

Рис.1. Главный экран панели управления

Каждый пользователь имеет свой флаг, которые говорил, что он является ботом, человеком или имеет другую метку. Некоторые пользователи помечены "ботами" (эти пользователи идентифицированы как боты системой и подтверждены с помощью человека). Другие пользователи могут быть помечены как "человек" или с некоторыми другими флагами (например, "несоответствие профиля изображения"). Во время кластеризации и обучения использовались эти метки, что позволяло открыть дополнительные группы пользователей, имеющие свои особенности.

Общая характеристика описывала, насколько полон профиль и описание отдельной личности (например, отношение подписок к подписчикам). Многие из этих переменных могли намекнуть аналитикам о причастности пользователя к ботам. Панель управления позволяла аналитикам запрашивать профили, сортировать их в порядке убывания столбцов.

На рис. 2 показаны детали пользователя gunslinger_mk1, который не имел изображение профиля, а потом оно у него появилось во время проведения соревнования. В верхней части экрана показано, что данный пользователь классифицируется как бот.

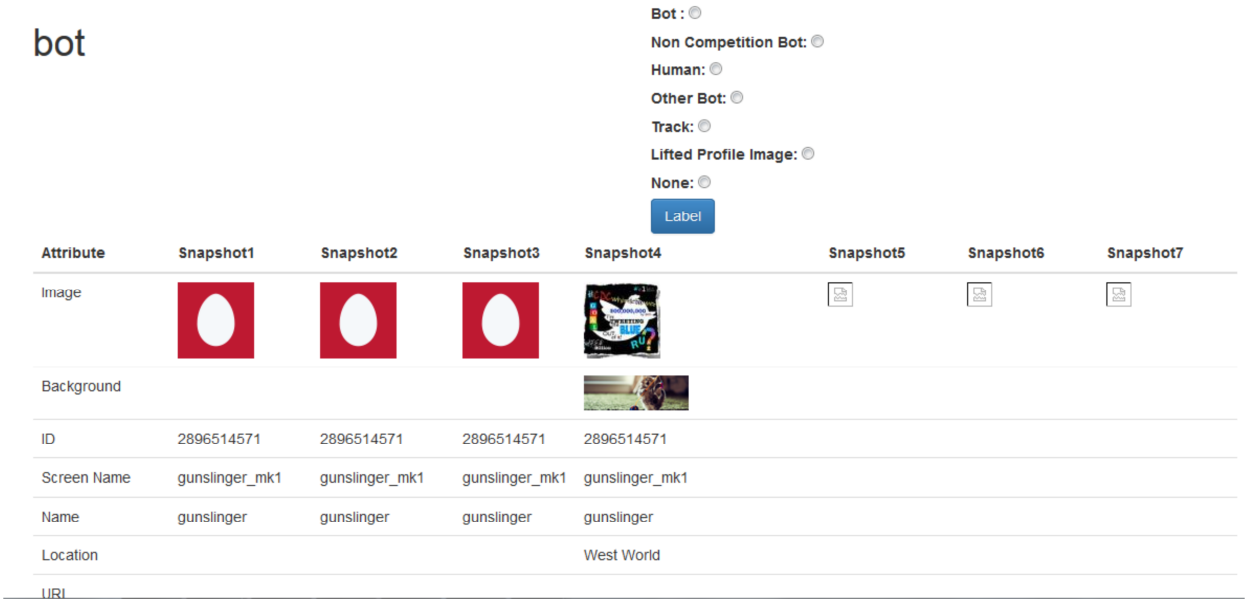

Рис.2. Детальный экран пользователя, являющегося ботом

Команды использовали различные дополнительные алгоритмы определения ботов. Например, были построены так называемые сети смежности хэштегов. Исходя из предоставленного списка хэштегов, связанных с конкретной темой, была построена сеть, узлы которой представляли собой уникальные хэштеги, а ребра имеют вес равный числу упоминаний этих двух хэштегов в твитере.

Эти сети использовались для выявления других хештегов, связанных с другими темами, чтобы обогатить список ключевых слов. Впоследствии сети были использованы, чтобы разделить пользователей на тех, кто занимает одну позицию по обсуждаемому вопрос и остальных пользователей.

Команды предположили, что было бы неэффективно для разработчиков ботов создавать ботов по одному за раз. Скорее всего, одна программа будет генерировать несколько ботов, варьируя один или несколько параметров профиля. Т.к. это приведет к сходству среди ботов, созданных при помощи одной и той же программы, были использованы два метода обнаружения.

Одна из команд сначала применила ортогональную неотрицательную матричную факторизацию (НМФ) к рассмотренным ранее особенностям, чтобы найти векторное представление характеристик каждого пользователя. Затем они использовали группирование на основе кластерного алгоритма обнаружения ботов, чтобы найти значительные отклонения в векторах, заданных в этом пространстве. Команда осуществила кластеризацию на микроуровне с помощью 2-х подходов. В результате

команда не сделала ни одной ошибки первого рода. В результате чего ими была получена превосходная точность.

Другая команда использовала хорошо известный алгоритм DBSCAN для генерации кластеров и затем выявляли которые из пользователей в этих кластерах вероятно являются ботами на основе анализа их особенностей и сходств между их особенностями и особенностями известных ботов.

В результате можно предложить общий алгоритм для обнаружения ботов:

1. Первичное обнаружение ботов.

Первый шаг заключается в определении нескольких ботов. Члены команд использовали четыре широких класса способов, чтобы провести первичное обнаружение ботов - эвристику, анализ поведения, лингвистический анализ, и нестыковки (например, профиль с фото бородатого пожилого человека и ником MaryJones17 - пользователь, который в основном говорит о колледже).

2. Кластеризация и анализ сетевой активности.

Данный шаг базируется на результатах, полученных на первом шаге. Предполагается, что боты взаимодействуют друг с другом, чтобы поднять количество подписчиков и ретвитов. В то время как большинство разработчиков ботов генерируют некоторые их параметры различными, другие их параметры могут быть одинаковыми и образовывать кластеры. Так кластеры, содержащие известных ботов, могут включать в себя других (неизвестных) ботов.

3. Классификация.

На этом шаге требуется классифицировать, к какому типу принадлежит тот или иной найденный бот: спам-бот, бот, который рассылает рекламу, или же бот, целью которого является навязывание определенного мнения.

В виду неспокойной политической обстановки, связанной с множеством локальных конфликтов разных стран, стоит ожидать лишь наращивания ботов, которые стремятся повлиять на умы пользователей социальных сетей. В этом контексте проблема их идентификации и устранения становится особенно острой и злободневной. Без сомнения, актуальность этого вопроса будет возрастать с каждым днем.

Список литературы

[1]. Robert Epstein, Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. URL: http://www.pnas.org/content/112/33/E4512.full.pdf (accessed 17.03.2016).

[2]. Subrahmanian V.S., Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, Rand Waltzman. The DARPA TWITTER BOT CHALLENGE. URL: http://arxiv.org/ftp/arxiv/papers/1601/1601.05140.pdfб, accessed 17.03.2016.

[3]. George Danezis, Prateek Mittal. SybilInfer: Detecting Sybil Nodes using Social Networks. Available at https://www.internetsociety.org/sites/default/files/SybilInfer%20Detecting%20Sybil%20Nodes%20using%20Social%20Networks%20(George%20Danezis).pdf, accessed 17.03.2016).

[4]. Linhong Zhu, Aram Galstyan, James Cheng, Kristina Lerman. Tripartite Graph Clustering for Dynamic Sentiment Analysis on Social Media. Available at: http://arxiv.org/pdf/1402.6010v3.pdf (accessed 17.03.2016)