AI and NLP for UN DPO Data

Dr. Margaret J. Foster
University of North Carolina, Chapel Hill
margaret.foster@unc.edu

Prepared for the United Nations
Department of Peace Operations (DPO)
Fall 2022

Overview

Part I: Technical Overview

Part II: Promises and Pitfalls in the A4P(+) Context

Part III: Techniques for Analysis and Presentation

Part IV: Overview of sources for additional data

Technical Overview: Terminology

Machine Learning:
Formally: Computer systems that can learn and adapt without explicit instructions
Informally: Pattern-finding

Artificial Intelligence (AI):
Computers implement utility functions to make decisions

Big Data:
Data arriving in great velocity and with great volume; often in complex forms.

Technical Overview: What is NLP/Text-as-Data?

Natural Language Processing (NLP):
Using linguistics, computer science, and AI to allow computer programs to understand human language as spoken and written.

Turns *texts* (and images!) into *data*

How?

Opportunities: Seeing the Bigger Picture Faster

- ML models can **summarize themes at scale**
- Good for systematic perspective on broader context
- Helps quantify trends and covariate relationships
- Efficiently access historical information
- Identifies new content embedded in reports

Opportunities: Example

Possible Use Cases: automatically code event participants over time

Report of the Secretary-General on the situation in Mali (9/2014)

4. On 15 and 16 June, the Ministers for Foreign Affairs and Ambassadors of Algeria, Burkina Faso, Chad, Mali, Mauritania and the Niger met in Algiers, affording my Special Representative the opportunity to reiterate the call of the Council for proper coordination among international actors in support of the Malian peace process.

Color key:
Whom
From where

Opportunities: Example II

Possible Use Cases: extract conflict events

Report of the Secretary-General on the situation in Mali (9/2014)

19. The breaches of the ceasefire increased security risks for civilians and resulted in human rights violations, in breach of article 10 of the preliminary agreement. After MNLA took the town of Aguelhok on 21 May, 66 Imghad Tuareg civilian men, fearing for their security, sought refuge in the MINUSMA camp. They stayed there until 3 July, when MINUSMA negotiated their safe return with MNLA and HCUA. The clashes that took place from 11 to 26 July in Anefis and Tabankort resulted in the killing of at least 4 civilians and the forced displacement of 56 women and 72 children. On 18 August, six men were severely beaten by MNLA in Ménaka (Gao region) for wearing t-shirts bearing the colours of the Malian flag. In Lerneb, MAA (Coordination) has curtailed the movements of members of the Arab community suspected of being MAA (Platform) sympathizers.

Color key:
When (start, end);
What happened
To Whom

Challenges

Challenge: NLP/Text-as-Data model *content*

…so users must consider omissions, biases, and frames

Report of the Secretary-General on the situation in Mali (12/2014)

16. On 6 November, the Malian defence and security forces took control of the southern bank of the Niger River in Didi, east of Timbuktu, while MAA (Coordination) and MNLA remained in control of the northern bank. In mid-November, the Coordination took control of Zarho (100 km east of Didi), and on 1 December the Platform seized Bamba (30 km east of Zarho). There was no significant violence recorded during those movements.

17. Extremist groups were suspected of killing 16 peacekeepers and injuring 14 others during the reporting period. Improvised explosive devices and anti-vehicle mines placed along routes used by MINUSMA severely hindered its operations. On

Challenges II

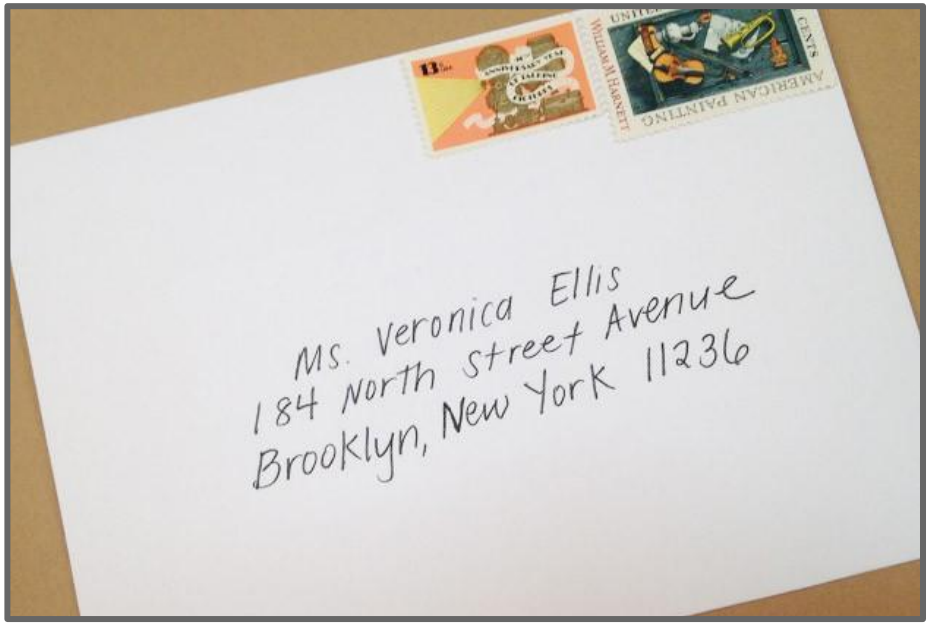
A non-exhaustive list of possible sources of bias:
Omission
Frame: Positivity/negativity biases
Missing-not-at-random
Noise
Unit of analysis mismatch
Algorithmic biases
Model vs data generating process

Challenges III

Illustration: missingness/noise not at random

Toy example: a wedding invitation for one

<u>Takeaway</u>
Understanding what you see/don't see requires domain knowledge



Challenges: Conclusion

ML/AI are math
Math is a model

Overcoming the Challenges

Strategies:

- Evaluate with expertise
- Triangulate with additional data inputs
- Treat ML outputs as *tools* and not *answers*

ML as a Tool: Thematic Summaries

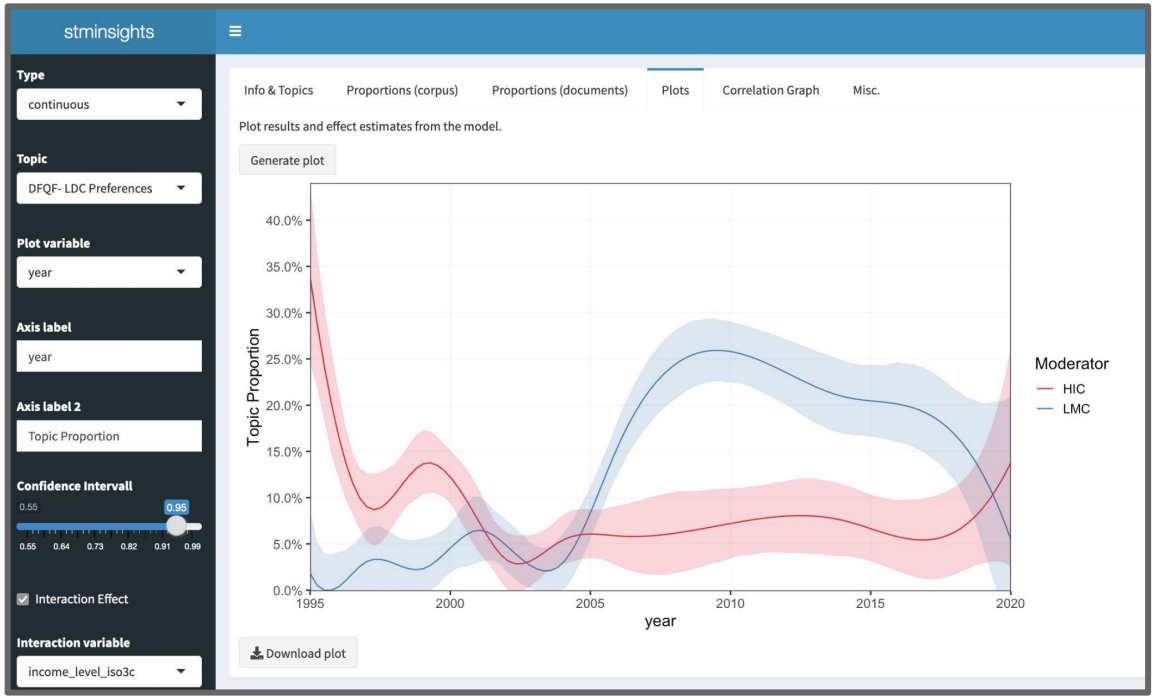
ML tools can identify and track themes

Model:
Structural Topic Model

Data:
Committee on Trade and Development
World Trade Organization

Visualization:
STM Insights

Prevalence of “LDC preferences” theme



ML As a Tool: Representative Passages

ML tools can help sift large quantities of data

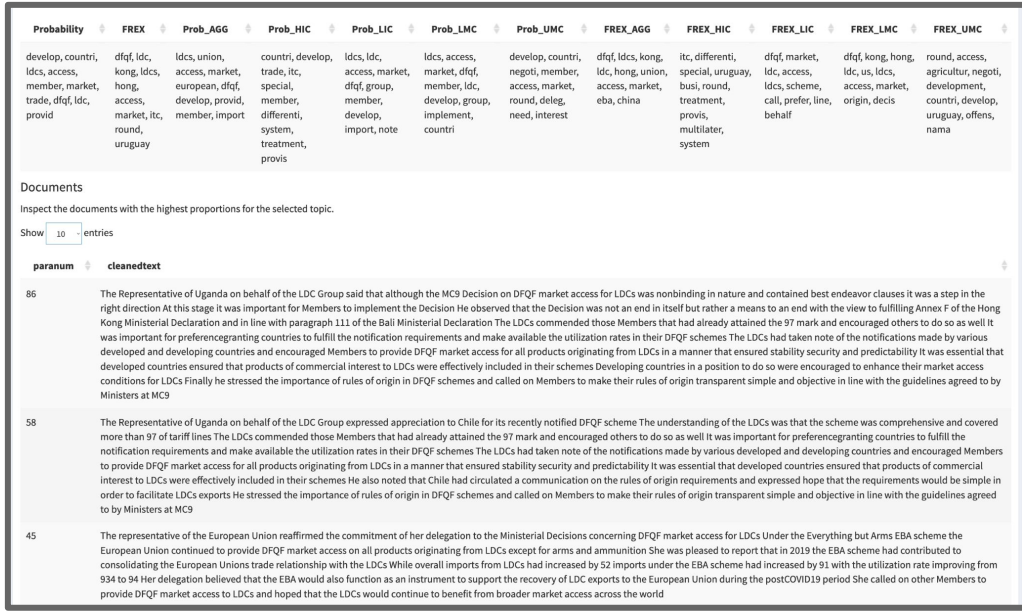
Task: take 8,500+ paragraphs, find 100 most strongly related to identified theme

<u>Model:</u>
Structural Topic Model

<u>Data</u>:
Committee on Trade and Development
World Trade Organization

<u>Visualization</u>:
STM Insights

Most representative paragraphs, LDC prefs.



Probability	FREX	Prob_AGG	Prob_HIC	Prob_LIC	Prob_LMC	Prob_UMC	FREX_AGG	FREX_HIC	FREX_LIC	FREX_LMC	FREX_UMC
develop, countri, ldcs, access, member, market, trade, dfqf, ldc, provid	dfqf, ldc, kong, ldcs, hong, access, market, itc, round, uruguay	ldcs, union, access, market, european, dfqf, develop, provid, member, import	countri, develop, trade, itc, special, member, differenti, system, treatment, provis	ldcs, ldc, access, market, dfqf, group, member, develop, import, note	ldcs, access, market, dfqf, member, ldc, develop, group, implement, countri	develop, countri, negoti, member, access, market, round, deleg, need, interest	dfqf, ldcs, kong, ldc, hong, union, access, market, eba, china	itc, differenti, special, uruguay, busi, round, treatment, provis, multilater, system	dfqf, market, ldc, access, ldcs, scheme, call, prefer, line, behalf	dfqf, kong, hong, ldc, us, ldcs, access, market, origin, decis	round, access, agricultur, negoti, development, countri, develop, uruguay, offens, nama

Documents

Inspect the documents with the highest proportions for the selected topic.

Show 10 entries

paranum	cleanedtext
86	The Representative of Uganda on behalf of the LDC Group said that although the MC9 Decision on DFQF market access for LDCs was nonbinding in nature and contained best endeavor clauses it was a step in the right direction At this stage it was important for Members to implement the Decision He observed that the Decision was not an end in itself but rather a means to an end with the view to fulfilling Annex F of the Hong Kong Ministerial Declaration and in line with paragraph 111 of the Bali Ministerial Declaration The LDCs commended those Members that had already attained the 97 mark and encouraged others to do so as well It was important for preferencegranting countries to fulfill the notification requirements and make available the utilization rates in their DFQF schemes The LDCs had taken note of the notifications made by various developed and developing countries and encouraged Members to provide DFQF market access for all products originating from LDCs in a manner that ensured stability security and predictability It was essential that developed countries ensured that products of commercial interest to LDCs were effectively included in their schemes Developing countries in a position to do so were encouraged to enhance their market access conditions for LDCs Finally he stressed the importance of rules of origin in DFQF schemes and called on Members to make their rules of origin transparent simple and objective in line with the guidelines agreed to by Ministers at MC9
58	The Representative of Uganda on behalf of the LDC Group expressed appreciation to Chile for its recently notified DFQF scheme The understanding of the LDCs was that the scheme was comprehensive and covered more than 97 of tariff lines The LDCs commended those Members that had already attained the 97 mark and encouraged others to do so as well It was important for preferencegranting countries to fulfill the notification requirements and make available the utilization rates in their DFQF schemes The LDCs had taken note of the notifications made by various developed and developing countries and encouraged Members to provide DFQF market access for all products originating from LDCs in a manner that ensured stability security and predictability It was essential that developed countries ensured that products of commercial interest to LDCs were effectively included in their schemes He also noted that Chile had circulated a communication on the rules of origin requirements and expressed hope that the requirements would be simple in order to facilitate LDCs exports He stressed the importance of rules of origin in DFQF schemes and called on Members to make their rules of origin transparent simple and objective in line with the guidelines agreed to by Ministers at MC9
45	The representative of the European Union reaffirmed the commitment of her delegation to the Ministerial Decisions concerning DFQF market access for LDCs Under the Everything but Arms EBA scheme the European Union continued to provide DFQF market access on all products originating from LDCs except for arms and ammunition She was pleased to report that in 2019 the EBA scheme had contributed to consolidating the European Unions trade relationship with the LDCs While overall imports from LDCs had increased by 52 imports under the EBA scheme had increased by 91 with the utilization rate improving from 934 to 94 Her delegation believed that the EBA would also function as an instrument to support the recovery of LDC exports to the European Union during the postCOVID19 period She called on other Members to provide DFQF market access to LDCs and hoped that the LDCs would continue to benefit from broader market access across the world

ML as a Tool: Sentiment in Different Settings

ML can be used to summarize sentiment(*)

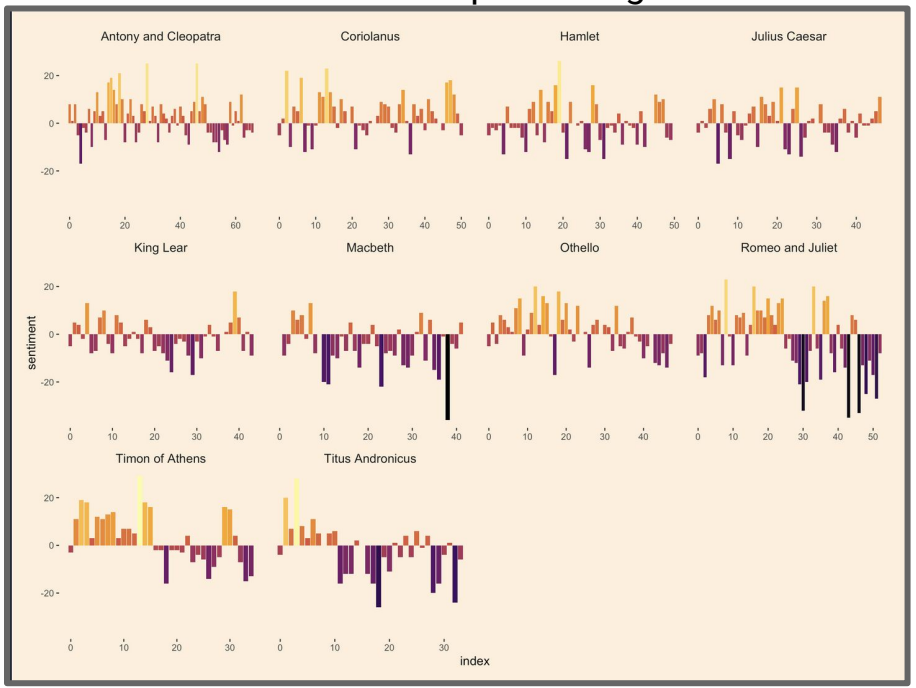
Results grouped as desired
(eg: country, year, mission, participants)

<u>Model</u>:
Sentiment analysis

<u>Data:</u>
Shakespeare tragedies

<u>Visualization</u>:
Tidytext

Sentiment in Shakespeare Tragedies



Source: https://peerchristensen.netlify.app/post/fair-is-foul-and-foul-is-fair-a-tidytext-entiment-analysis-of-shakespeare-s-tragedies/

Other ML Tools for Texts

Some other useful tools

- Sentiment /Frequency analysis (https://github.com/juliasilge/tidytext)
- Keyword-assisted topic model (https://keyatm.github.io/keyATM/)
- Word and language embedding (https://huggingface.co/docs/transformers/model_doc/bert)
- Text networks (https://cbail.github.io/textasdata/text-networks/rmarkdown/Text_Networks.html)

Additional Data Sources for UN/A4P(+): Text

The *UNSC Debates* Corpus (Schoenfeld et al. 2021)
What: 82,000+ speeches from 5,748 UNSC Meetings (1995-2020)
Codebook + data available via Harvard Dataverse:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KGVSYH

***UN General Debate* Corpus (Baturo, Dasandi, and Mikhaylov 2017)**
What: 7300+ country statements at UNGA (1970-2014)
Visualization tool: http://ungd.smikhaylov.net/
Data available via Harvard Dataverse:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y

***PeaceKeeping Operations* Corpus (Amicarelli and Di Salvatore 2021)**
1,455 reports covering 68 UN Peacekeeping Missions (1994-2020)
Data is available via: https://www.prio.org/journals/jpr/replicationdata

Sources compiled by Evgenija Kroeker

Additional Data Sources for A4P(+): Event Data I

Peacekeeping Specific:

International Peace Institute (IPI) Datal; monthly mission personnel summaries

Peacekeeping Mandates (PEMA) Dataset; UNSC resolutions

Geocoded Peacekeeping Operations (Geo-PKO) Dataset; Mission deployment maps; UNSG mission progress reports

UCDP Peacemakers at Risk (PAR) Dataset; Factiva (News archive database); UN, NGO and open-source reports

UN Peace Initiatives (UNPI) Dataset; Repertoire of the Practices of the Security Council, UNGA Yearly Reports

Sources compiled by: Evgenija Kroeker, Maline Meiske & Andrea Ruggeri

Additional Data Sources for A4P(+): Event Data II

General Conflict Events

Armed Conflict Location & Event Data Project (ACLED)

Uppsala Conflict Data Program (UCDP)

Social conflict analysis dataset (SCAD)

Global terrorism database (GTD)

Sexual Violence in Armed Conflict (SVAC) Dataset

Varieties of Democracy (V-DEM)

Sources compiled by: Evgenija Kroeker, Maline Meiske & Andrea Ruggeri

Questions?

Questions? Comments?

Contact: margaret.foster@unc.edu