



Text as a Network: Analysis of COVID-19 related Tweets

J.D. Moffitt

jdmoffit@cs.cmu.edu

CASOS Center, Institute for Software Research
Carnegie Mellon University

CASOS Summer Institute 2020



Carnegie Mellon

Center for Computational Analysis of
Social and Organizational Systems
<http://www.casos.cs.cmu.edu/>



Agenda

- Objectives
- Case Study Background & Data
- Text as a Network Refresher
- Hands on with NetMapper & ORA for Text analysis
- Reference Slides



June 2020

2



Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Objectives of this case study


- In the context of the COVID-19 pandemic:
 - How can we use Dynamic Network Analysis tools to examine the Twitter conversation around COVID-19 as a bioweapon?
 - How can we discover emerging topics, individuals, groups, or organizations through twitter discourse?

CASOS
June 2020 3

Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Known COVID19 Mis-/Dis-Information Campaigns

1. Stories relating inaccurate information about cures or preventative measures
2. Stories relating inaccurate information about the nature of the virus
3. Stories relating inaccurate information that are conspiracy stories



CASOS
June 2020 4



Carnegie Mellon
Institute for SOFTWARE RESEARCH

Data: COVID19 Related Tweets (Bioweapon)

Raw Data:

- Tweets collected from global Twitter stream based on keywords:

NCoV2019 covid-19 covid19 coronavirus NCoV
coronavirus covid 19 wuhanvirus 2019nCoV wuhan virus

- Using regular expressions, further filtered tweets for only those containing the word bioweapon

bioweapon bioweapons Lab
bat bio-weapon 5G

- Resulting in:
 - ~97,000 tweets from 16-29 February 2020
 - ~200,000 tweets from 01-31 March 2020

Data Processing:

Parsed tweet objects from 150 to 11 attributes → Filtered for tweets in English → Conducted feature engineering for network/key entity Analysis → Extracted tweet text for content analysis

CASOS
June 2020

5

Carnegie Mellon
Institute for SOFTWARE RESEARCH

Method / Analysis

Network Construction:

- Create edge lists from tweets
- Re-tweet Network
- Reply Network
- Mention Network

Key Entity Analysis:

- Identify Who is important
 - Dynamically → Changes over time
 - Statically → In a given period
- Identify/Analyze what the important entities are saying

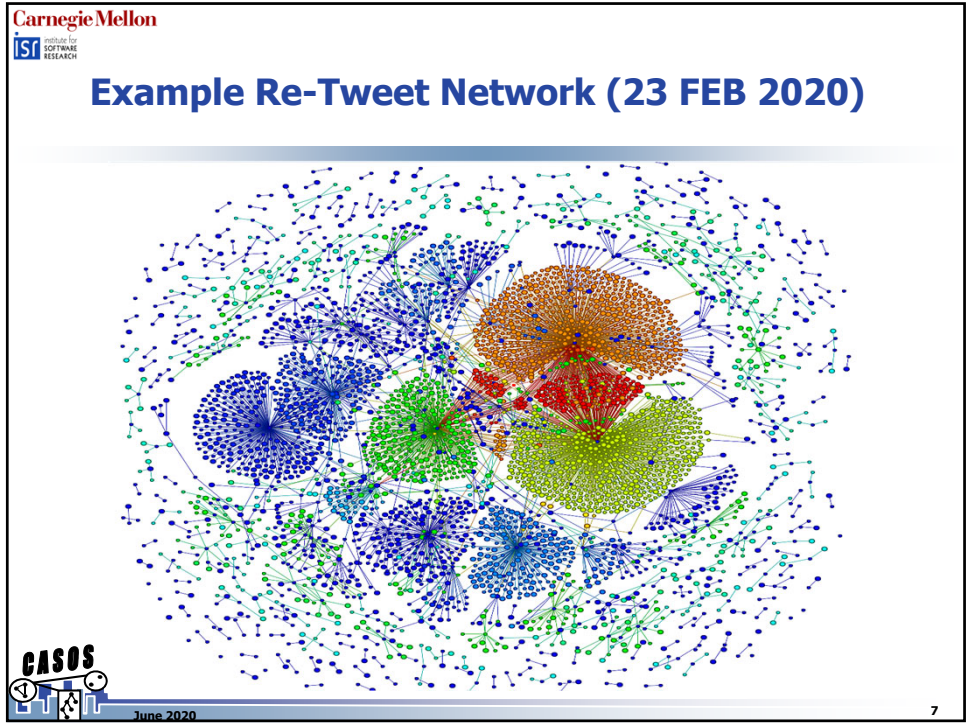
Data-frame	Predicted Bot %	Predicted Not bot %
FEB-Bioweapon	36%	64%
FEB-Covid19	39%	61%
FEB-5G	34%	66%
MAR-Bioweapon	33%	67%
MAR-Covid19	32%	69%
MAR-5G	30%	70%

Network Metric Analysis:

- Nodes/links/density/etc.
- Composition of nodes (who is in convo)
 - Bots / Countries / Agent type

CASOS
June 2020





Carnegie Mellon
ISI Institute for SOFTWARE RESEARCH

Example Re-tweet Key Entity Text (FEB 2020)

February 2020 Bio-weapon Re-tweet Key Entities (Change in Key Entities Report)

Screen Name	Who Are they (Profile + Internet Search)	What they Said	Analysis
[REDACTED]	English Youtube personality, considered far-right or alt-right, conspiracy theorist	CNN/NY Times journo says coronavirus might have emerged from lab in Wuhan after someone was bitten by a bat during a test. For the past month, the media has been defaming anyone who suggested the same as a "dangerous" fake news conspiracy theorist.	User's post was re-tweeted 1,608 times on 18FEB20. This appears to be an attempt to highlight to hypocrisy of information spread.
[REDACTED]	Suspended account	They cannot contain the truth that #CoronavirusOutbreak originated in a lab, so their excuse is that a bat peed on a scientist who didn't wash their hands. Dumb. Why does a deadly disease have a PATENT IN THE FIRST PLACE? #GatesFoundationhttps://t.co/bvobqe51i	This user's account is suspended. I was able to piece together some information because user's continue to mention this user on Twitter. It appears this user is of a QAnon flavor. Users continue to use Stormisupon us hashtags and there are co-uses of Obamagate hashtags. This post had 6,097 retweets on 17 FEB 20.
[REDACTED]	Daily newspaper in new york city; tabloid-ish	Don't buy China's story: The coronavirus may have leaked from a lab https://t.co/macW75ame https://t.co/mVMX0x09W	This is a media account. Link directs users to an article on its website. The post received 812 retweets on 20FEB20.
[REDACTED]	Digital Journalist, CNNi, Covering Asia, formerly SCMPNews.	This is beyond shocking China shut down the lab that published the world's first genome sequence of the #coronavirus last month, barring its scientists from finding ways to contain the outbreak Their only crime? Publishing the sequence before authorities	This user appears to be a reporter for CNN International. They have a verified check. The tweet links to an article to South China Morning Post. The tweet was re-tweeted 3,662 on 29FEB20.
[REDACTED]	account suspended	Don't forget Epstein didn't kill himself...And don't forget the Coronavirus was created in a lab in Wuhan. And now it's being used as a bio-weapon to remove / murder the Chinese protestors, create a demand for a new vaccine (\$), & manufacture public fear to hurt the economy.	This users account is suspended. The user appears to be affiliated with Qanon. The tweet was re-tweeted 2,400 times on 29FEB20.

CASOS

June 2020

8



Carnegie Mellon Institute for Software Research

Why Text?

- Text is a cheap easy way to store large volumes of information
 - Books
 - Documents (legal, annual reports, transcripts, mission statements)
 - News
 - Blogs
 - Social Media
- Information can be extracted from Text:
 - Content Analysis (word counts, parts of speech, concepts)
 - Key Entity Analysis (Find people, Organizations, Locations)
 - Topic Analysis (#'s, hot topics, themes, groups of topics)
 - Semantic Network Analysis (mental models of text usage)
 - Meta-Network Analysis
 - Sentiment Analysis

CASOS June 2020 9

Carnegie Mellon Institute for Software Research

Text in Network Terms

- Nodes
 - Concepts
 - Words
 - Phrases
- Link / Edges
 - Link between two+ concepts
 - i.e. a statement
- Network
 - Union of all statements in a text
 - A Map
- Meta-network
 - Map + Taxonomy

Legend:
○ → Agent
□ → Organization
⬡ → Task
△ → Resource
⬡ → Location

CASOS June 2020 10



Carnegie Mellon Institute for SOFTWARE RESEARCH

Semantic Network vs Meta-Network

- **Semantic Network:**
 - One mode network (concepts & connections)
 - Cognitive / Mental Model that can:
 1. Represent the author's reality
 2. Represent the author's knowledge & Information on a topic
- **Meta-Network:**
 - Cross-classify nodes in semantic network into categories
 - Requires Mapping of Words to Categories (explicit or algorithms)
 - Allows Analyst to:
 1. Who is linked to orgs, resources, tasks
 2. What resources or knowledge are needed for what task
 3. Agent characteristics
 4. Types of orgs, locations, etc.

CASOS June 2020 11

Carnegie Mellon Institute for SOFTWARE RESEARCH

Turning Text into Networks

Tip: Analyst can refine thesauri and delete lists after observing NetMapper outputs and reprocess text with new inputs

- **Preprocess** (Choose your favorite tool)
 - Text
 - Source
 - Reduction
 - Normalization
 - Links
 - Domain / subject expertise
 - Develop initial scheme for how concepts are linked
 - Can adjust pre- & post-processing
- **NetMapper**
 - Thesauri
 - Link relevant concepts
 - Ontology cross-classification
 - Reduce noise by combining common spellings, mis-spellings
 - Built-in or User-defined
 - Delete Lists
 - Remove words that do not contribute to analysis
 - Built-in or User-defined
- **ORA**
 - Analysis
 - Attribute Addition
 - Geo-location
 - Membership and belief inference


CASOS June 2020 12



Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

Hands on Exercise

Getting stuck after skipping the tutorial



Can you repeat the part of the stuff where you said all about the things?

every time i skip the tutorial

CASOS

June 2020

13

Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

Hands on Exercise

1. Process raw .txt file in NetMapper
2. Refine Thesaurus and Delete Lists
3. Create Semantic and Meta-Networks by day for tweets from (14-29 FEB 2020)
4. Load Networks into ORA for Analysis
5. Refined Thesaurus and Delete Lists in ORA
6. Explore ORA Reports that Aide in Text Analysis

CASOS


June 2020

14



Carnegie Mellon
Institute for SOFTWARE RESEARCH

Reference Slides



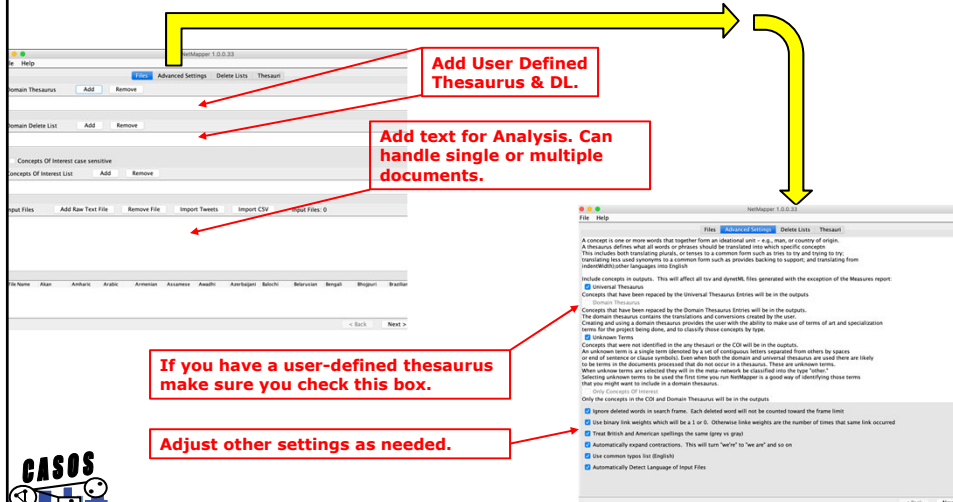
CASOS

June 2020

15

Carnegie Mellon
Institute for SOFTWARE RESEARCH

Reference Slide: NetMapper



NetMapper 1.0.0.33

File Help

File: [Home/Concepts](#) [Delete Lists...](#) [Thesaur...](#)

A concept is one or more words that together form an abstract idea - i.e. topic, or portion of reality. A thesaurus defines what all words or phrases should be translated into which specific concepts. This includes both translating directly, or creating the concept from such as time by and trying to try, translating the word systems to a common form such as provides backing to support, and translating from indistinguishable languages into English.

Include concepts in outputs. This will affect all txt and dynamic files generated with the exception of the Measures report.

Universal Thesaurus

Concepts that have been rejected by the Universal Thesaurus Entries will be in the outputs.

Concepts that have been rejected by the Domain Thesaurus Entries will be in the outputs.

The domain Thesaurus contains the translations and concepts created by the user.

Adding and using a custom thesaurus provides the user with the ability to make use of terms of art and specialization terms for the project being done, and to classify these concepts by type.

Unknown Terms

Concepts that were not identified in the any thesaur or the COI will be in the outputs.

An entry item in a single entry identified by a set of concepts with an universal thesaurus are used there are likely other entries items are created they will be the main network for identified into the type. Some missing unknown terms to be used the first time you run NetMapper is a good way of identifying these terms that are might want to include in a Domain Thesaurus.

Only the concepts in the COI and Domain Thesaurus will be in the outputs.

Only the concepts in the COI and Domain Thesaurus will be in the outputs.

Ignore defined words in search frame. Each defined word will not be counted toward the frame limit.

Use binary link weights which will be a 1 or 0. Otherwise link weights are the number of times that same link occurred.

Use Dutch and American spellings the same (yes or no)

Automatically expand contractions. This will turn "we're" to "we an" and so on.

Use common types list (English)

Automatically Detect Language of Input Files

NetMapper 1.0.0.33

File Help

Back Next

CASOS

June 2020

16



Reference Slide: NetMapper

Search Window Type: Word

Search Window Width: [Slider]

Sentiment Window Width: 3

Choices made here depend on type and size of document. For larger documents it may be prudent to search by sentence, and for smaller text by word. Analysts should experiment/refine to find best settings for their text.

- Search Window Type: Sentence vs Word
- Search Window Width: 1 to N
- Sentiment Window Width: 1 to N

Concept	Frequency	Mean Sentiment	Uncertainty
laboratory	6372	-0.086003	0.00367604
coronavirus	6307	-0.062443	0.00360607
Peoples_Republic_of_China	25032	-0.330619	0.00588433
Wuhan	23764	0.00702696	0.00652507
bat	23726	-0.150418	0.00621376
bioweapon	18854	-0.193256	0.00644762
originate	14658	-0.14395	0.00782274
chinese	12398	-0.0630956	0.00811939
etc	12142	-0.0468007	0.008154
zonal	9784	-0.0555802	0.00975029

laboratory	location	hour	event
laboratory	unknown	denial	task
analogue	resource	mem	belief
Peoples_Republic_of_China	location	laboratory	location
carry	task	ingredient	resource
denial	task	coronavirus	unknown
ita	unknown	Wuhan	location
bat	organization	bat	task
bioweapons	resource	belief	belief
animal	resource	sale	task
tule	task	coronavirus	unknown

concept count	reading difficulty	named entity	abusive
306327	0.109863	1898	39
exclusive	poweranger	powerencourage	powerfear
1	1	2	3
powerforbidden	powergreed	powerlust	powersafety
2	1	1	1
absolutist	equivocal	connective	positive
2	1	1	702
negative	1st person	2nd person	3rd person
698	8		
all caps	avg sentence length	# sentences	avg word length
5795	134.278	1060	7.40512

17

Reference Slide: ORA Edit Nodes

To Delete or Merge Nodes:

1. Select node(s) of interest
2. Right Click
3. Choose Appropriate Action

- Delete selected nodes
- Keep only selected nodes
- Merge selected nodes
- Move selected nodes
- Show all nodes
- Show only selected nodes
- Hide selected nodes

18



Carnegie Mellon
ISRI Institute for SOFTWARE RESEARCH

Reference Slide: ORA Reports Used

Semantic Network Report

Topic Analysis Report

Change in Key Entities Report

June 2020 19

