Natural Language Processing

Review of applications and opportunities for IEG

Aivin V. Solatorio

Data Scientist | DECAT

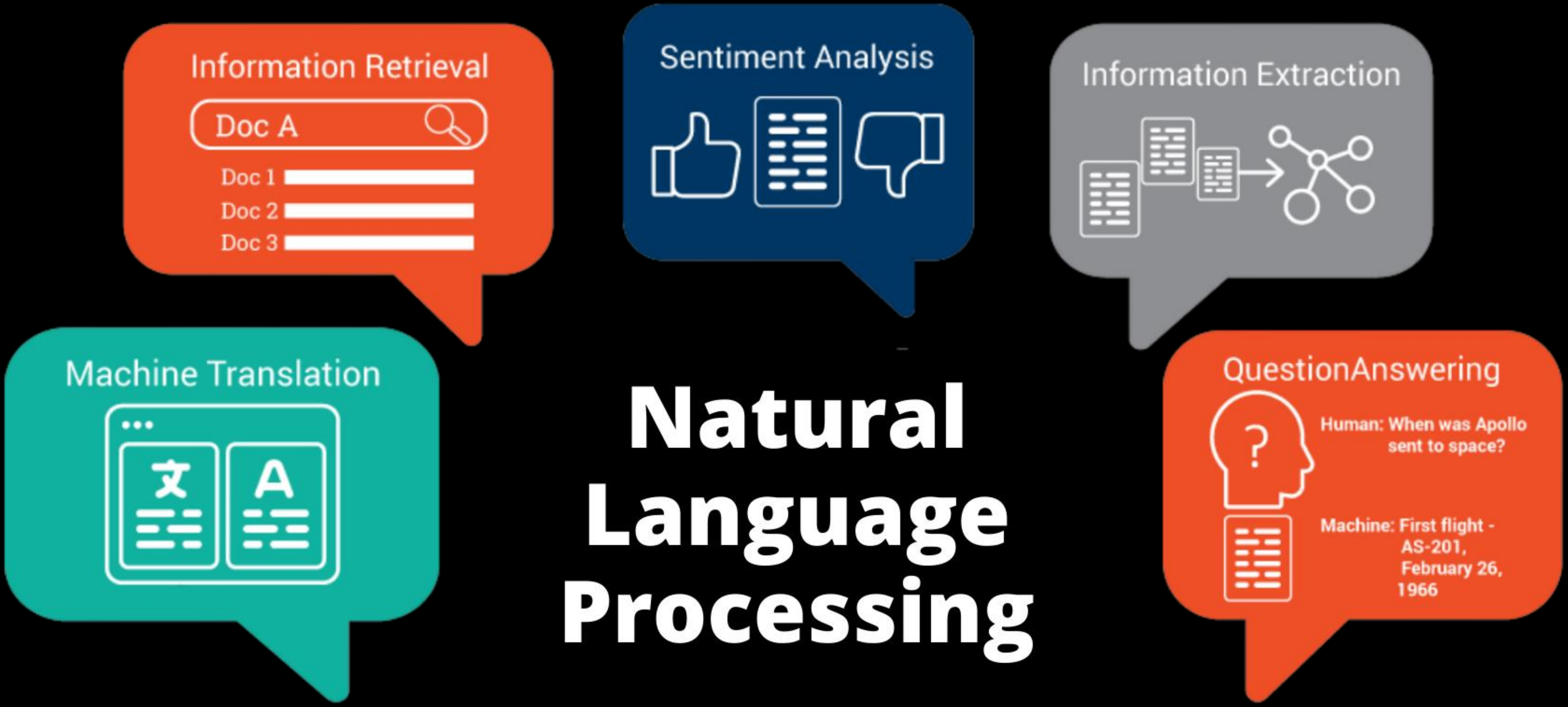
asolatorio@worldbank.org

@avsolatorio

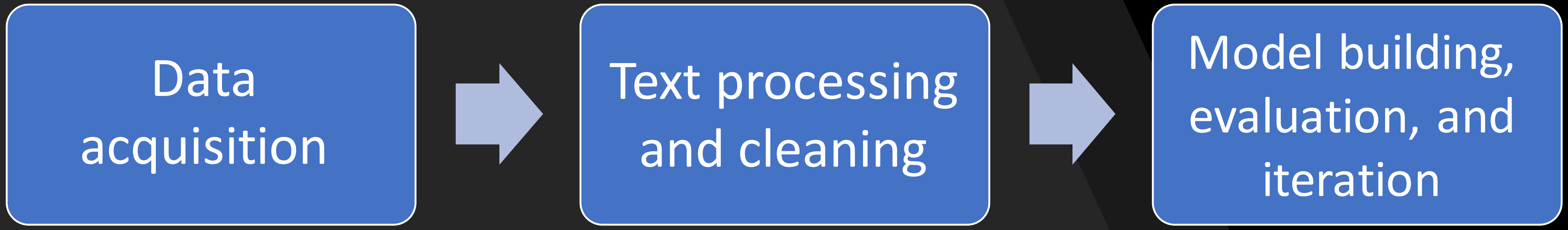
POLL: How familiar are you with NLP?

1. I have never heard of it
2. I have heard about it but I don't know what it is
3. I am familiar with NLP but not a practitioner
4. I have worked in projects that use NLP
5. I am an expert in NLP

What is Natural Language Processing (NLP)?



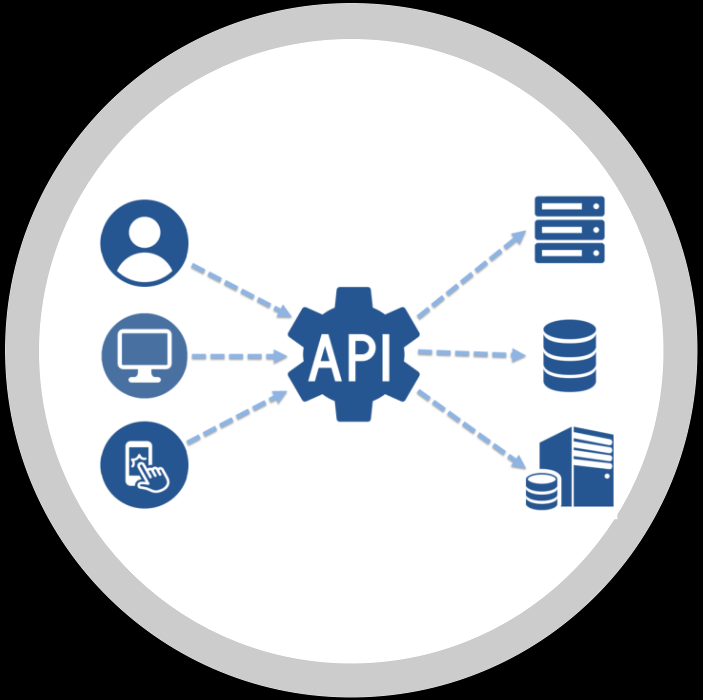
Basic NLP Workflow



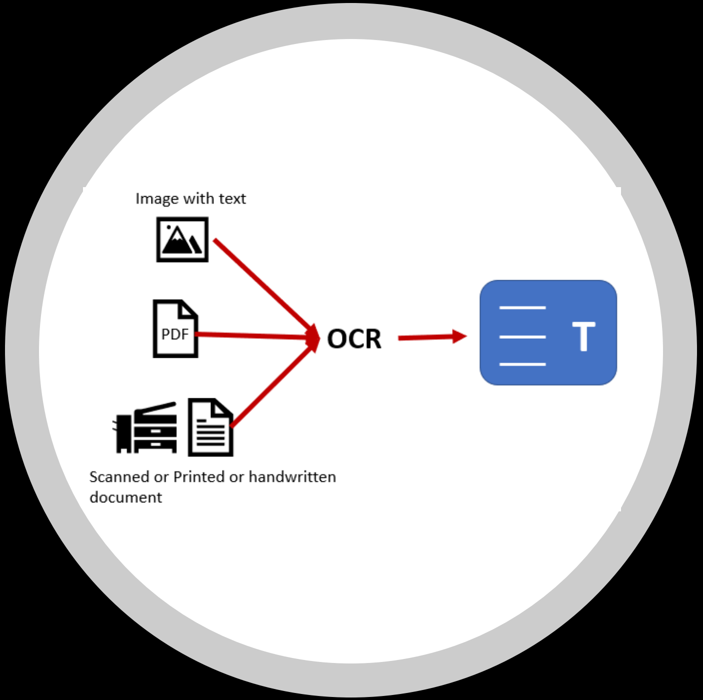
Data acquisition

Text processing and cleaning

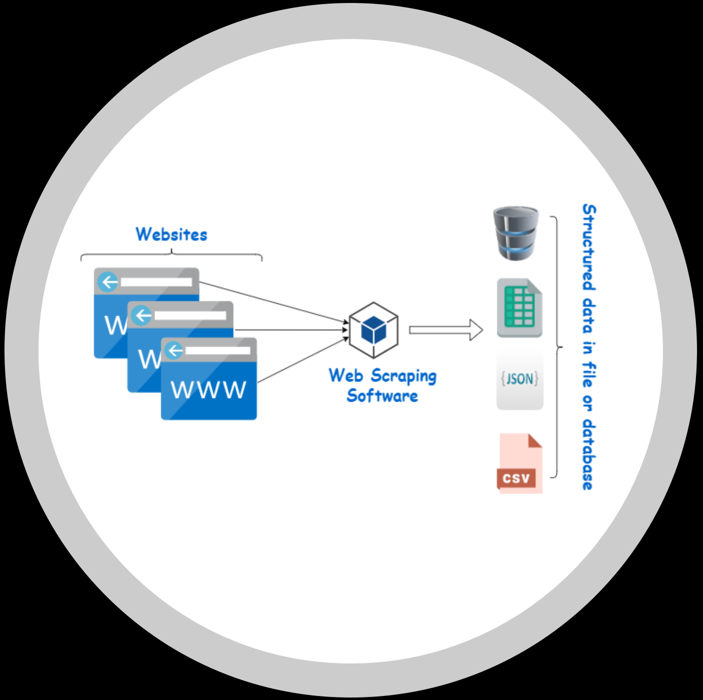
Model building, evaluation, and iteration



API scraping

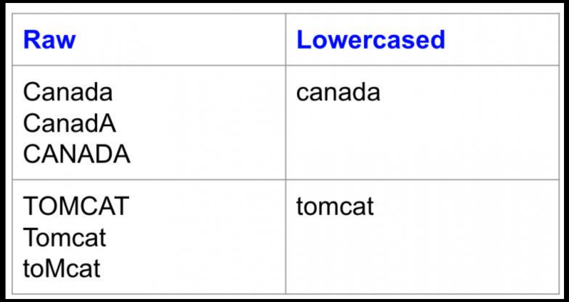


Document parsing



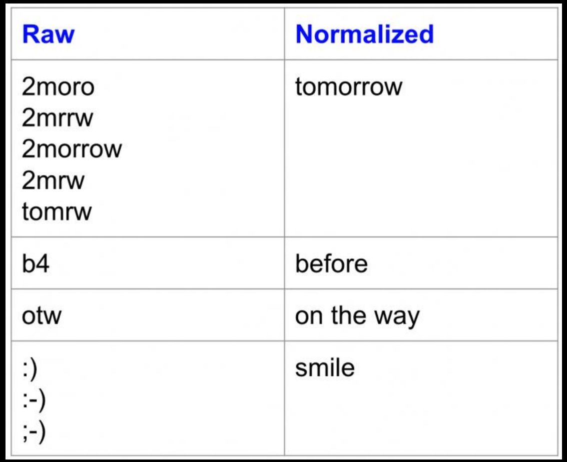
Web scraping

Data acquisition



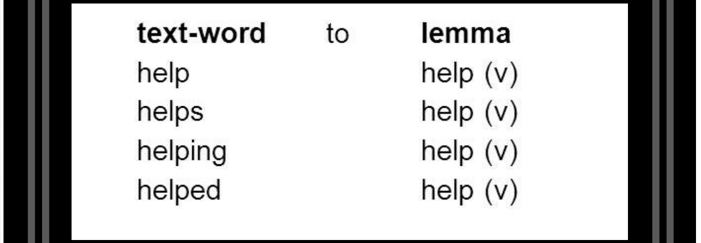
Raw	Lowercased
Canada CanadA CANADA	canada
TOMCAT Tomcat toMcat	tomcat

Lowercase text



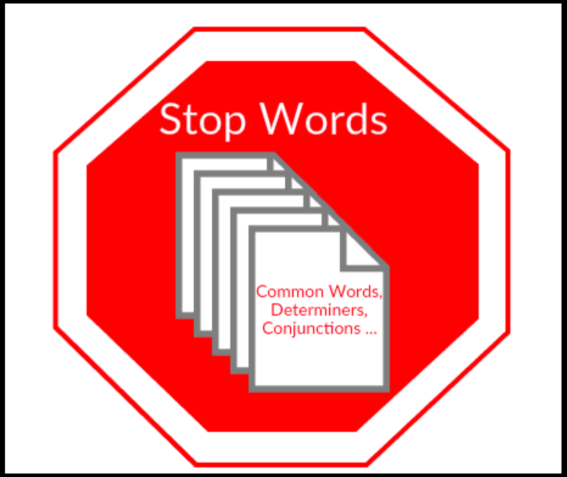
Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

Token normalization



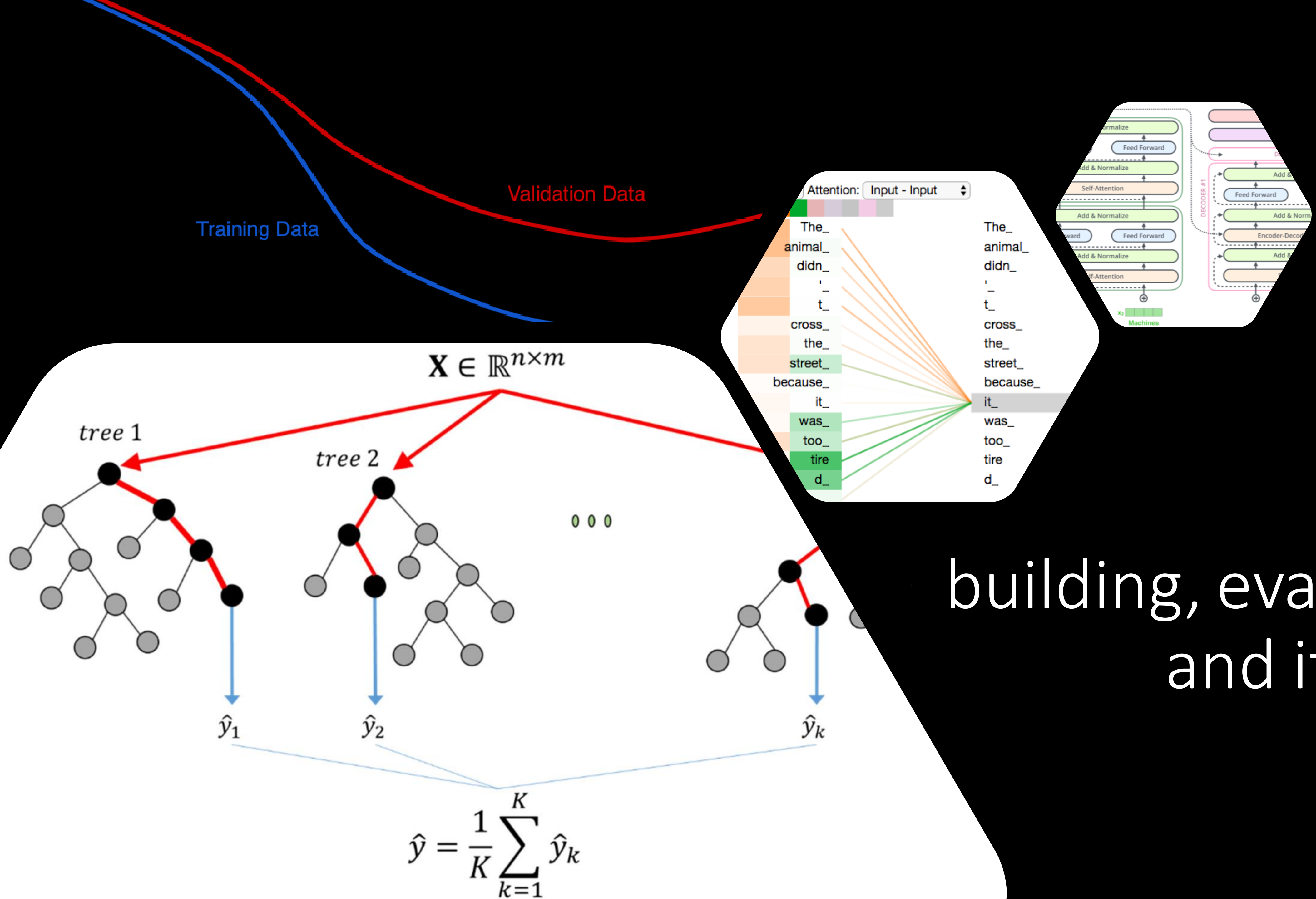
text-word	to	lemma
help		help (v)
helps		help (v)
helping		help (v)
helped		help (v)

Word lemmatization



Stopwords removal

Basic text processing and cleaning



Model building, evaluation, and iteration

NLP Applications

- Data discovery
- Systematic country diagnostics
- Climate change co-benefits prediction

Data Discovery

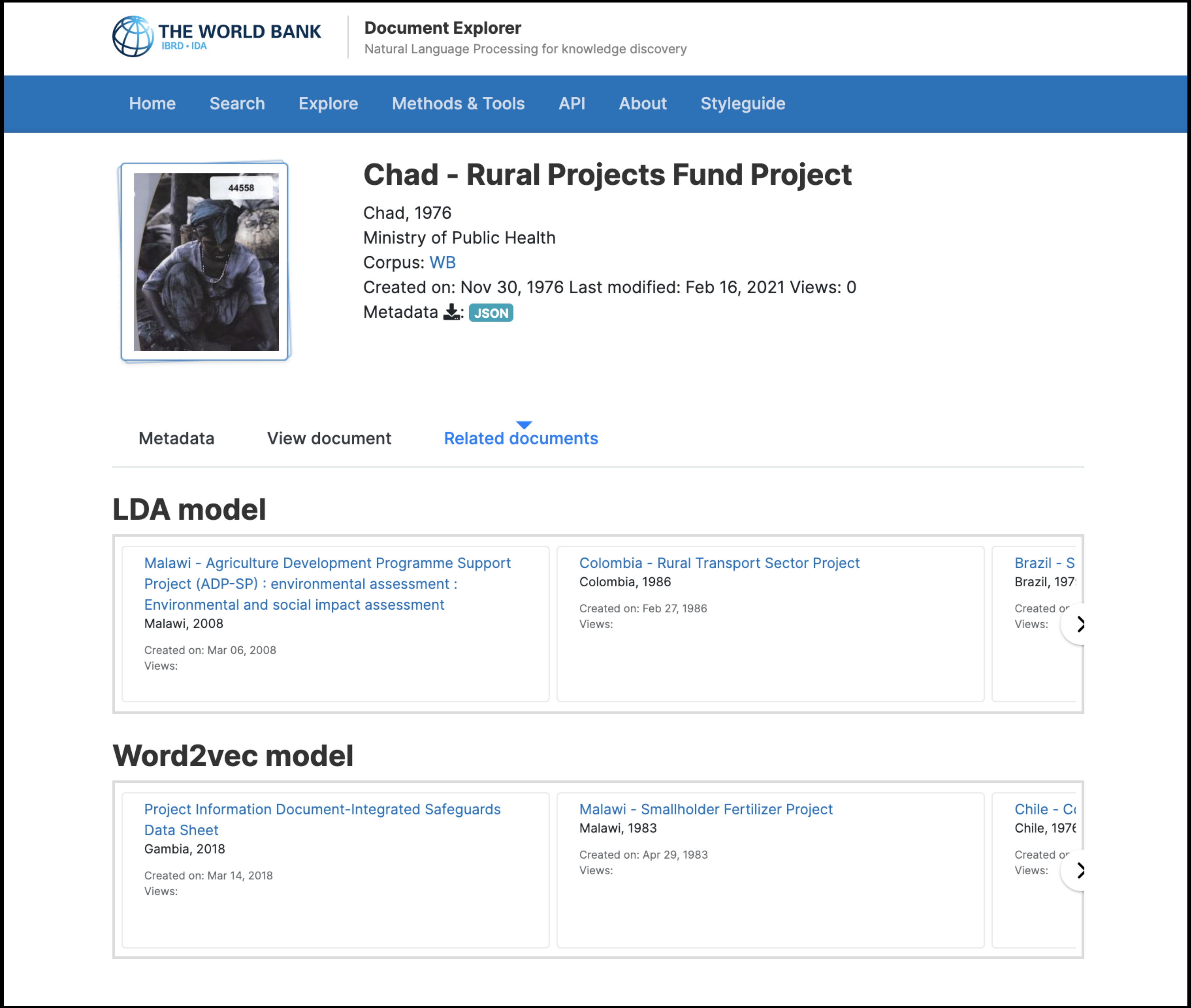
Why is data discovery important?

- Different types of data are available across our institution and beyond.

- This is an asset when we can find the best data or document for our needs out of all the available data.

- However, it **proves to be a liability** when these data only reside in silos and databases that are **not equiped with discovery tools** to allow users to use them.

Document discovery platform



THE WORLD BANK
IBRD • IDA

Document Explorer
Natural Language Processing for knowledge discovery

Home | Search | Explore | Methods & Tools | API | About | Styleguide

44558

Chad - Rural Projects Fund Project

Chad, 1976
Ministry of Public Health
Corpus: WB
Created on: Nov 30, 1976 Last modified: Feb 16, 2021 Views: 0
Metadata: JSON

Metadata | View document | Related documents

LDA model

Malawi - Agriculture Development Programme Support Project (ADP-SP) : environmental assessment : Environmental and social impact assessment
Malawi, 2008
Created on: Mar 06, 2008
Views:

Colombia - Rural Transport Sector Project
Colombia, 1986
Created on: Feb 27, 1986
Views:

Brazil - S
Brazil, 197
Created on
Views:

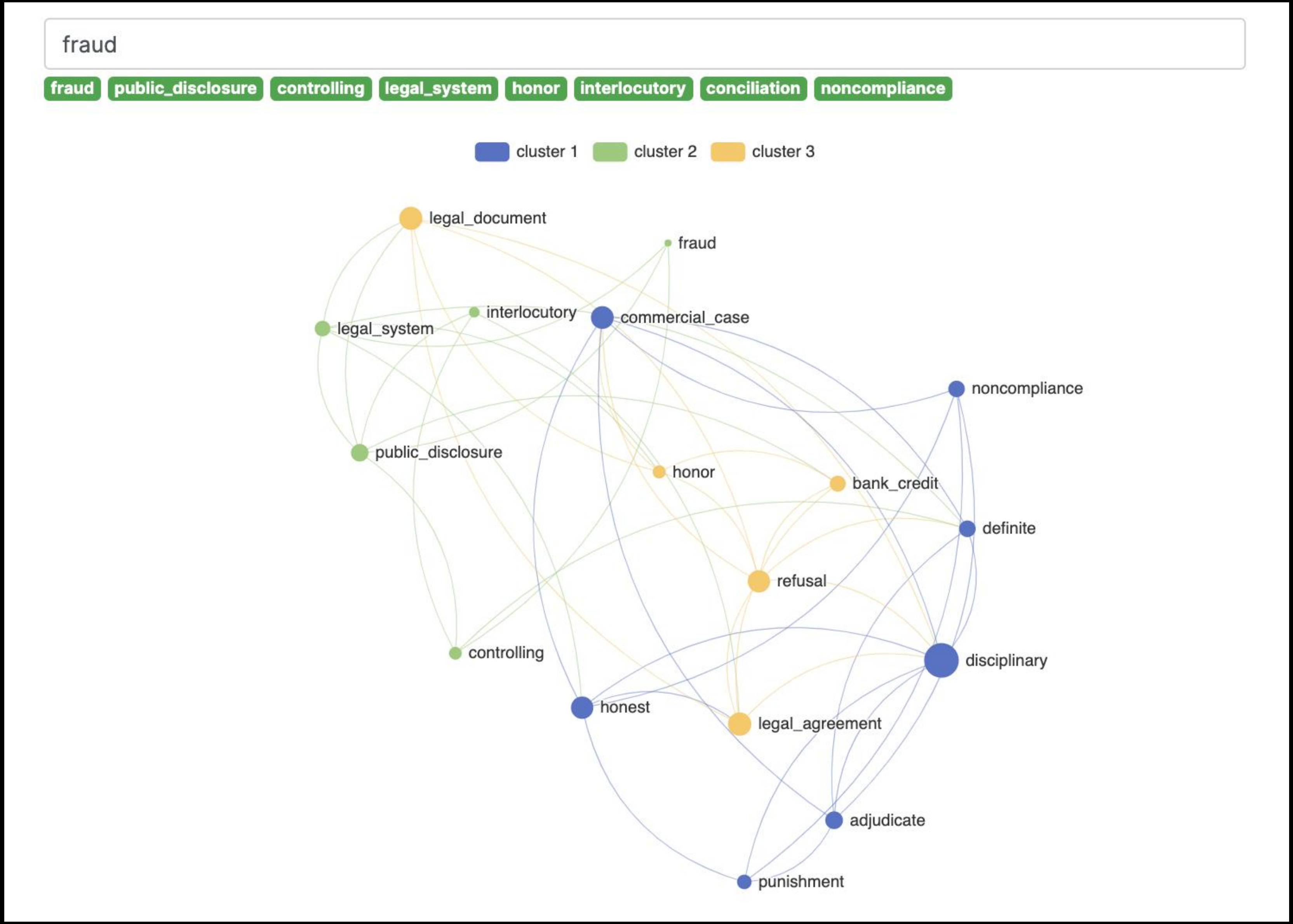
Word2vec model

Project Information Document-Integrated Safeguards Data Sheet
Gambia, 2018
Created on: Mar 14, 2018
Views:

Malawi - Smallholder Fertilizer Project
Malawi, 1983
Created on: Apr 29, 1983
Views:

Chile - C
Chile, 197
Created on
Views:

Semantic understanding



fraud

How could data discovery be relevant for IEG?

- May be able to provide a tool for more comprehensive review of reports.
- Discover related projects based on previous reports.
- Curate audit materials based on project topics.

Systematic Country Diagnostics

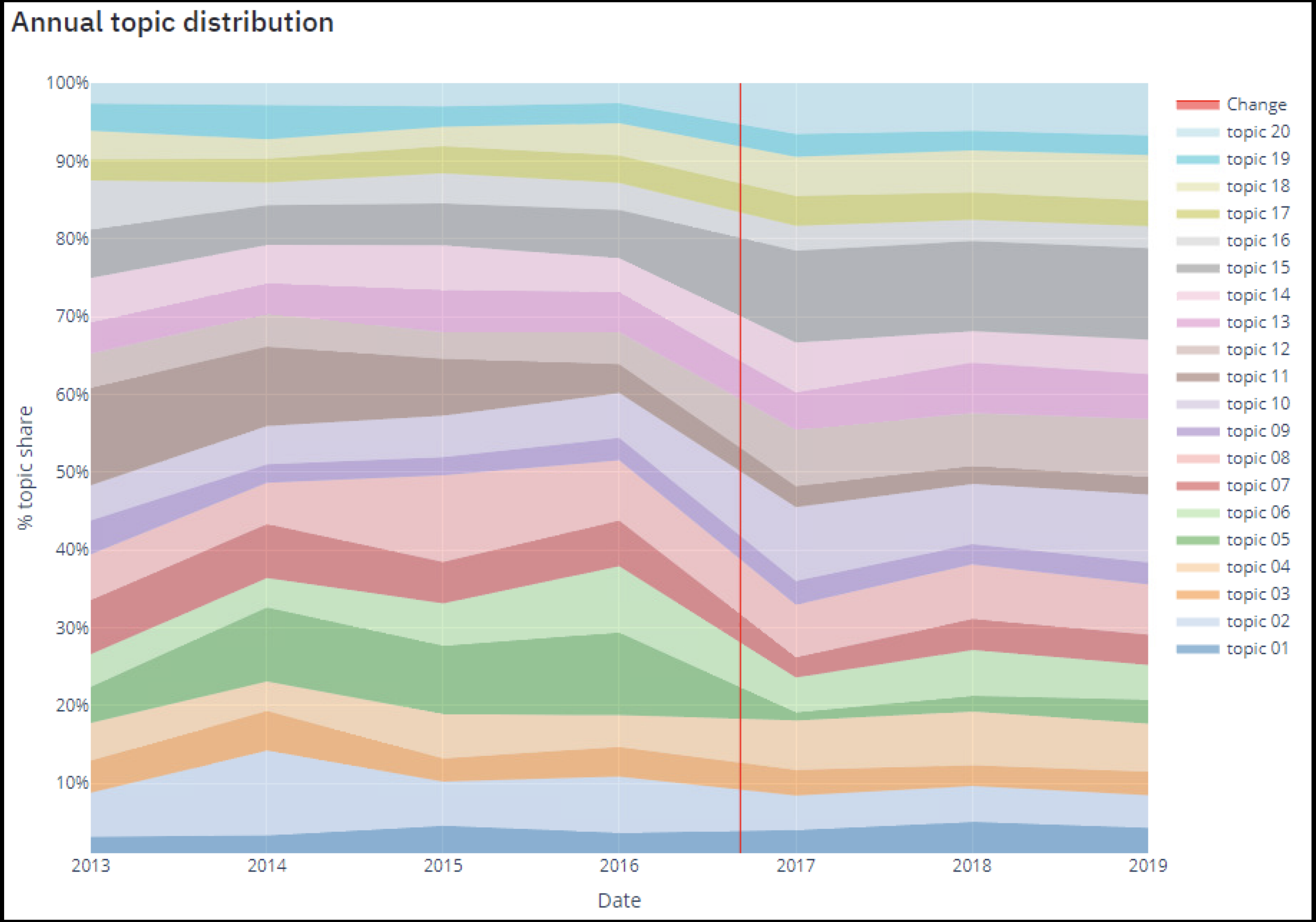
Systematic Country Diagnostics

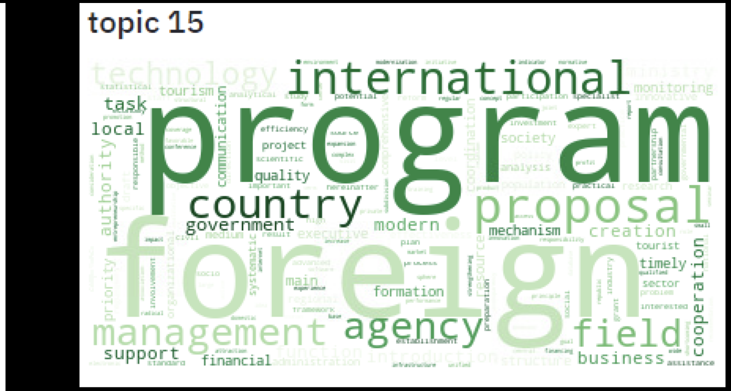
- Use topic modeling to understand changes in legislative agenda across change in government.

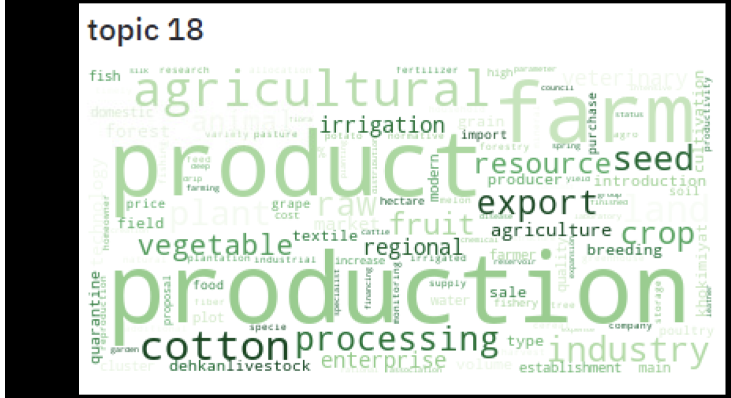
- Analyze the general sentiment in the country from global news data.

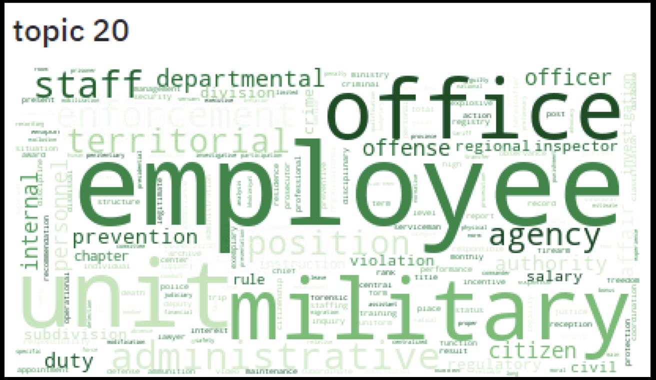
- Demonstrate the feasibility of proxy methods to understand certain country-level information using NLP.

Topic evolution and changes in legislation priorities

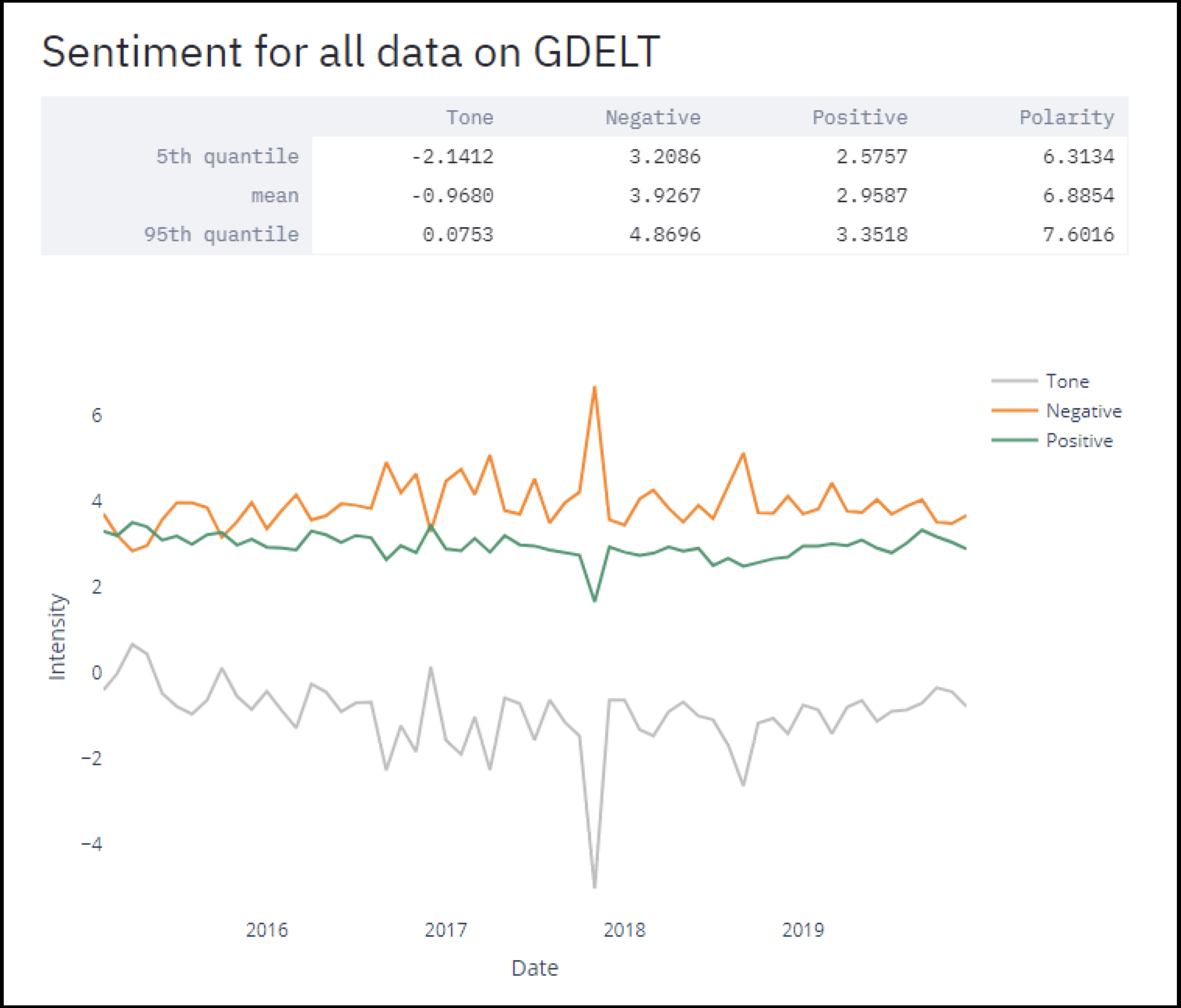

Annual topic distribution


topic 15


topic 18


topic 20

Capturing sentiment from news



Sentiment for all data on GDELT

	Tone	Negative	Positive	Polarity
5th quantile	-2.1412	3.2086	2.5757	6.3134
mean	-0.9680	3.9267	2.9587	6.8854
95th quantile	0.0753	4.8696	3.3518	7.6016

How might IEG benefit from sentiment and topic models?

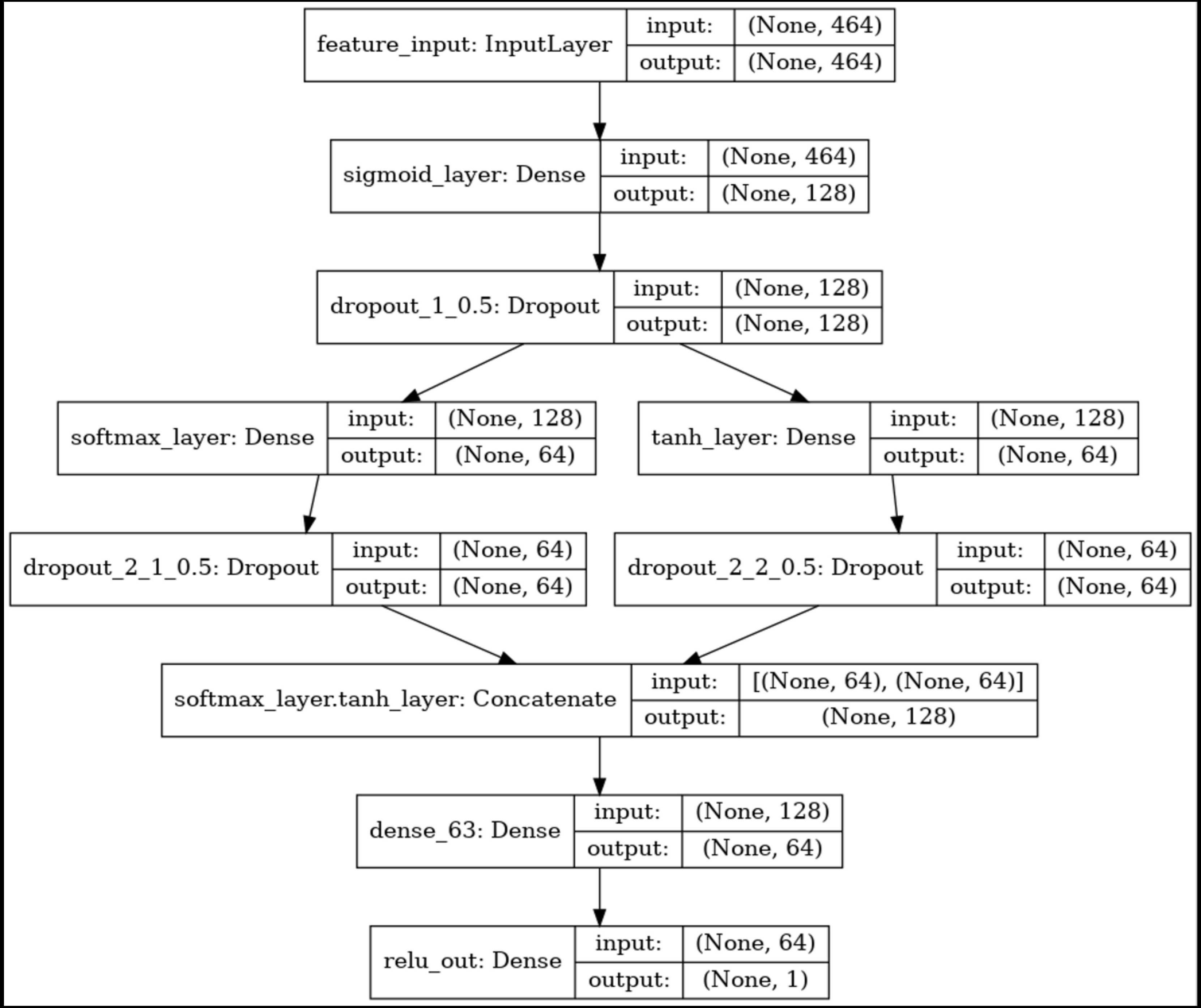
- Monitor impact of projects based on social media sentiment.
- Improve the management of projects by using topic models.
- Understand trends between successful and unsuccessful projects.

Climate change co-benefits prediction

Climate change co-benefits prediction

- Process and perform text mining on project documents and enrich the available data using other sources.
- Use pretrained NLP models and machine learning to predict climate change co-benefits.
- The project was exploratory in nature. A baseline exists and our model has outperformed this baseline.

Neural network model



feature_input: InputLayer | input: (None, 464) | output: (None, 464)

sigmoid_layer: Dense | input: (None, 464) | output: (None, 128)

dropout_1_0.5: Dropout | input: (None, 128) | output: (None, 128)

softmax_layer: Dense | input: (None, 128) | output: (None, 64)

tanh_layer: Dense | input: (None, 128) | output: (None, 64)

dropout_2_1_0.5: Dropout | input: (None, 64) | output: (None, 64)

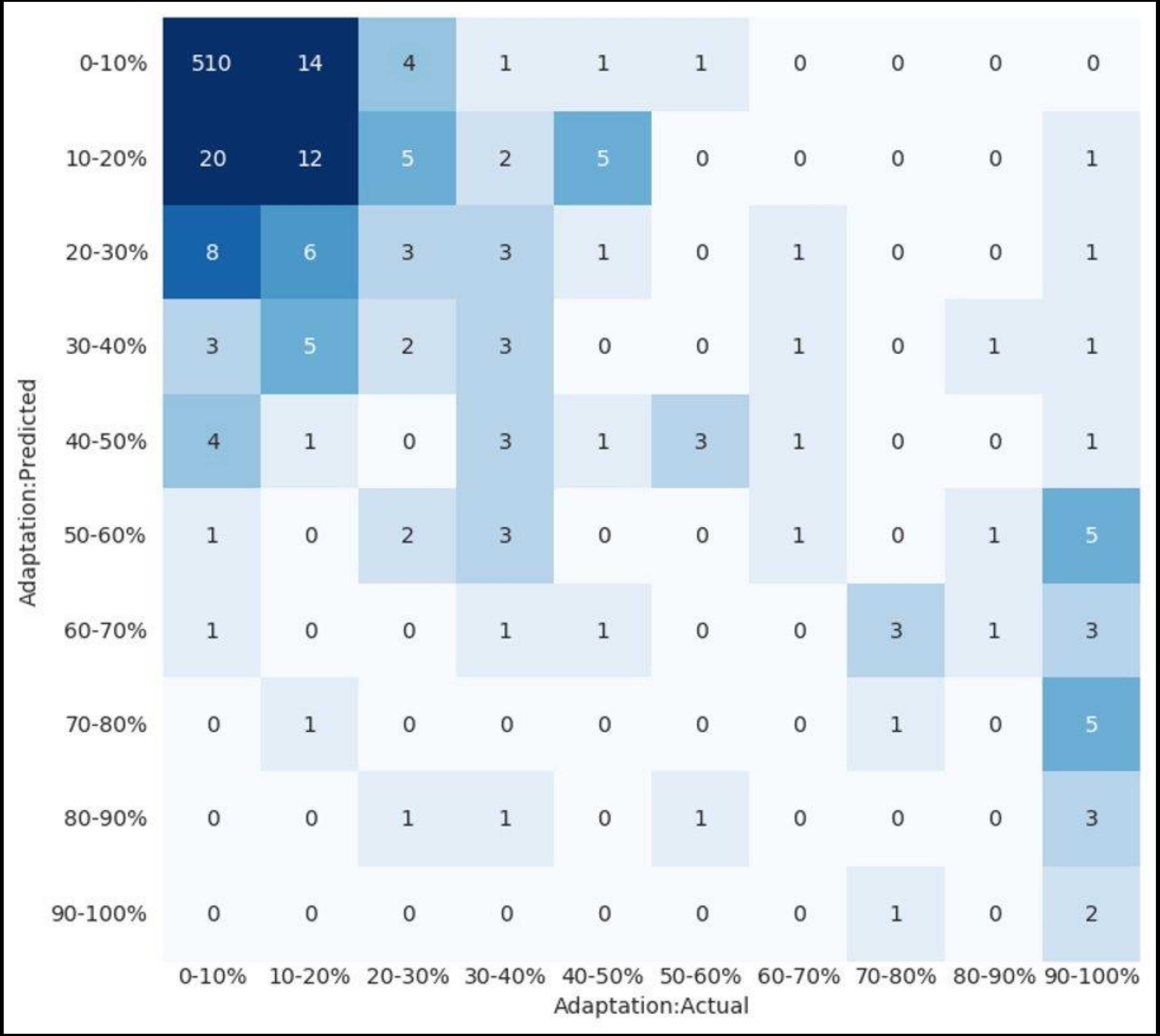
dropout_2_2_0.5: Dropout | input: (None, 64) | output: (None, 64)

softmax_layer.tanh_layer: Concatenate | input: [(None, 64), (None, 64)] | output: (None, 128)

dense_63: Dense | input: (None, 128) | output: (None, 64)

relu_out: Dense | input: (None, 64) | output: (None, 1)

Predictive performance



	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
0-10%	510	14	4	1	1	1	0	0	0	0
10-20%	20	12	5	2	5	0	0	0	0	1
20-30%	8	6	3	3	1	0	1	0	0	1
30-40%	3	5	2	3	0	0	1	0	1	1
40-50%	4	1	0	3	1	3	1	0	0	1
50-60%	1	0	2	3	0	0	1	0	1	5
60-70%	1	0	0	1	1	0	0	3	1	3
70-80%	0	1	0	0	0	0	0	1	0	5
80-90%	0	0	1	1	0	1	0	0	0	3
90-100%	0	0	0	0	0	0	0	1	0	2

Climate change co-benefits prediction

- While we have shown feasibility of improved performance, the degree of accuracy was still not at the desired level for automation.

- The primary reason for poor performance was the limited volume of available data.

How machine learning with NLP could be beneficial in IEG?

- Automate data extraction from text.
- Detect fraud or potential issues in projects given historical data.

POLL: Which NLP application will be immediately useful for IEG?

1. May be able to provide a tool for more comprehensive review of reports.
2. Discover related projects based on previous reports.
3. Curate audit materials based on project topics.
4. Monitor impact of projects based on social media sentiment.
5. Improve the management of projects by using topic models.
6. Understand trends between successful and unsuccessful projects.
7. Automate data extraction from text.
8. Detect fraud or potential issues in projects given historical data.

Natural Language Processing

Review of applications and opportunities for IEG

Aivin V. Solatorio

Data Scientist | DECAT

asolatorio@worldbank.org

@avsolatorio