Zaburzenie procesów demokratycznych przez boty

Śledź, Ścieszka, Rydzewski

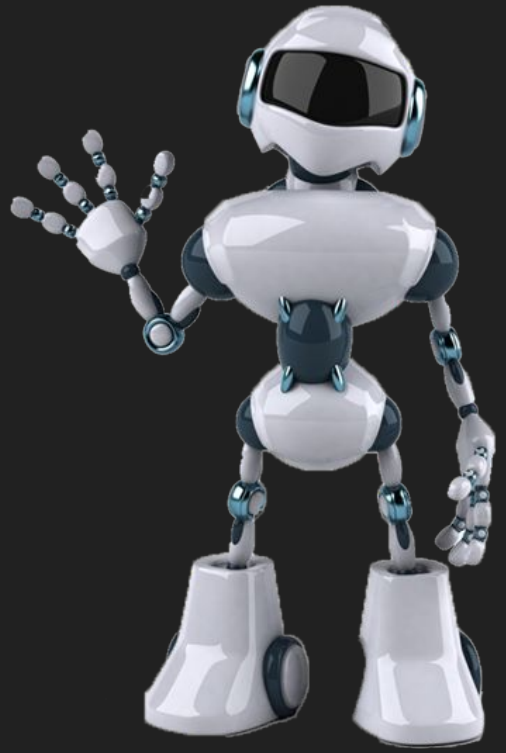
Kraków, grudzień 2022

Wstęp

Wiele różnych organizacji takich jak ISIS, ale także politycy lub media państwowe prowadzą kampanię mające na celu wpływać na opinie użytkowników. Takie działania są niebezpieczne, ponieważ mogą zaburzyć procesy demokratyczne. Aby im przeciwdziałać rośnie zapotrzebowanie na oprogramowania identyfikujące tzw. “influence bots”. Jednym z pomysłów na walkę z botami jest wydarzenie organizowane przez agencję rządową DARPA (Defense Advanced Research Projects Agency) mające na celu spośród wielu drużyn wyłonić oprogramowanie w najlepszy sposób radzące sobie z wykrywaniem influence botów na platformie twitter.

Czym są boty?

Bot – skrót od słowa robot; w świecie technologii informatycznych oznaczający program funkcjonujący w przestrzeni przeznaczonej z założenia dla człowieka i symulujący zachowanie żywego użytkownika. Szczególnym typem botów są zdolne do konwersacji chatboty i voiceboty.



Rodzaje botów na twitterze

- **Spamboty** - rozsiewają spam na dane tematy
- **Payboty** - nielegalnie zarabiają pieniądze poprzez np. kopiowanie linków do stron internetowych i przekierowywanie użytkownika przez strony, za wejścia na które twórca bota dostaje pieniądze
- **Influence boty** - ich zadaniem jest wpłynięcie na decyzję użytkownika poprzez wyrażanie propagandowych poglądów

Do czego są wykorzystywane boty na twitterze

Według badań ponad 8.5% wszystkich użytkowników twittera to boty. Większość z nich służy do spamu, jednak część to influence boty - które są groźniejsze i mogą być niebezpieczne dla wolności wyrażania własnego zdania. Przykładami działania takich botów są radykalne poglądy przekazywane przez ISIS, używane po to, by zachęcać młodzież do ich wspierania lub boty wykorzystywane przez rząd rosyjski, by szerzyć propagandę na temat ich ataku na Ukrainę

Co to jest DARPA?

Defense Advanced Projects Agency Badania (DARPA) jest agencją Departamentu Obrony Stanów Zjednoczonych odpowiedzialną za badania i rozwój nowych technologii dla celów wojskowych. Do dziś DARPA była źródłem rozwoju wielu technologii, które miały poważne konsekwencje na całym świecie, w tym sieci komputerowe (w szczególności ARPANET, który ostatecznie stał się Internetem) i NLS (akronim reprezentujący w języku angielskim wyrażenie „ System ON-Line”, w języku francuskim dosłownie „system online”), który był zarówno pierwszym systemem hipertekstowym, jak i ważnym prekursorem interfejsów graficznych, które stały się dziś wszechobecne.

Twitter bot detection challenge

Stworzona przez DARPA czterotygodniowa konkurencja testująca metody utworzone podczas programu SMISC (SocialMedia in Strategic Communications) wykrywania influence botów. Zadaniem uczestników było wykrycie botów szerzących pro-szczepionkowe poglądy.

Główne zadania uczestników:

- oddzielić influence boty od innych typów botów
- oddzielić influnece boty wypowiadające się na dany temat od innych botów
- oddzielić influnece boty szerzące pro-szczepionkowe poglądy od tych wypowiadających się negatywnie oraz neutralnie na ich temat

Uczestnicy wyzwania oraz wyniki

W konkurencji brało udział 6 drużyn: University of Southern California, Indiana University, Georgia Tech, Sentimetrix, IBM oraz Boston Fusion.

Drużyny nie znając dokładnej liczby miały wykryć 39 pro-szczepionkowych botów.

Wyzwanie wygrał Sentimetrix, który pobił inne drużyny o 6 dni z dokładnością 39/40 trafień. Uniwersytet Południowej Kalifornii (University of Southern California) osiągnął najwyższą dokładność 39/39 trafień.

Tabela wyników oraz sposób ich obliczania

	Misses	Hits	Guesses	Accuracy	Speed	Final Score
Sentimetrix	1	39	40	38.75	12	50.75
USC	0	39	39	39	6	45
DESPIC	7	39	46	37.25	6	43.25
IBM	4	39	43	38	5	43
B. Fusion	9	39	48	36.75	5	41.75
G. Tech	56	38	94	24	0	24

Kolumna “*Accuracy*” - precyzja, przedstawia wartość (*h-0.25m*), gdzie *h* to ilość trafień, *m* to ilość pomyłek. Kolumna “*Speed*” określa ilość pozostałych dni do zakończenia wyzwania w momencie wykrycia przez drużynę wszystkich botów.

Finałowy rezultat był obliczany w sposób:
*FinalScore = Hits - 0.25*Misses + Speed*

Przygotowanie środowiska do wyzwania

Do wyzwania przygotowano syntetyczne środowisko emulujące API Twittera. Zestaw danych składał się z:

- 7038 kont użytkowników
- profile użytkowników
  - zredagowano profile użytkowników w formacie użytkowników Twittera (zdjęcie profilowe, URL, liczba znajomych i obserwujących i krótkie bio)
- 4095083 tweetów ze znacznikami czasowymi (timestamp)

Po rozpoczęciu wyzwania drużyny miały możliwość przesyłania odpowiedzi, które serwer sprawdzał i oznaczał odpowiedź jako prawidłowa/nieprawidłowa

Podejście drużyn do znajdowania botów

Każdy z trzech zwycięskich zespołów stwierdził, że same techniki uczenia maszynowego byłyby niewystarczające z powodu braku danych szkoleniowych, Jednak półautomatyczny proces, który obejmował uczenie maszynowe, okazał się przydatny.

Żaden zespół nie rozpoczął od uczenia nienadzorowanego (unsupervised). USC wykorzystał nienadzorowane wykrywanie wartości odstających w połączeniu z innymi dowodami. Żaden zespół nie uznał za przydatne istniejących metod Sybil detection. Każdy zespół skorzystał z wcześniejszych badań nad influence-botami.

Stworzenie zbioru trenującego

Prawie każdy z zespołów wykorzystał wcześniejszą pracę do zbudowania profilu użytkownika. Sentimetrixowy SentiBot uczył się identyfikować influence-boty na podstawie zbioru zawierającego ponad 17 milionów użytkowników i 25 milionów tweetów.

Tweet syntax

Kilka elementów tweeta, na które zwrócono uwagę:

- czy składnia przypomina tę, którą posługują się programy do generacji tekstu
- średnia liczba hashtagów, oznaczeń innych użytkowników, ilość znaków specjalnych
- średnia liczba retweetów
- informacja, czy tweety są geo-enabled
- procent tweetów, które kończą się kropką, hashtagiem, lub linkiem

Semantyka tweetów

Kilka elementów, na które zwracano uwagę:

- liczba tweetów użytkownika powiązanych z tematem szczepień
- średni wynik sentymentu (związany z tematem szczepień)
- najczęstszy temat tweetów danego użytkownika
- liczba języków, w którym został wygenerowany tweet
- niespójność sentymentu

Zmiana zachowania użytkownika w czasie

Kategoria ta jest istotna, ponieważ taktyka, którą mogą obierać influence-boty, to np. skupienie uwagi, oraz zaangażowanie grupy antyszczepionkowej, aby następnie zmienić pogląd w celu zmiany poglądów użytkowników, których uwagę udało się pozyskać.

Przykładowe badane cechy:

- zmiana sentymentów,
- liczba tweetów na dzień,
- czas trwania sesji (sesja trwająca np. dzień jest podejrzliwa),
- procent straconych obserwujących

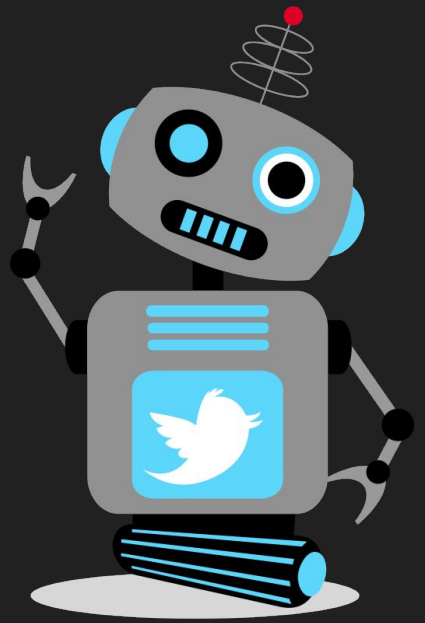
Cechy profilu użytkownika

Przykładowe cechy brane pod uwagę przy analizie profilu użytkownika:

- czy użytkownik ustawił zdjęcie profilowe, jeśli tak, to czy jest ono w bazie stoków zdjęciowych
- czy nazwa użytkownika przypomina nazwy generowane przez programy do generowania nazw
- liczba tweetów, retweetów, odpowiedzi, oznaczeń
- liczba obserwujących i obserwowanych
- dostępność koordynatów gps
- liczba urządzeń (aplikacja mobilna, przeglądarka internetowa)

Metody identyfikowania botów

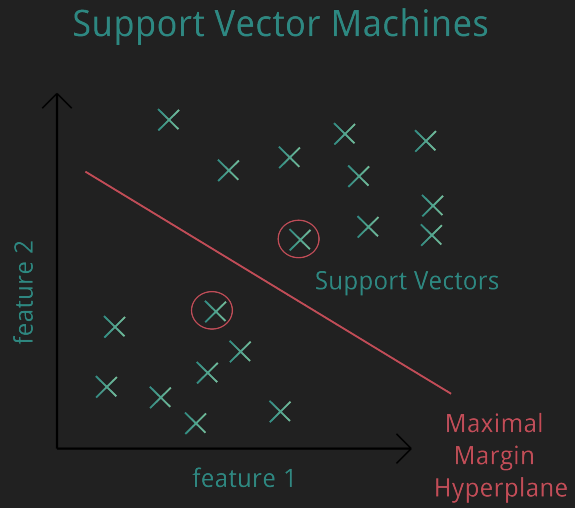
Każdy zespół wykorzystał wiedzę z pracy wykonanej w przeszłości do manualnego znalezienia początkowego zbioru botów.



Metody identyfikowania botów

Sentimetrix wykorzystał:

- grupowanie
- analizę sieci (network analysis) do zidentyfikowania dodatkowej liczby botów
- SVM (Support Vector Machine) do przewidywania pozostałej liczby botów



Metody identyfikowania botów

Uniwersytet Południowej Kaliforni

- wykrywanie wartości odstających (outlier detection) za pomocą analizy treści i kontroli manualnej - wykrywanie początkowych botów
- analiza sieciowa (network analysis), analiza treści i sentymentu, oraz pół-nadzorowane grupowanie (semi-supervised clustering) - wykrywanie dodatkowych botów

USC
UNIVERSITY
OF SOUTHERN
CALIFORNIA

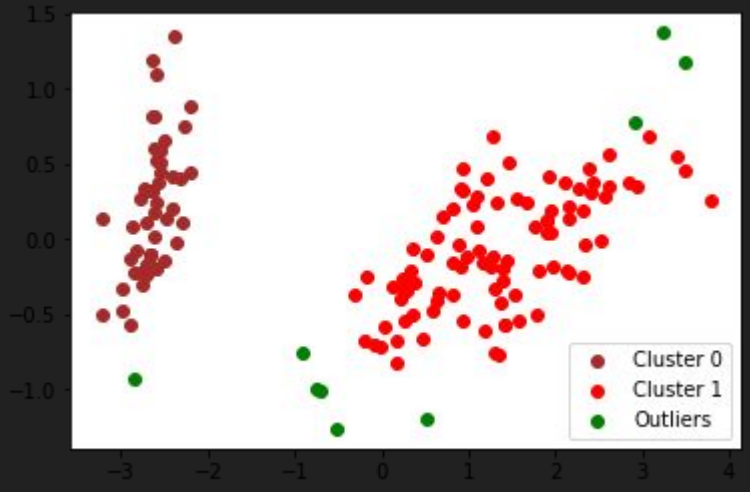
Algorytmy analizy botów stosowane w wyzwaniu

Zespół z Indiany do identyfikacji botów wykorzystał:

- sieci współwystępowania hashtagów (hashtag co-occurrence networks),
- miary odległości
  - podobieństwo cosinusowe
  - odległość Jaccarda (Jaccard distance)
- strategię predykcji online opartą na wieloramiennym bandycie (multi-arm bandit)

Algorytmy analizy botów stosowane w wyzwaniu

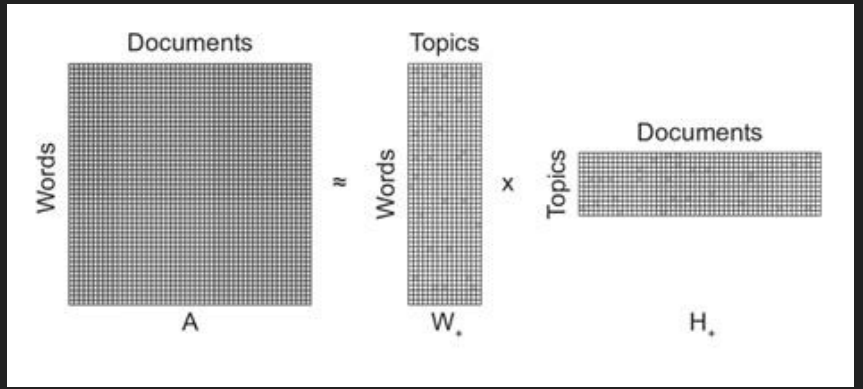
Uniwersytet Południowej Kaliforni (USC) i **Sentimetrix** wykorzystały techniki wykrywania wartości odstających (outlier detection)



Algorytmy analizy botów stosowane w wyzwaniu

USC aby znaleźć wartości odstające w niskowymiarowej (low-dimensional) przestrzeni latentnej (latent space) zastosował:

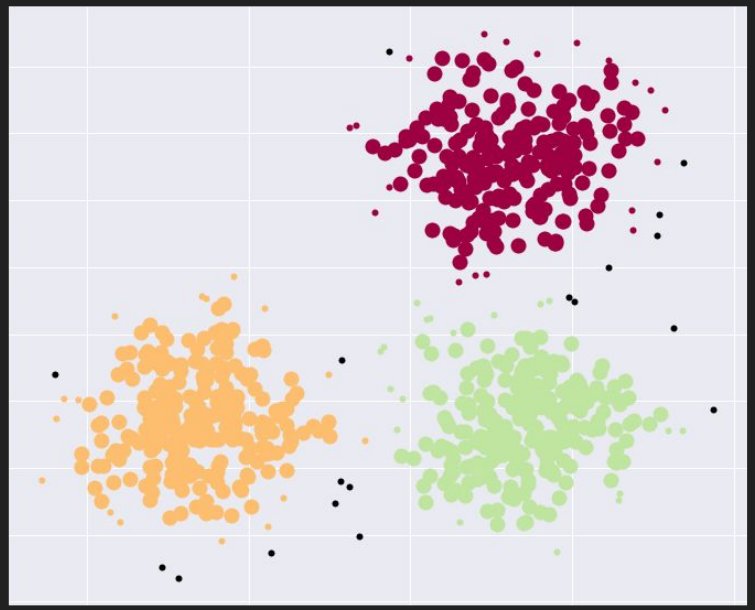
- ortogonalną faktoryzację macierzy nieujemnej (*Non-Negative Matrix Factorization - NMF*)
- wykrywanie wartości odstających oparte na klastrowaniu (clustering-based outlier detection),



Algorytmy analizy botów stosowane w wyzwaniu

Sentimetrix w celu identyfikacji prawdopodobnych botów (likely bots):

- wykorzystał algorytm *DBSCAN* do generowania klastrów
- analizował cechy użytkownika (user features)
- analizował podobieństwa do znanych botów

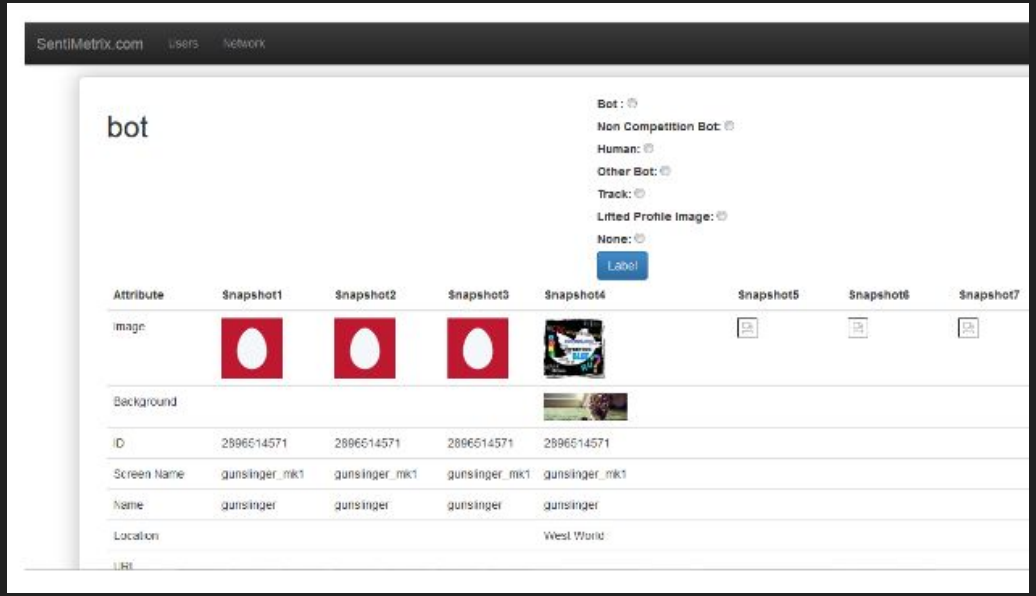


Algorytmy analizy botów stosowane w wyzwaniu

Podejście zespołu z Indiany do oceny botów oraz technika USC do wykrywania wartości odstających osiągnęły w konkursie doskonałe wyniki dokładności

Narzędzia wykorzystane do identyfikacji botów

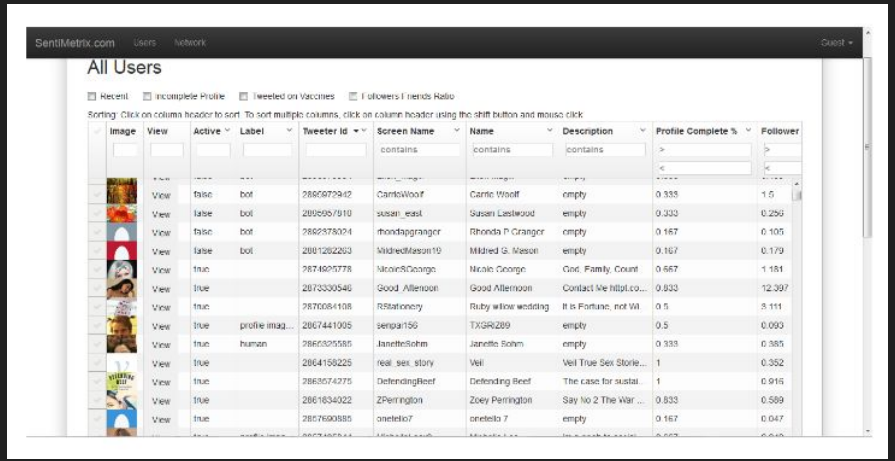
Zespoły zastosowały panele informacyjne (dahsboards) do listowania użytkowników w celu identyfikowania podejrzanych kont



Narzędzia wykorzystane do identyfikacji botów

Panel **Sentimetrixa** zawierał szczegółowe informacje dotyczące każdego konta użytkownika:

- kompletność profilu,
- opisy i dodatkowe zmienne, które mogą sugerować podejrzaną aktywność



Narzędzia wykorzystane do identyfikacji botów

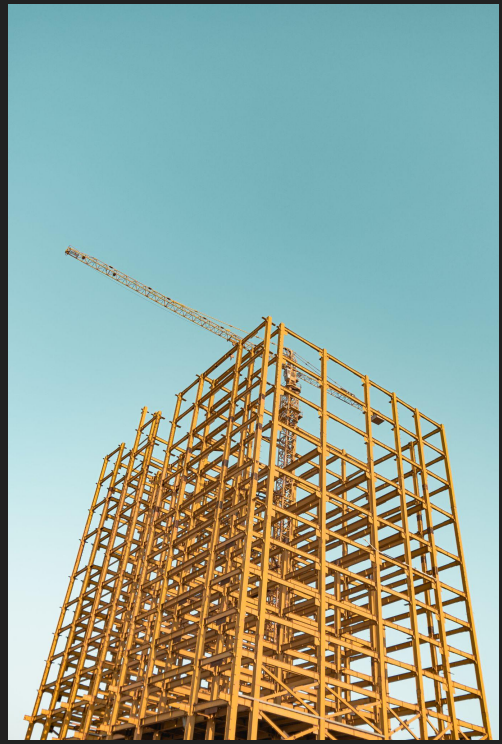
Zastosowanie panelu pozwoliło analitykom na:

- przeszukiwanie profili,
- sortowanie profili,
- przeglądanie zrzutów sieci,
- przeglądanie szczegółów dotyczących użytkowników.

Stworzenie frameworku do wykrywania influence botów

Wykrywanie opiera się na półautomatycznym procesie, który wykorzystuje cztery techniki:

- wykrywanie niespójności (inconsistency detection) i modelowanie zachowań (behavioral modeling),
- analiza tekstu (text analysis),
- analiza sieci (network analysis)
- uczenie maszynowe (machine learning)

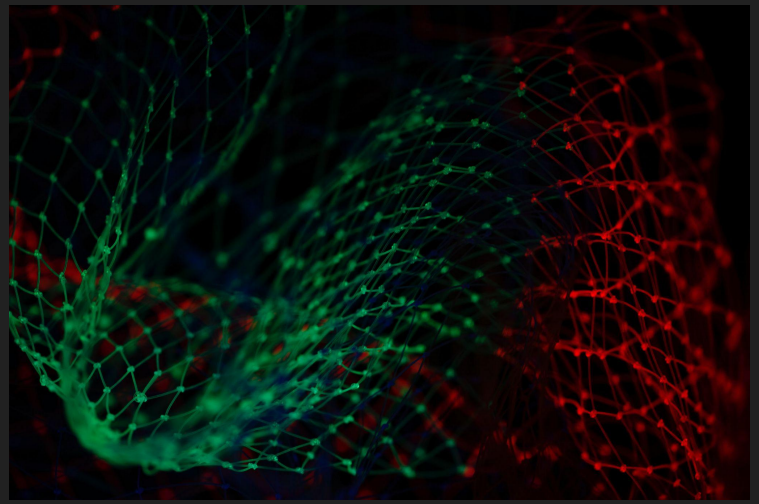


Stworzenie frameworku do wykrywania wpływu botów

Krok 1 - wstępne wykrywanie botów
Initial Bot Detection

Identyfikacja kilku botów przy użyciu:

- heurystyki (heuristics),
- zachowań (behaviors),
- wskazówek językowych (linguistic cues),
- niespójności (inconsistencies).



Stworzenie frameworku do wykrywania wpływu botów

Krok 2 - grupowanie, wartości odstające i analiza sieci (Clustering, Outliers, and Network Analysis)

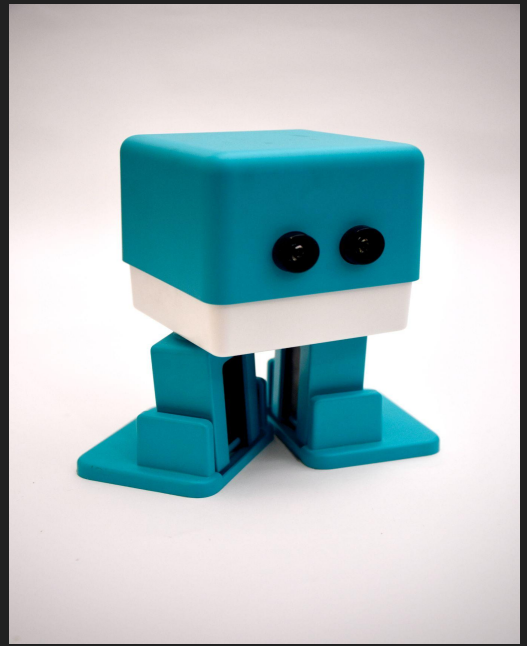
- wykorzystanie zidentyfikowanych botów do znalezienia skupisk innych botów
- wykorzystanie analizy wartości odstających do zidentyfikowania dodatkowych botów

Stworzenie frameworku do wykrywania wpływu botów

Krok 3 - klasyfikacja/analiza wartości odstających (Classification/Outlier Analysis)

Identyfikacja dodatkowych botów po zidentyfikowaniu określonej liczby botów i ludzi poprzez:

- wykorzystanie standardowych klasyfikatorów
- analiza wartości odstających



Stworzenie frameworku do wykrywania wpływu botów

- system musi być częściowo nadzorowany, z ludzką oceną
- wykorzystanie interfejsów:
  - odpowiadają na pytanie dlaczego dane konto jest uważane za bota
  - dają informacje zwrotne (feedback) w celu poprawy dokładności

Wnioski

- Twórcy botów stają się coraz bardziej wyrafinowani.
- Możemy spodziewać się rozprzestrzenienia się wpływu botów na media społecznościowe.
- Potrzeba ulepszenia narzędzi analitycznych
- Opisane metody mogą być używane do wykrywania zarówno całkowicie zautomatyzowanych botów, jak i zarządzanych przez ludzi (human-orchestrated) operacji wpływu (influence operations).

Możliwość rozwoju na przyszłość

- Automatyzacja procesu wstępnego wykrywania botów.
- Opracowanie zestawu narzędzi dla etapu "Klasteryzacja, odstające elementy i analiza sieci"
- Użycie lepszych metod wizualizacji.
- Użycie tradycyjnych klasyfikatorów, w celu wygenerowania dodatkowych kandydatów botów po odkryciu wystarczającej liczby botów i prawdziwych kont (benign accounts).

Literatura

- https://arxiv.org/ftp/arxiv/papers/1601/1601.05140.pdf
- https://www.k2bots.ai/abotcadlo/bot
- https://pl.frwiki.wiki/wiki/Defense_Advanced_Research_Projects_Agency
- https://www.bbc.com/news/technology-60790821
- https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00338-3