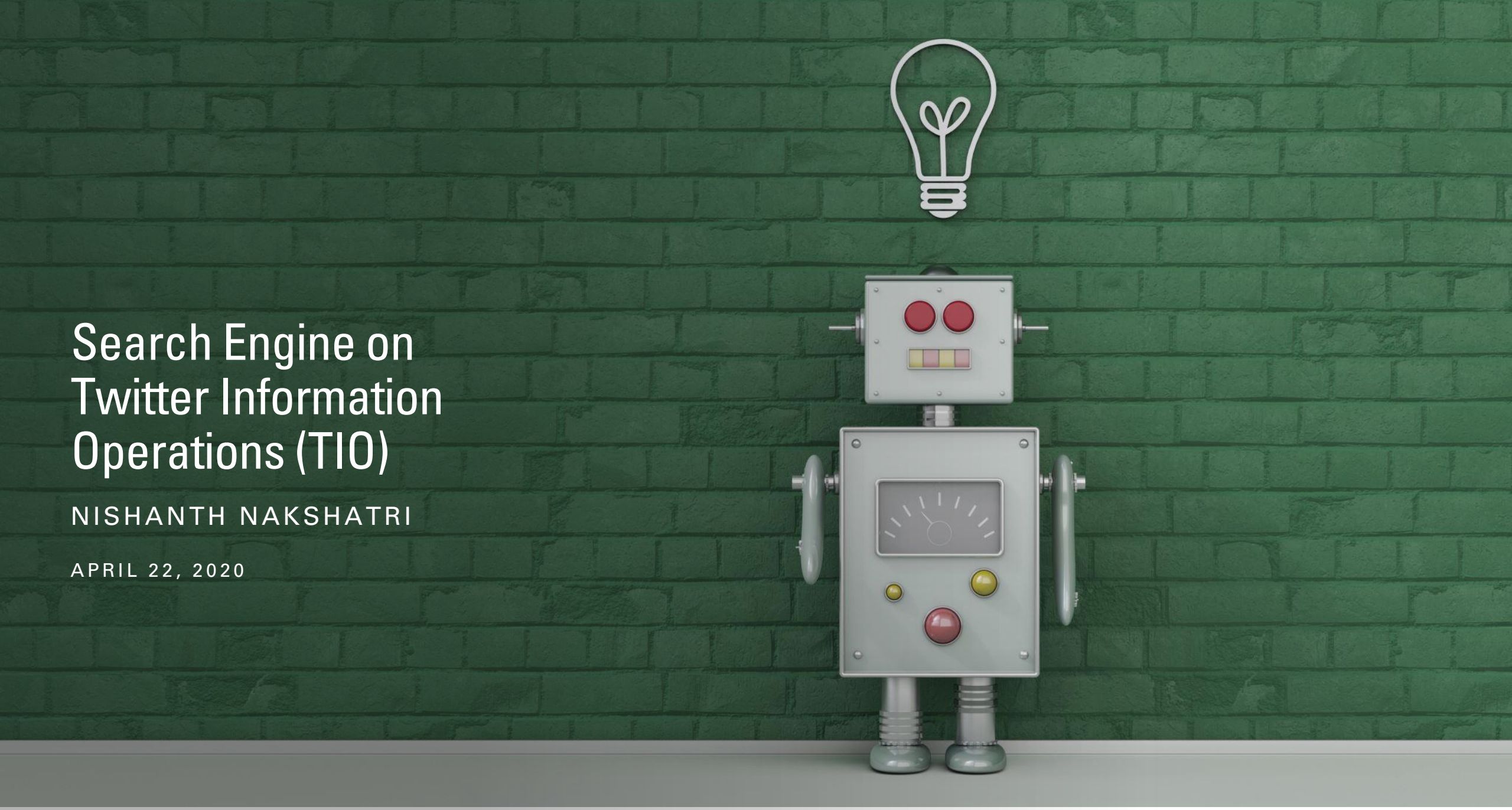
Search Engine on Twitter Information Operations (TIO)

NISHANTH NAKSHATRI

APRIL 22, 2020



Overview

- Customers
- SRS – Search Engine Requirement Specifications
- About the Dataset
- Research Interests
- Statistics about the dataset used for the project
- Workflow
- ElasticSearch APIs Explored
- Search Engine Comparison and Evaluation
- Demo
- Future Work

Customers

1. Christopher Griffin, Associate Research Professor at the Applied Research Laboratory (ARL) at Penn State.
2. Sarah Rajtmajer, Assistant Professor in the College of Information Sciences and Technology at Penn State.

Search Engine Requirement Specifications

- Language agnostic search capability on the Twitter Information Operations Dataset
- Ability to identify Tweet Bursts over a specified timeframe
- Basic User Interface
- Ingest all the data pertaining specific countries of interest

About the Dataset

- Tweets and media associated with known state-backed information operations on Twitter
- Entire Dataset can be found here
- Countries of interest: **Russia**, **Iran** & **Venezuela**
- More information about the dataset can be found here
- ** All accounts have been removed from Twitter
- ** Twitter Attribution is not 100% accurate
  - Relying primarily on researchers, government and other law enforcement agencies to inform about twitter attribution efforts.

Research Interests

- Identify attempted influence campaigns
- Investigation into foreign interference in political conversations on Twitter
  - Identify coordinated international campaigns
- Encourage open research and investigation of these behaviors from researchers around the globe
- Relevant Publications:
  - Rajtmajer, Sarah, et al. "A Dynamical Systems Perspective Reveals Coordination in Russian Twitter Operations." arXiv preprint arXiv:2001.08816 (2020)
  - Griffin, Christopher, and Brady Bickel. "Unsupervised Machine Learning of Open Source Russian Twitter Data Reveals Global Scope and Operational Characteristics." arXiv preprint arXiv:1810.01466 (2018)

Statistics about the dataset used

- 7,539 User Accounts
- 19,510,805 million tweets (full-tweets)
- 70 languages
- ~11.5 gb of data

Workflow

Data Pre-processing	Data Ingestion & Indexing	Build APIs for custom search & insights	UI Integration
• Data type issues • Null values	• Python's elasticsearch API • Index on the entire dataset • Built-in tokenizer for many languages	• Search on tweet-text • Search based on userid • Search for tweet bursts based on a threshold • Search for geo-based twitter activity	• Simple UI to search • Communicate results in an effective manner

ElasticSearch APIs Explored

- Search API
  - Search
  - Multi-Search API
  - Count
- Bucket Aggregations
- Terms Aggregations
- SQL API

Comparison & Evaluation Component

- First of its kind – a niche search engine!
- No other search engine based on information operations released by twitter
- **Google Custom Search Engine?**
  - Not the right fit for the problem
  - Google Custom search engine retrieves information from current Twitter network
  - Information operations dataset comprises of users banned from Twitter

Search Engine Live Demo

Link to Search Engine

Future Work

- Extend this work on other countries
- Include media and try indexing the image content
- Make this search engine publicly available for researchers