NATIONAL UNIVERSITY

Computing Society of the Philippines
Special Interest Group - Natural Language Processing (SIG-NLP)

A Primer to Natural Language Processing Research

Joseph Marvin Imperial

Department of Computer Science
NU Human Language Technology Lab
National University

Background

Faculty Member
Department of Computer Science, National University

Lab Coordinator
NU Human Language Technology Lab

MS Computer Science at De La Salle University
BS Computer Science at Bicol University

My research interests are in NLP for education, specifically on **readability assessment and text complexity** of children's books. I have also worked with NLP for understanding social media texts.

www.josephimperial.com

Goals

NATIONAL UNIVERSITY

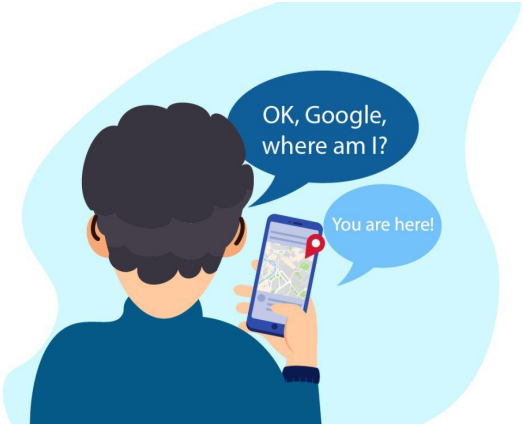
1. To understand the basics and purpose of NLP.
2. To be informed of the interesting research problems and opportunities in NLP research using PH languages.

Natural Language Processing

Natural Language Processing (NLP) is a subfield of computer science concerned with the **interactions** between computers and human (natural) languages.

We want enable computers to **process human language** in the form of text or voice data and to understand its full meaning, complete with the speaker or writer's intent and sentiment.

Computers + language = magic?



Natural Language Processing

"We want to build models that enable people to **read news** that was not written in their language, or **ask questions** about their health when they don't have access to a doctor, and many more."

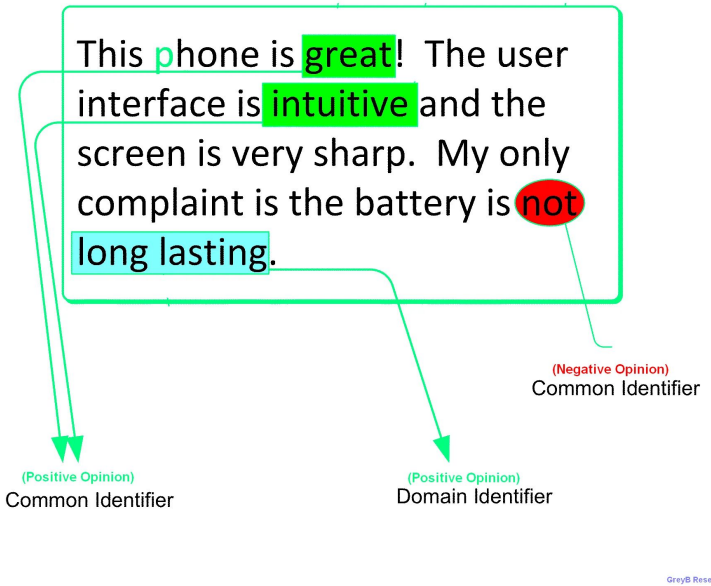
- **Sebastian Ruder**, NLP Researcher at DeepMind

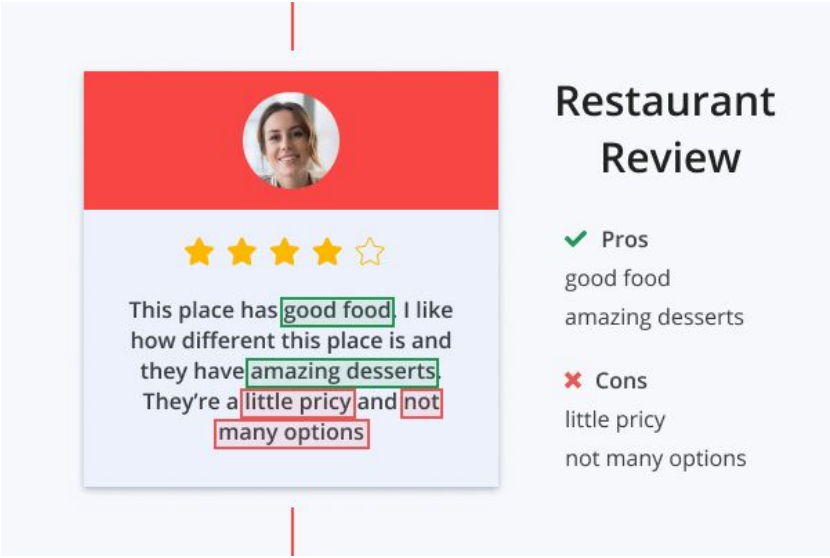
NLP Applications

NATIONAL UNIVERSITY

1. Sentiment analysis of product reviews.
2. Question answering / chatbot systems for customer help
3. Machine translation for travelers
4. Topic modelling for thematic extraction
5. ...

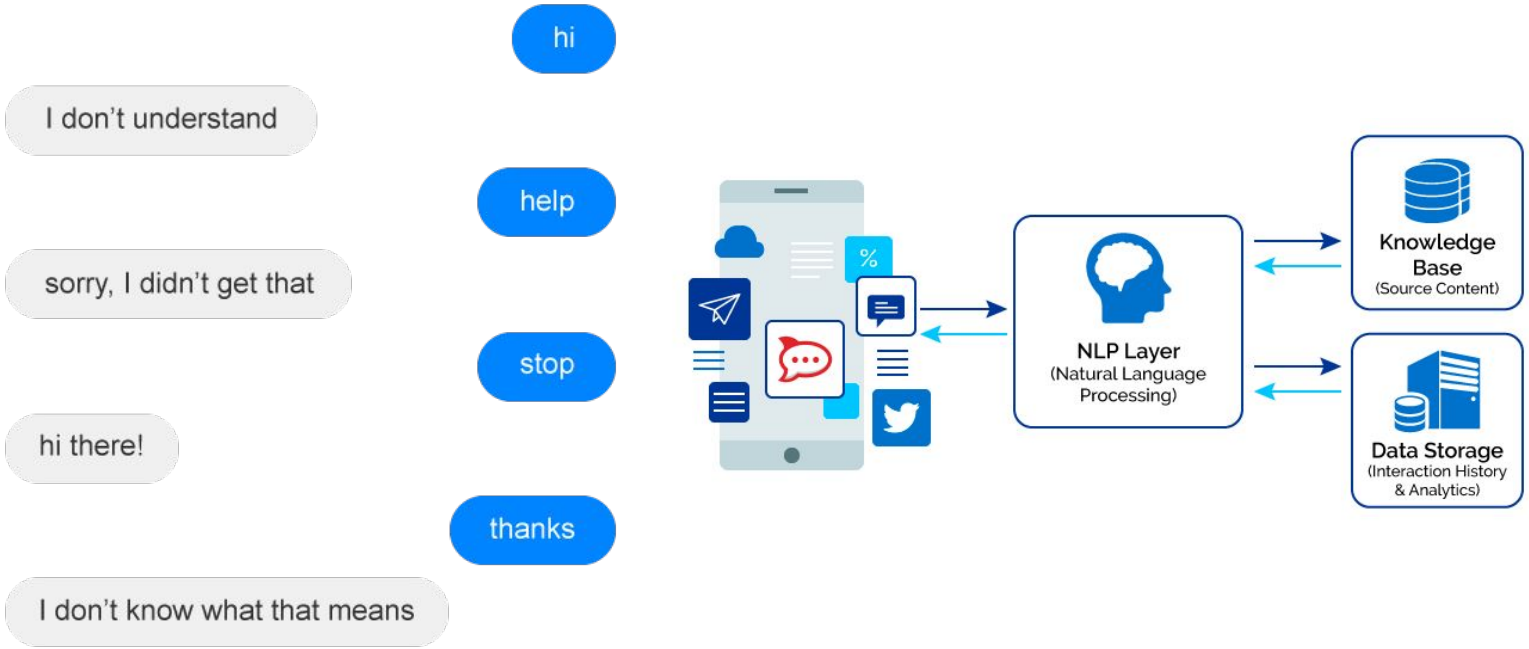
Sentiment Analysis



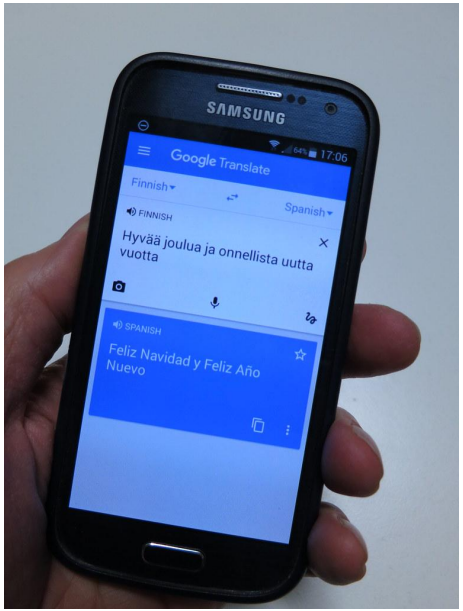


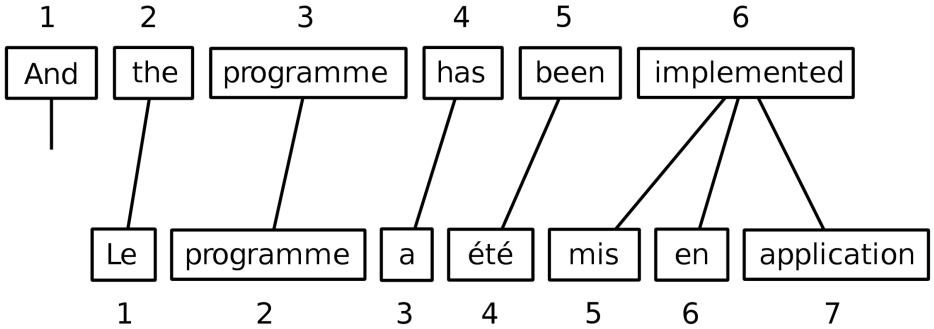
Source: Skyl.AI

Chatbots


hi
I don't understand
help
sorry, I didn't get that
stop
hi there!
thanks
I don't know what that means
NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)

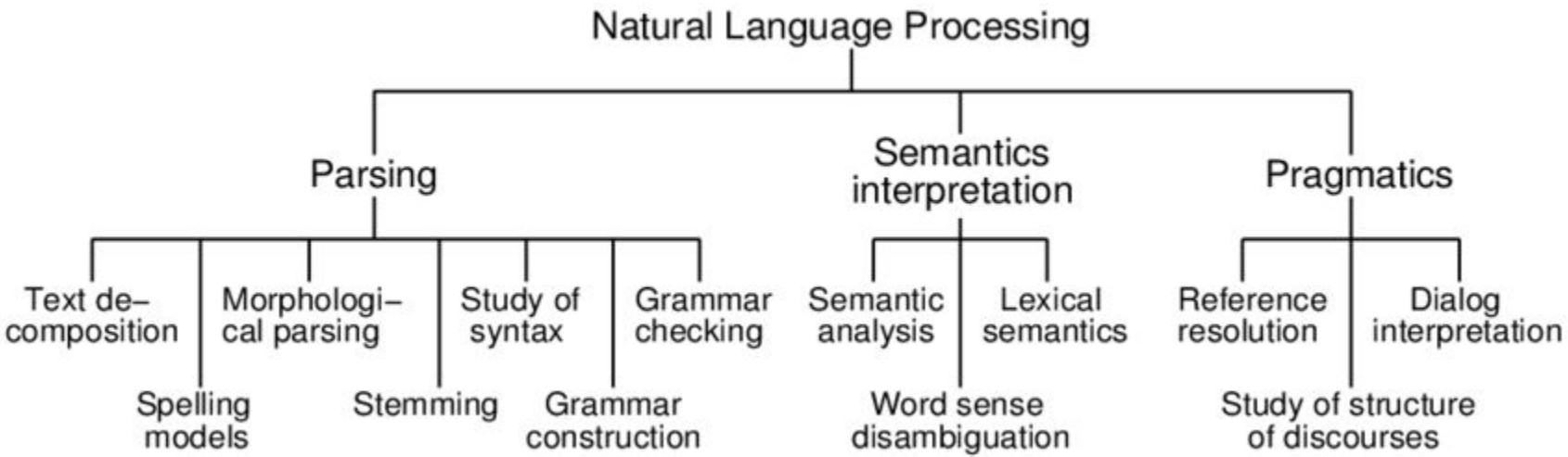
Translator Systems





Source: Skyl.AI

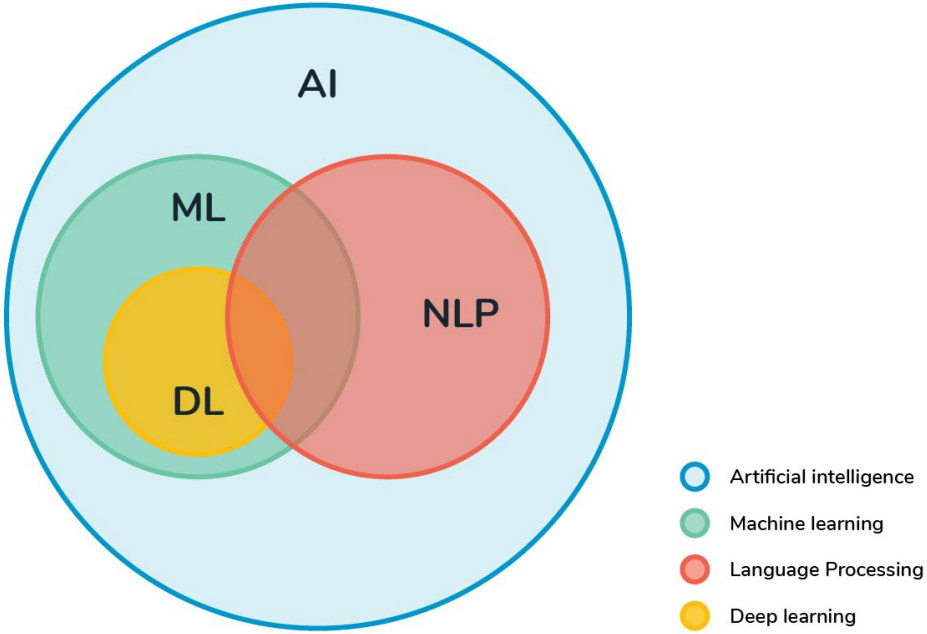
Natural Language Processing



A Primer to Natural Language Processing Research

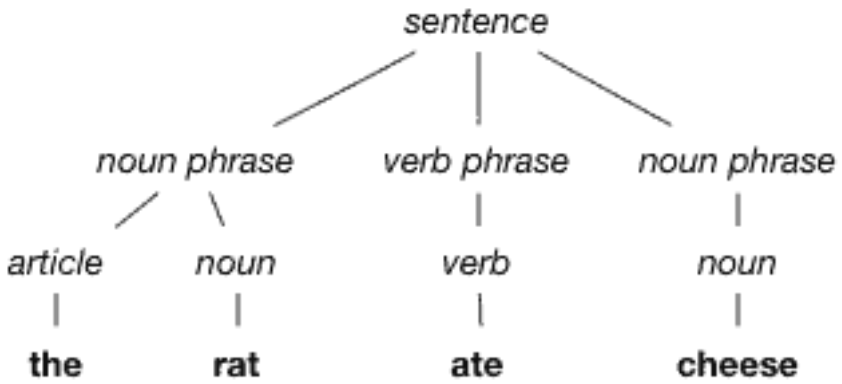
Computationally intelligent methods for qualitative data analysis. Lensu, A. (2002).

Natural Language Processing



Rule-Based NLP

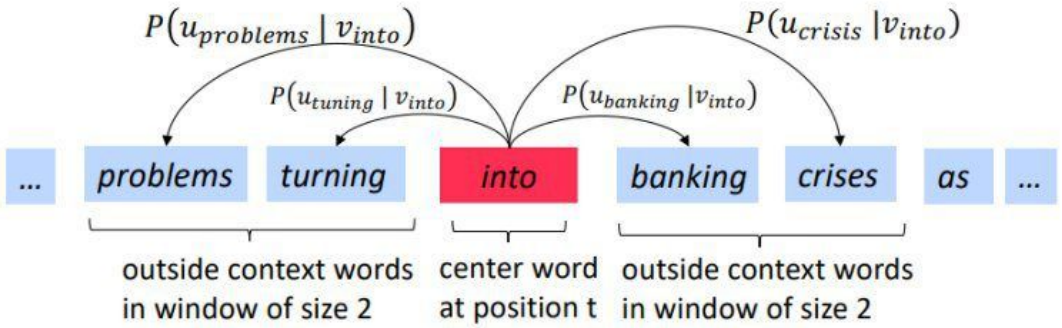
- One of the oldest approaches in NLP.
- Tried and true, proved to work well.
- Often rely and focus on pattern-matching or parsing.



Machine Learning + NLP

- Makes use of statistical learning algorithms such as probabilistic modelling, likelihood maximization, and linear classifiers to recognize patterns from features of texts.

- Involves training a model on parameters followed by fitting on test data (typical of machine learning systems in general).



Deep Learning + NLP

- Makes use of neural network-based methods where feature extraction is skipped as the networks are powerful enough learn important features.

- High demand for data when using complex networks.

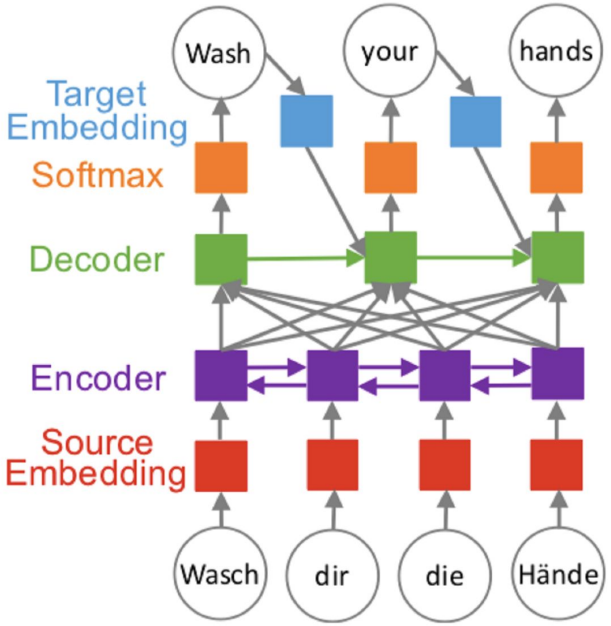
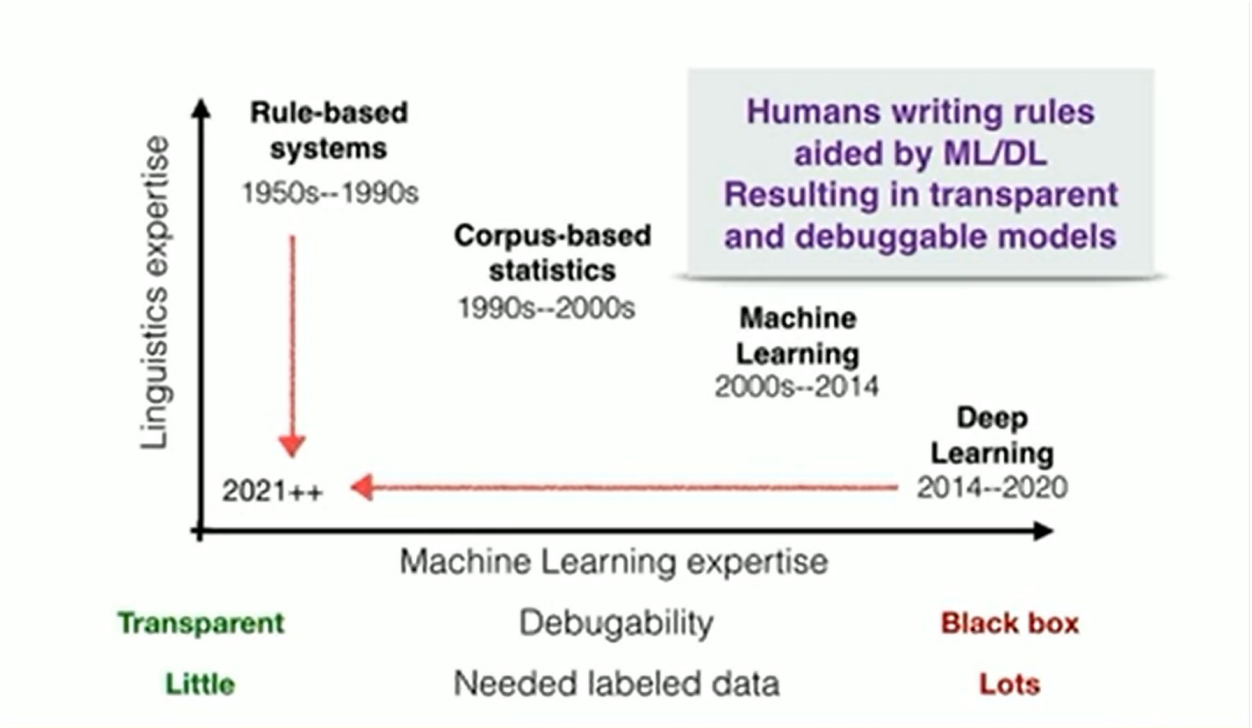
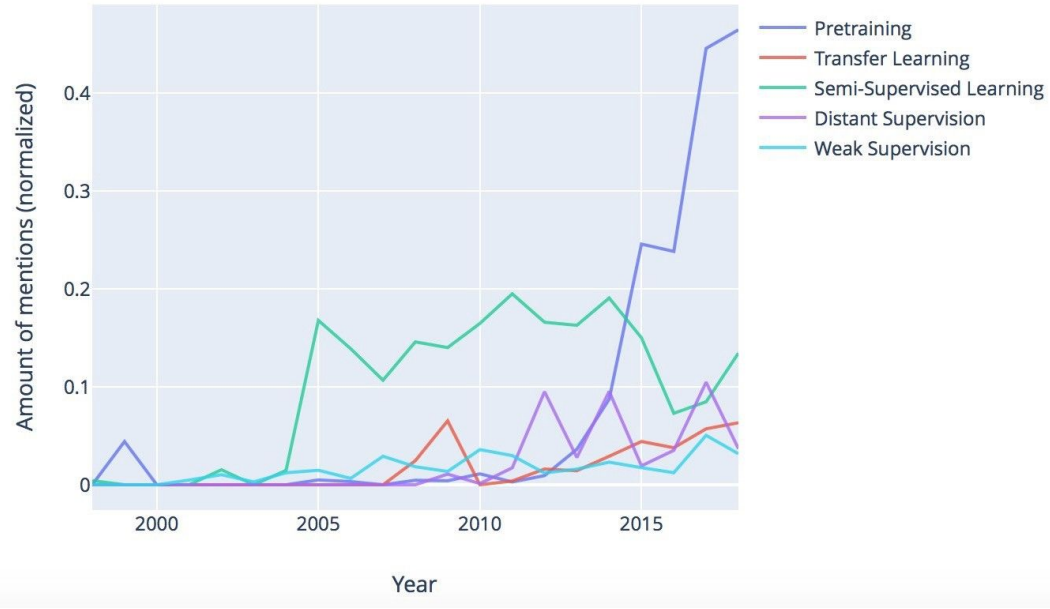


Figure 1: Visualization of an NMT system segmented into components.

Growth of NLP Research

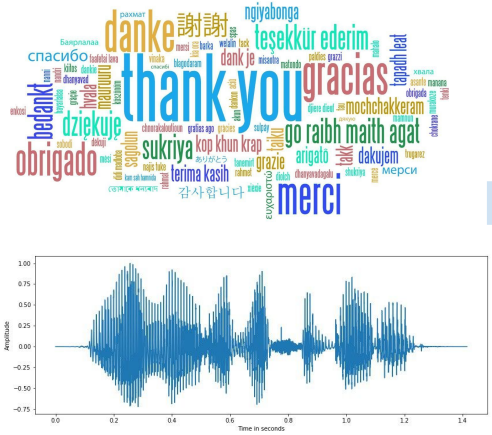


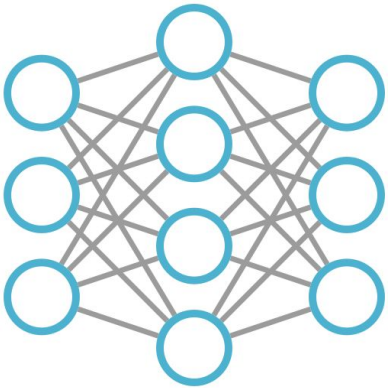
Growth of NLP Research



With the introduction of the Internet, NLP became even more important with the flood of textual information becoming machine accessible.

The NLP Research Framework



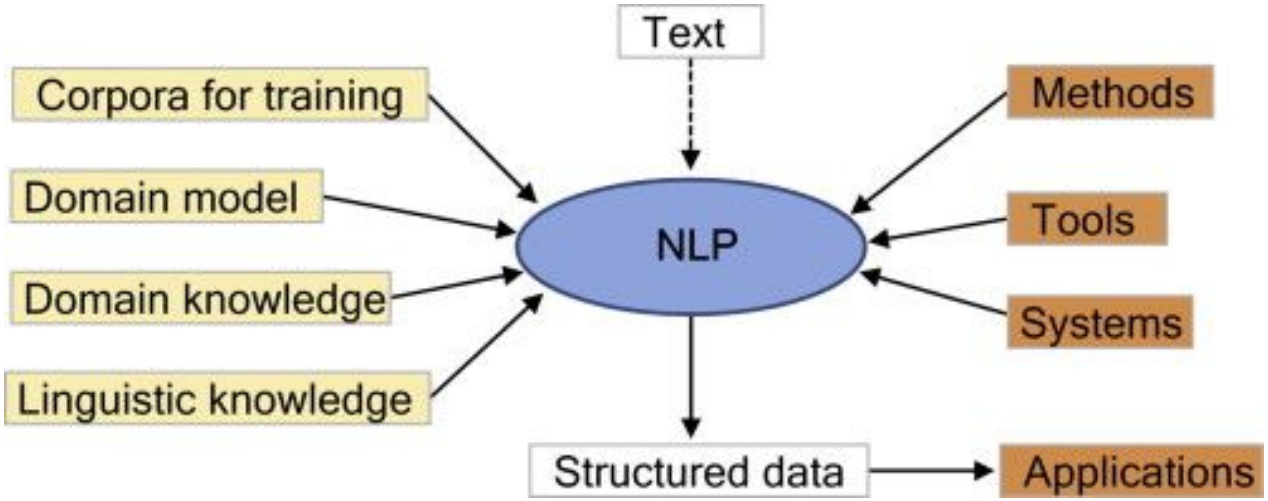


Text or speech data (audio files, tweets, articles, etc)

Algorithms (rule-based, ML, DL)

Trained model

The NLP Research Framework



A Primer to Natural Language Processing Research

Friedman, C., Rindflesch, T. C., & Corn, M. (2013).

NATIONAL UNIVERSITY

Understanding Sentiments of Filipinos During Typhoons

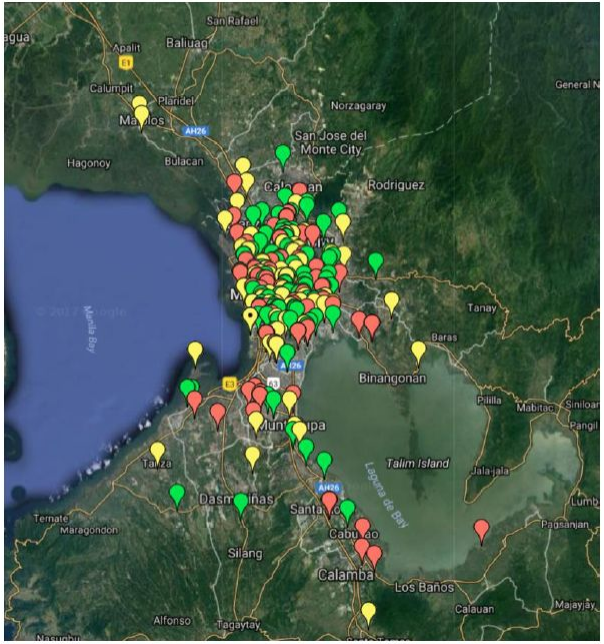
An undergraduate thesis by Joseph Marvin Imperial, Shiela Mae Mazon, Jerome Orozco at Bicol University. Advised by Dr. Lany Maceda.

A Primer to Natural Language Processing Research

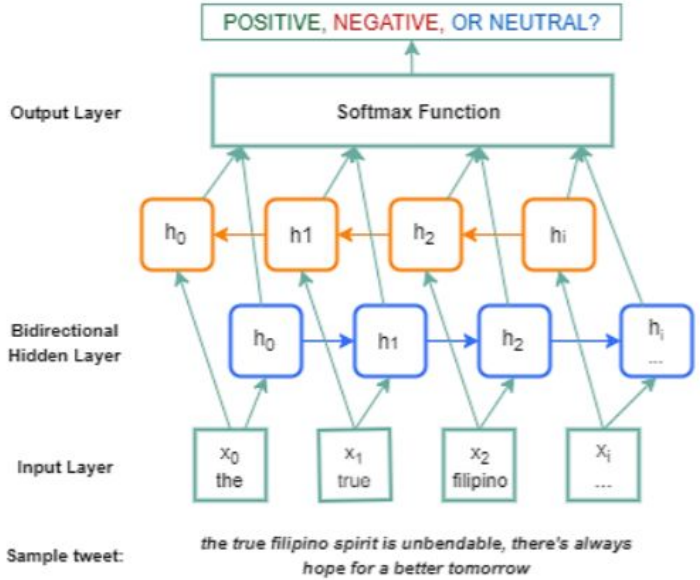
Understanding Sentiments of Filipinos During Typhoons

We analyzed over 92,040 tweets using hashtags **#YolandaPH**, **#BangonPH**, and **#BangonPilipinas** starting from November 1, 2013 to January 1, 2014.

We used deep learning technique (RNN) to understand the semantics behind each tweet, classifying it into three emotional categories: **positive**, **negative**, or **neutral**.



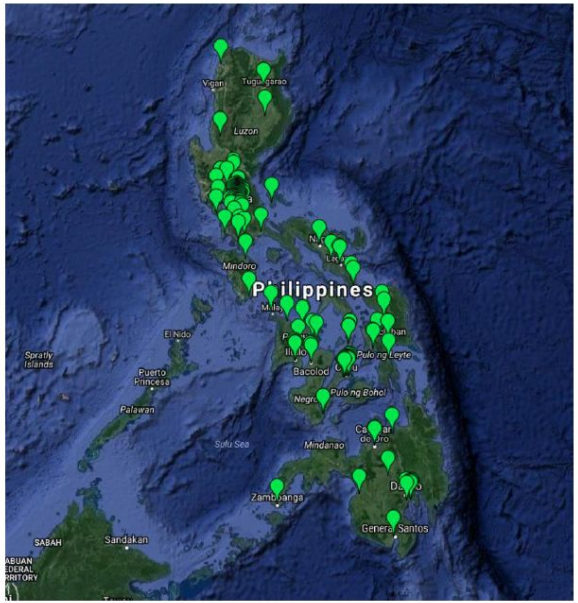
Understanding Sentiments of Filipinos During Typhoons

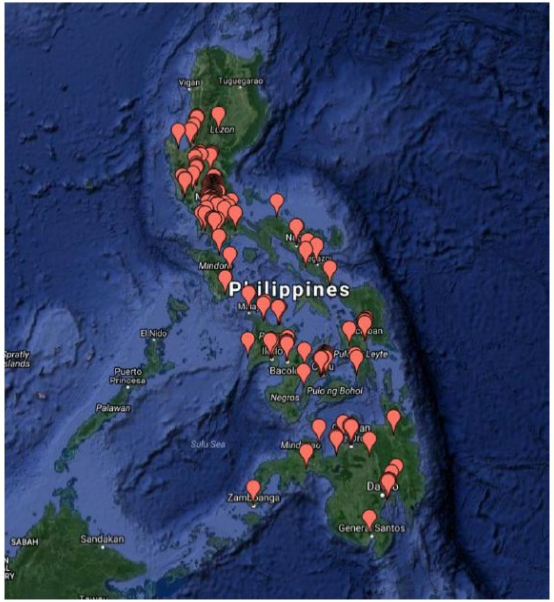


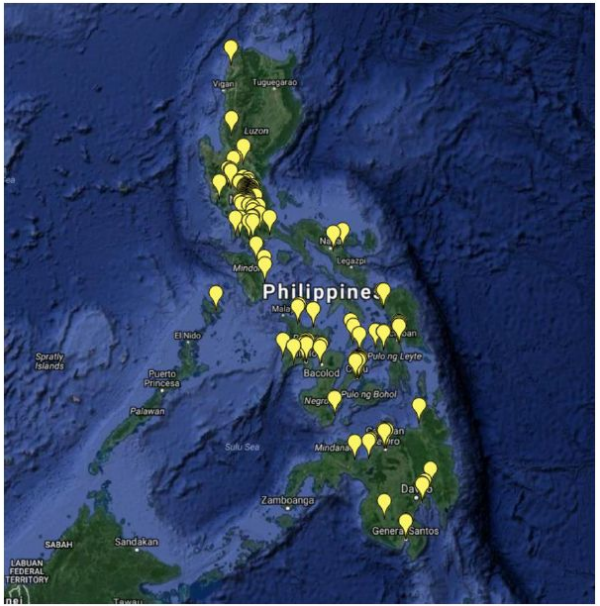
A Primer to Natural Language Processing Research

Understanding Sentiments of Filipinos During Typhoons

NATIONAL UNIVERSITY







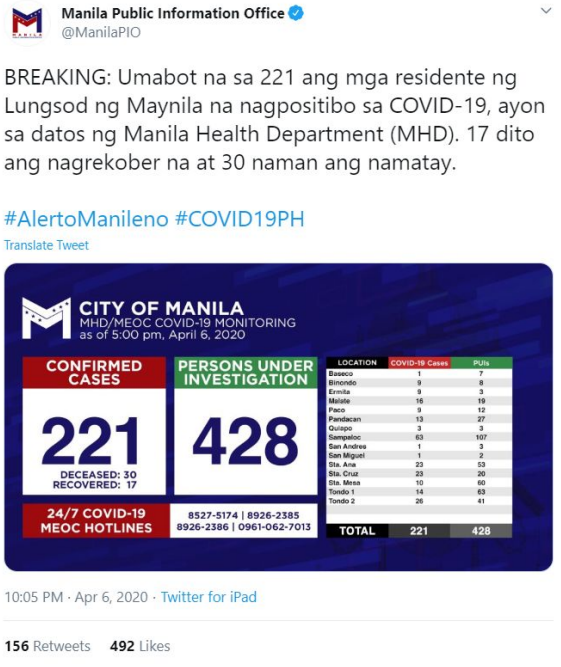
A Primer to Natural Language Processing Research

Modeling Themes from COVID-19 Public Reports

A research article written by Joseph Marvin Imperial. Published in Medium and DLSU's COVID Research Portal.

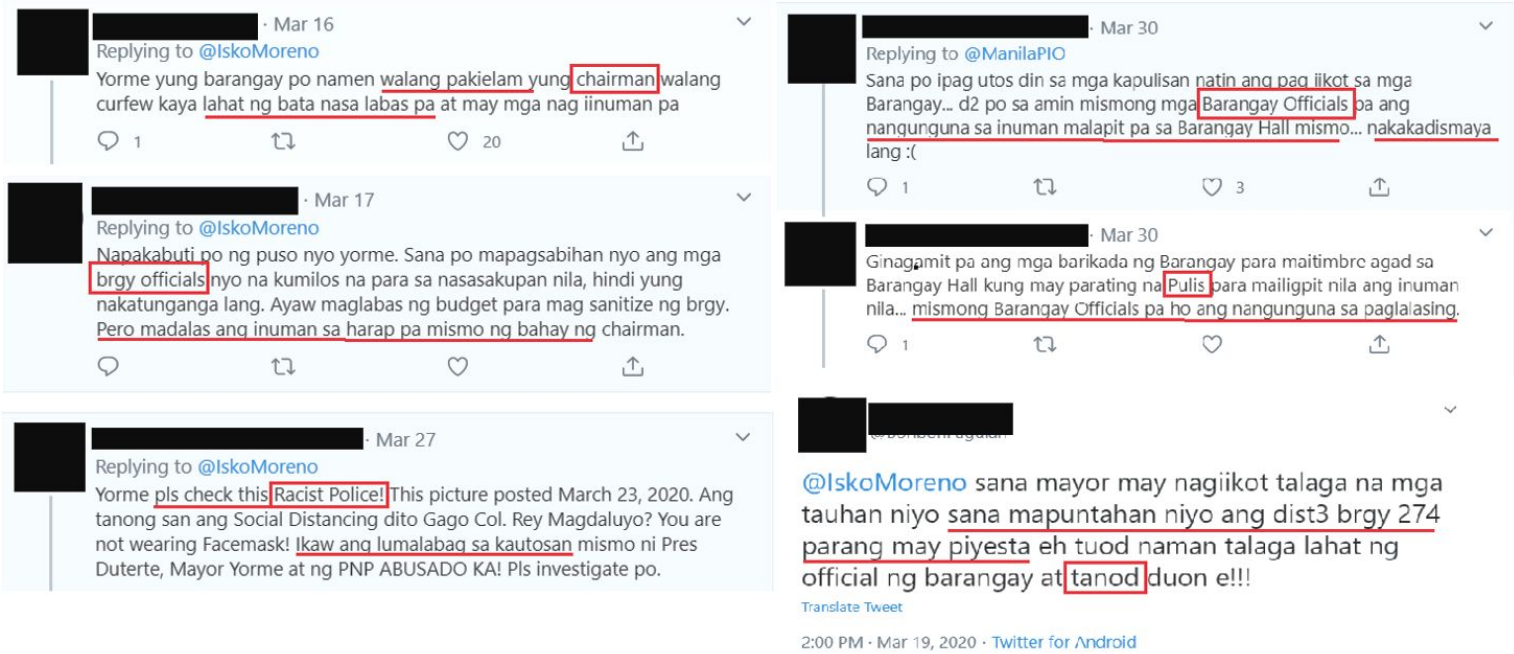
A Primer to Natural Language Processing Research

Modeling Themes from COVID-19 Public Reports



Modeling Themes from COVID-19 Public Reports

NATIONAL UNIVERSITY



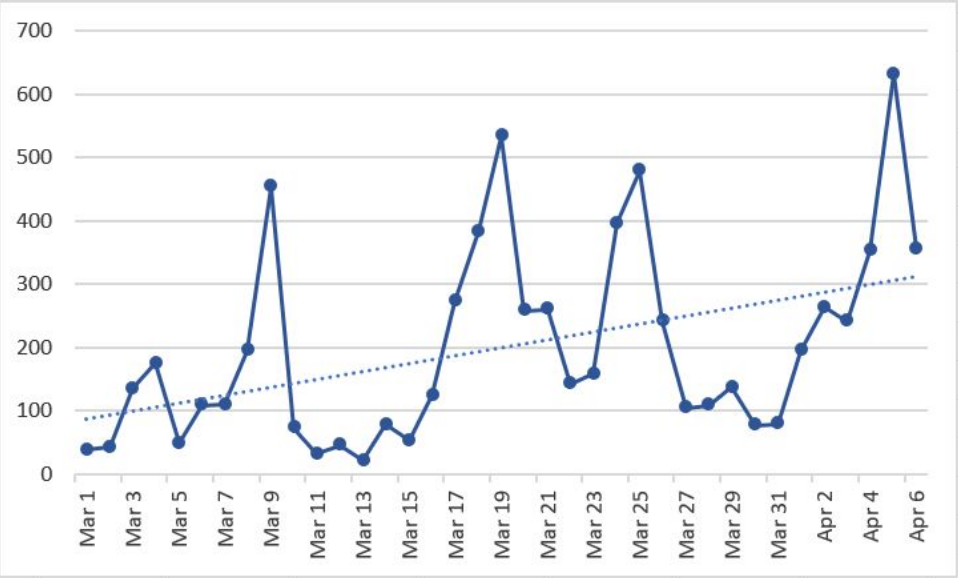
Modeling Themes from COVID-19 Public Reports

NATIONAL UNIVERSITY

	A	B	C	D	E	F	G	H	I	J	K
1	id	conversati	created_a	date	time	timezone	user_id	username	name	place	tweet
2	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	22:09:50	Malay Pen	4.15E+09	[redacted]	[redacted]		Very good project, Yorme!
3	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	21:58:15	Malay Pen	1.10E+18			on	@IskoMoreno Yorme, Manila Post Office @Ppostofficial. Tulong po. Paki bantayan sana ang mga sulat naming mga OFW. Nagpadala
4	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	21:37:55	Malay Pen	1.72E+08			a	@IskoMoreno GOOD MORNING YORME. WHY THE VENDORS, ILLEGAL PARKINGS, LINIS MO HARAP MO, HOMELESS PROGRAMS AR
5	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:13:43	Malay Pen	1.18E+18				@IskoMoreno nag-quit na po siya, he's already resting yorme. Thank you sa pagnotice mo hehe. https://twitter.com/potatogani/st
6	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:10:27	Malay Pen	1.08E+18				Tinapos nya na journal nya yorme, pero tinapos nya na rin buhay nya.
7	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:00:48	Malay Pen	1.18E+18				@IskoMoreno yorme thank you for this, isa sa mga kagustuhan niya ay nangyari. Now he can rest in peace. we love u @potatogani
8	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	18:34:24	Malay Pen	1.04E+18				@IskoMoreno ay oh ngaun k Lang nakita!!!! Meron p Lang ganto s Idol yormeðŸ˜Š pic.twitter.com/T5reKoVkGP
9	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	16:38:14	Malay Pen	1.10E+18			gins	Yorme paano kung mag lagay ng mga halaman sa bangketa ang barangay... Baka ma-obstruction na naman at hakutin ng DPS.
10	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	15:39:43	Malay Pen	1.10E+18				mas mainam po Yorme at magiging inspired ang bawat tao kung ang mabebenipisyo ng reward na yan ay ang mismong taong-barang
11	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	15:09:28	Malay Pen	1.01E+18			° Ajumma	Yorme, dami na naman pong mga nakapark sa gedli baka naman po pede mamompyang ng baranggay Chairman.
12	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	14:58:33	Malay Pen	1.31E+09				Talo na ang Pacheco at bandang Madrid, puro kasi tae ng aso't pusa ang nasa kalsada nila, walang kusa. Doon pa naman ako banda u
13	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	13:21:47	Malay Pen	8.87E+17				Yorme, you may have good intentions but this feels inappropriate and open to corruption. Public funds must be used by all, not won
14	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	22:09:50	Malay Pen	4.15E+09				Very good project, Yorme!
15	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	21:58:15	Malay Pen	1.10E+18			on	@IskoMoreno Yorme, Manila Post Office @Ppostofficial. Tulong po. Paki bantayan sana ang mga sulat naming mga OFW. Nagpadala
16	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	21:37:55	Malay Pen	1.72E+08			a	@IskoMoreno GOOD MORNING YORME. WHY THE VENDORS, ILLEGAL PARKINGS, LINIS MO HARAP MO, HOMELESS PROGRAMS AR
17	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:13:43	Malay Pen	1.18E+18				@IskoMoreno nag-quit na po siya, he's already resting yorme. Thank you sa pagnotice mo hehe. https://twitter.com/potatogani/st
18	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:10:27	Malay Pen	1.08E+18				Tinapos nya na journal nya yorme, pero tinapos nya na rin buhay nya.
19	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	20:00:48	Malay Pen	1.18E+18				@IskoMoreno yorme thank you for this, isa sa mga kagustuhan niya ay nangyari. Now he can rest in peace. we love u @potatogani
20	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	18:34:24	Malay Pen	1.04E+18				@IskoMoreno ay oh ngaun k Lang nakita!!!! Meron p Lang ganto s Idol yormeðŸ˜Š pic.twitter.com/T5reKoVkGP
21	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	16:38:14	Malay Pen	1.10E+18			gins	Yorme paano kung mag lagay ng mga halaman sa bangketa ang barangay... Baka ma-obstruction na naman at hakutin ng DPS.
22	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	15:39:43	Malay Pen	1.10E+18				mas mainam po Yorme at magiging inspired ang bawat tao kung ang mabebenipisyo ng reward na yan ay ang mismong taong-barang
23	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	15:09:28	Malay Pen	1.01E+18			° Ajumma	Yorme, dami na naman pong mga nakapark sa gedli baka naman po pede mamompyang ng baranggay Chairman.
24	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	14:58:33	Malay Pen	1.31E+09				Talo na ang Pacheco at bandang Madrid, puro kasi tae ng aso't pusa ang nasa kalsada nila, walang kusa. Doon pa naman ako banda u
25	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	13:21:47	Malay Pen	8.87E+17				Yorme, you may have good intentions but this feels inappropriate and open to corruption. Public funds must be used by all, not won
26	1.23E+18	1.23E+18	1.58E+12	01-Mar-20	16:23:18	Malay Pen	1.22E+18				pwede nyo po ba turuan mayor ng paranaque? traffic, illegal parking, pasaway ma triccycle especially sa better living po. madumi kal

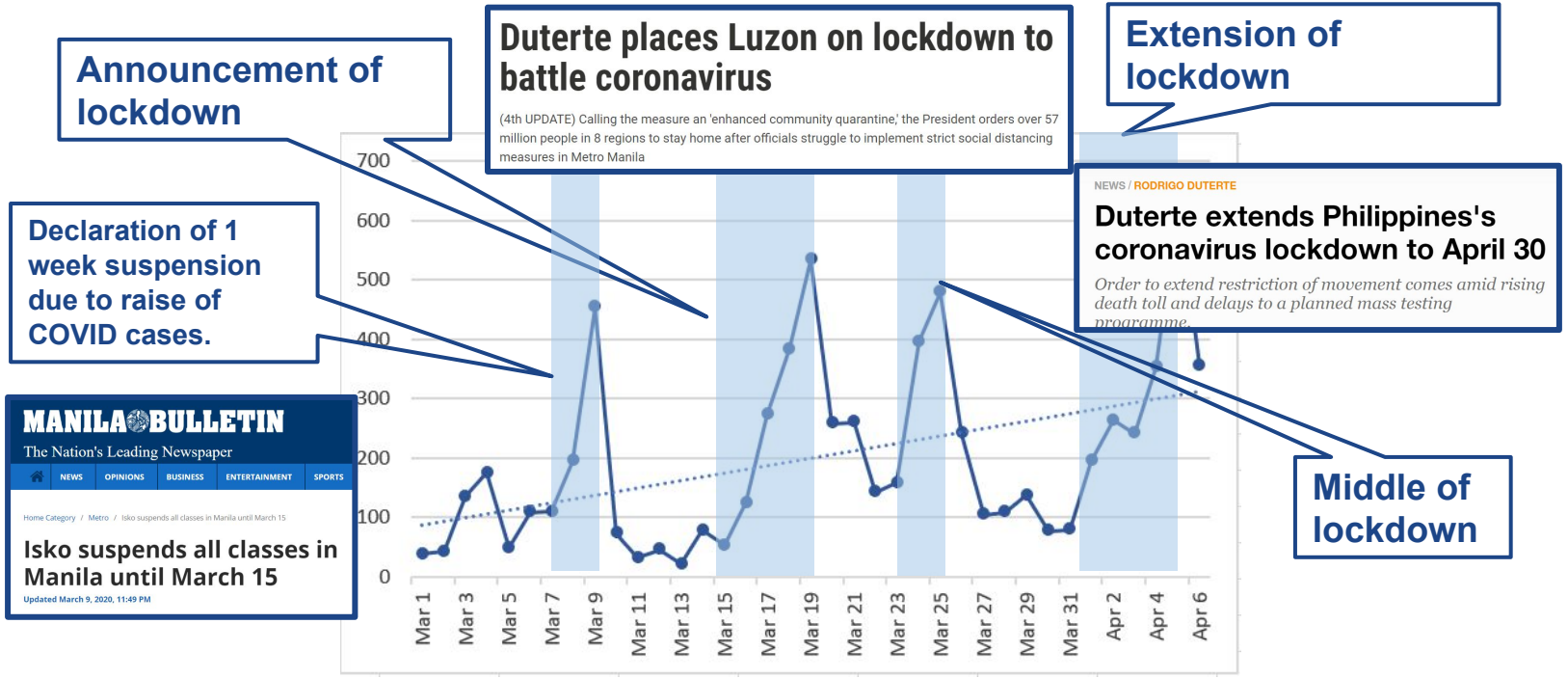
Modeling Themes from COVID-19 Public Reports

NATIONAL UNIVERSITY



A Primer to Natural Language Processing Research

Modeling Themes from COVID-19 Public Reports



Modeling Themes from COVID-19 Public Reports

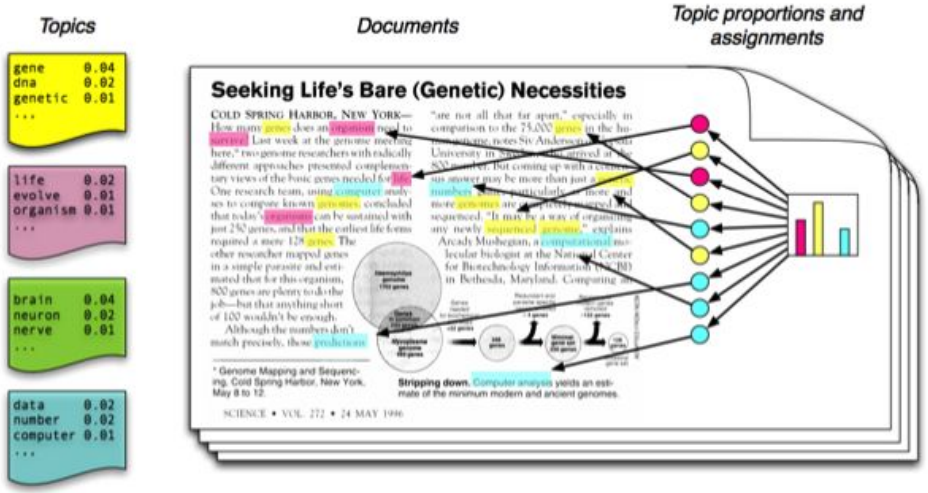
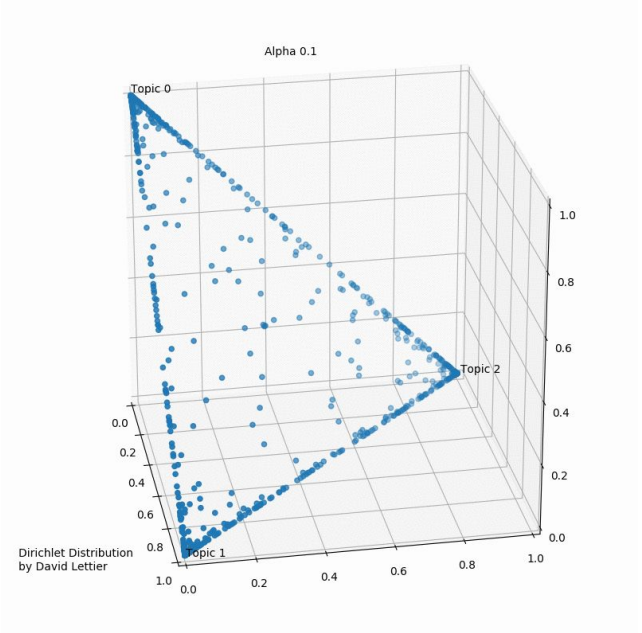


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.



A Primer to Natural Language Processing Research

Modeling Themes from COVID-19 Public Reports

Using **BTM topic modelling** revealed four central themes from the public reports:

1. Continuous declaration of needs (relief goods, cash).
2. Abuse and negligence of barangay captains and police.
3. Use of tricycles for ferrying COVID patients.
4. Petition to suspend classes in NCR indefinitely.

Modeling Themes from COVID-19 Public Reports

BRGY #	COMPLAINT
33	brgy 33 pasuyo po dito maayos. **madami pa din pong tambay sa labas** regardless anung oras
83	brgy 83 zone 7 po is **p500 lang binigay instead of 1k**. sabi po ng asawa ng chairman, city hall daw po nag-utos na pagkasyahin na lang and pinipilit po nila na 460 daw po ang number ng families dun, iba po sa pinost niyo po
96	brgy. 96, zone 8, district 1, tondo, manila. **indi daw kasama nanay ko though senior sya** at wala sa listahan nila.
101	barangay 101 lng ata walang balak sa ganyan yorme hahahaah **hanggang pirma lng tapos walang dararing** hahahah
174	brgy 174 **never pa kami nabbgyan dito simula nung lockdown** po
182	brgy 182 simula ng ecq inerereklamo ko na . **until now, walang aksyon**. pyesta araw araw. may darating na pulis, tago lahat ng tao. di pa nakakaliko ng kanto labasan uli. **may bingo, inuman, bukas na computer shop, tong-its**. takot ang barangay kasi baka di na sila iboto...
342	brgy 342, **4 senior citizen wala pang natatanggap** mg gates o php1,500 na sinabi
500	brgy 500 po **god bless** dps

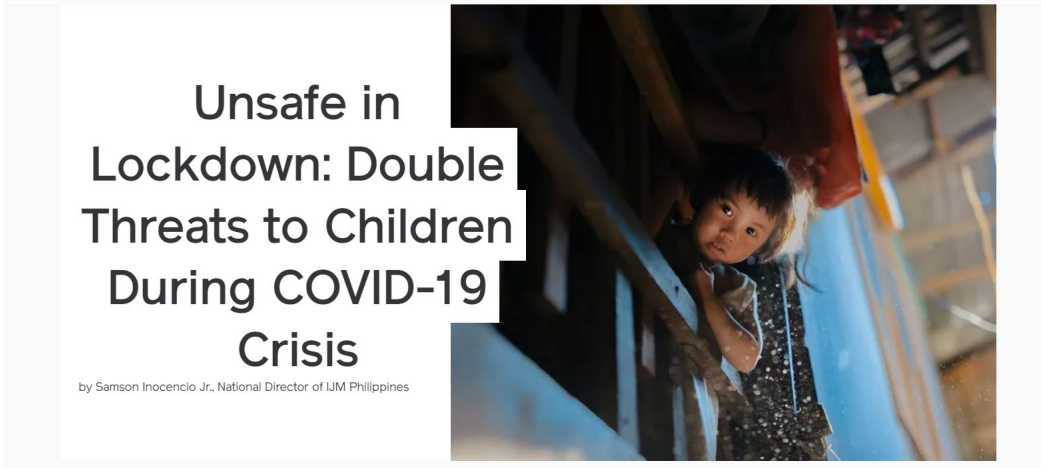
NATIONAL UNIVERSITY

Investigating Child Pornography and Exploitation on Twitter During Lockdown

A data science project by Michael Ibañez, Ranz Sapinit, Lloyd Reyes, and Mohammed Hussien. Advised by Joseph Marvin Imperial.

A Primer to Natural Language Processing Research

Investigating Child Pornography and Exploitation on Twitter During Lockdown



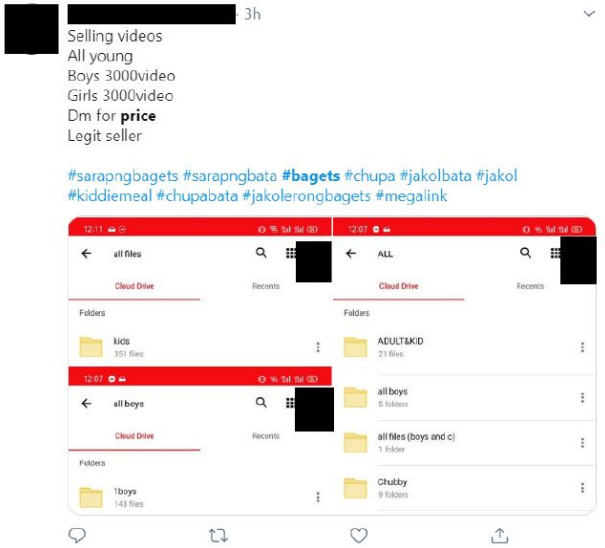


Fig. 1. Example tweet showing a peddler marketing child pornographic material on Twitter.

Investigating Child Pornography and Exploitation on Twitter During Lockdown

Fig. 2. Word cloud visualization of top-occurring / high-frequency words from the data.

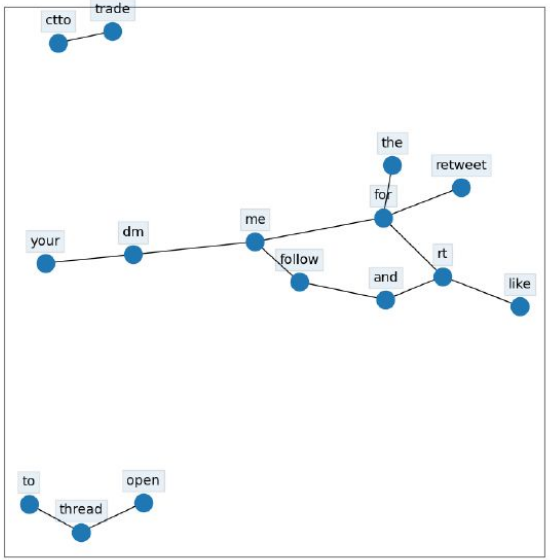


Fig. 3. Trigram network indicating terms that frequently co-occur in child pornographic tweets.

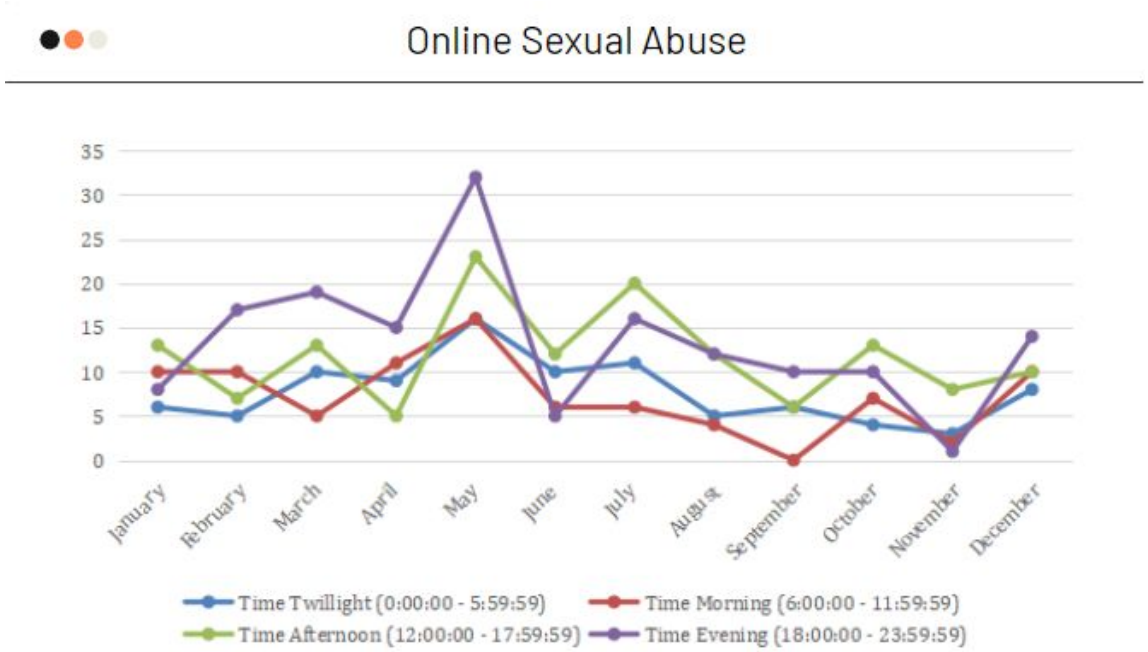
Investigating Child Pornography and Exploitation on Twitter During Lockdown

NATIONAL UNIVERSITY

TABLE I
ROLES OF SEXUAL PREDATORS BASED ON THEMATIC WORD USAGE.

Persona	Vocabulary
The Propagator	rt, follow, vid, retweet, videos, like, link, jakol, post, comment
The Peddler	bagets, dm, price, pic, area, php, avail, looking, pls, willing
The Social	tara, face, pogi, pm, jakol, jan, pwede, want, sino tayo
The Voyeur	boso, face, tara, pic, dm, bagets, baby, sarap, cr, boys

Investigating Child Pornography and Exploitation on Twitter During Lockdown



A Primer to Natural Language Processing Research

The NLP Research Primer

NATIONAL UNIVERSITY

The primer contains short discussions on basic NLP processes, example published NLP papers by Filipino students and researchers, open-source codes and repositories, and links to online tools to help you kickstart your own NLP project.

bit.ly/nlp-primer-ph

NLP Mini Projects

NATIONAL UNIVERSITY

1. **Hate speech classifier** - what words make a text convey hate speech or not?
2. **Next word prediction using language models** - if you train a language model from a large text, it can probabilistically guess the next word.
3. **Spell checker with edit-distance formulas** - edit distance formulas calculate changes in texts such as alterations, deletions, and insertions.
4. **Sentiment analyzer using emotion dictionaries** - what words make a text convey strong emotions?
5. **Static string-based chatbots** - use string matching and similarity algorithms to intelligently match answers to queries.

NLP "Thesis-Worthy" Projects

NATIONAL UNIVERSITY

1. **A readability assessment model for X language** - can we develop an automated machine-learning based model for identifying readability of text in other PH languages? Cebuano, Bikolano, Ibanag, Chavacano, etc.

2. **An automated essay scoring tool for X language** - given an essay written by a student, how can we teach machines to automatically grade based on factors that make a good essay such as being cohesive, well organized, contains relevant topics, and grammatically correct?

3. **Suicide ideation detection from social media posts** - can we teach a machine to identify signals that a person is forming / about to form suicidal thoughts based from their public social media posts? Note that this is not a prescriptive approach.

Advice

NATIONAL UNIVERSITY

1. Assess your skill set and domain knowledge.
2. Have a clear goal and well-defined problem.
3. Ensure resources are available.
4. Research how you will represent your data as features.
5. Inspect your model and evaluate correctly with existing baselines.

NATIONAL UNIVERSITY

Computing Society of the Philippines
Special Interest Group - Natural Language Processing (SIG-NLP)

A Primer to Natural Language Processing Research

Joseph Marvin Imperial

Department of Computer Science
NU Human Language Technology Lab
National University