Lessons from the Words of Peace project

Larry S. Liebovitch

https://sites.google.com/view/Larry-phd

Adjunct Senior Research Scholar, Advanced Consortium on Cooperation, Conflict and Complexity, at the Earth Institute, Columbia University

Professor Emeritus, Physics, Queens College, City University of New York

More about this work is at:
https://arxiv.org/abs/2305.12537

Lessons from the Words of Peace project

Larry S. Liebovitch

https://sites.google.com/view/Larry-phd

BS Physics City College, City University NY	**PhD Astronomy** Harvard University	**Asst. Professor** Columbia University College of Physicians & Surgeons
Professor Psychology Interim Director Center for Complex Systems & Brain Sciences Florida Atlantic University	**Professor of Physics & Psychology** Dean of the Division of Mathematics & Natural Sciences Queens College, City University NY	**Adjunct Senior Research Scholar** AC4, Climate School, Columbia University

Previous Work

Physical	Biological	Social
Nonlinear Oscillators (autoresonance)	**Gene Regulatory Networks**	Models of Sustainable Peace
Dynamics of Coupled Maps	Predicting Drug Interactions	**Emotions in Psychotherapy**
Kinetics of Protein Ion Channels	**Timing of Heart Arrhythmias**	Dynamics on Small World Networks
Motions in Proteins	Flow of Ions and Water through Cells & Tissues	Analysis of Difficult Conversations
Motions of Stars & Gas in Galaxies	Why Eyes are Round	Models of Conflicts
Fractals	Error Correcting Codes in DNA	**Commodity Prices in Ancient Babylon**
Chaos Theory	Anoxia in the Turtle Brain	Distribution of Artifacts at Archeological Sites

Current Work
computer science, data science

Sustaining Peace **natural language processing, machine learning**
-2023. L. S. Liebovitch et al. Machine learning to determine the word differences in media in lower and higher peace countries and a quantitative peace index.
https://arxiv.org/abs/2305.12537

Flow of Information in Networks **numerical dynamical simulations**
-2022. E. Jacobo-Villegas et al. Conflict Dynamics in Scale-Free Networks with Degree Correlations and Hierarchical Structure.
-2015. I. Y. Fernandez-Rosale et al. The dynamic consequences of cooperation and competition in small-world networks.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0126234
https://www.mdpi.com/1099-4300/24/11/1571

When Neural Networks Produce "hallucinations" **dynamics of neural networks**
-2023. W. Powers, et al. Tools and Visualizations for Exploring Classification Landscapes. IEEE 57th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA
https://ieeexplore.ieee.org/document/10089673

Goals of the Sustaining Peace Project

- **Positive Peace**
  identify the social process that support and sustain peace

- **Peace Speech**
  use machine learning to find the words that reflect/drive peace

- **Create New Measures of Peace**
  from machine learning models of media and social media

- **Real Time Dashboard**
  assessment tools for policy makers, academics, journalists

Positive Peace

Most previous peace research studies have analyzed peace only in a negative way, as the absence of conflict or violence.

A growing effort is to understand "positive peace", the social systems that generate and sustain peaceful societies.

Douglas P. Fry. 2005. The Human Potential for Peace: An Anthropological Challenge to Assumptions about War and Violence. Oxford University Press, Oxford UK.

Peter T. Coleman and Morton Deutsch. 2012. Psychological Components of Sustainable Peace (Peace Psychology Book Series). Springer, New York, NY.

Paul F. Diehl. 2016. Exploring peace: Looking beyond war and negative peace. International Studies Quarterly 60 (2016), 1–10. https://doi.org/10.1093/isq/sqw00

Gary Goertz, Paul F. Diehl, and Alexandru Balas. 2016.The Puzzle of Peace: The Evolution of Peace in the International System. Oxford University Press, Oxford UK

Youssef Mahmoud and Anupah Makoond. 2017. Sustaining peace: What does it mean in practice? International Peace Institute (April 2017), 1–5.

Examples of Positive Peace

Iroquois Great League of Peace (1450-1750)

Myths: war hatchets buried and washed away by an underground river
Mores: reinforced at an annual peace ceremony
Conflict Management: at the village, tribal, and confederacy levels
Women: authority to remove chiefs

Contemporary Peace Systems

Fry. 2012. Science 336:879-884.

Location	***Peace System***
Australia	Peoples of the Great Western Desert (the Mardudjara/Mardu, Gugadji, Walmadjeri, and Pintupi)
Canada	Montagnais-Naskapi and East Main Cree of the Labrador Peninsula
India	Nilgiri Plateau/Hills societies (Toda, Kota, Badaga, and Kurumbas)
Malaysia	Central Peninsular Orang Asli societies (e.g., Batek, Jahai, Semai, Chewong, and Btsisi)
Greenland	Native Inuit populations Iroquois Great League of Peace
US and Canada	Iroquois Great League of Peace (Cayuga, Mohawk, Oneida, Onondaga, Seneca, and Tuscarora)
Brazil	Ten tribes of the Upper Xingu River basin
Europe	European Union (27 member countries and growing)

Traditional Science

Top-Down Approach: start with ideas

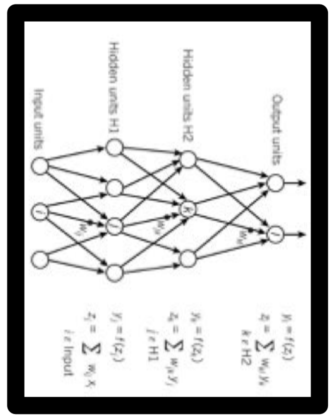
hypothesis → equations → analytics/code → solution

F=ma → $f = m\dfrac{d^2x}{dt^2}$ → $\begin{array}{l} \mathbf{x}(n+1) = \mathbf{x}(n) + \mathbf{v}(n)dt + (1/2)\mathbf{a}(n)dt^2 \\ \mathbf{v}(n+1) = \mathbf{v}(n) + (1/2)[\mathbf{a}(n+1) + \mathbf{a}(n)]dt \end{array}$ → **x(t)**

New Data Science

Bottom-Up Approach start with data

Supervised ML model
Train sample (x_{in}, y_{out})
(x_{in}, y_{out}), (x_{in}, y_{out}), (x_{in}, y_{out}) . . →

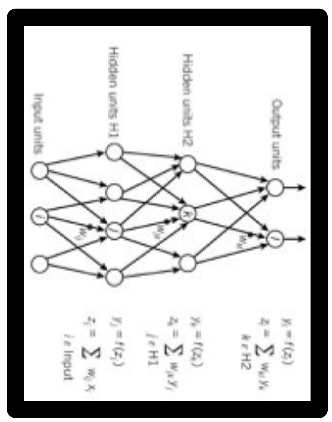


Adjustable Parameters (w_k)

Find (w_k) best fit training (x_{in}, y_{out})

Given a new
(x_{in}) →

we don't solve the problem, we create a model that gives good answers, trained from previous examples.



Predict a new
(y_{out})

hope model generalizes to get a good (y_{out})

Methods

Traditional Science last 200 years

top down: theory to data

Conjecture what words are different in conflicts and peace, and then search for the frequency of those words in the data.

New Data Science now

bottom up: data to theory

Use natural language processing and machine learning to find the words that best predict if a country is lower-peace or higher-peace.

Advanced Consortium on Cooperation, Conflict, and Complexity
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Summary of Results

Sustaining Peace Project
Machine Learning

2020

- Liebovitch, Powers, Shi, Chen-Carrel, Loustaunau, Coleman
- NOW Dataset

Effective predictions of lower-peace and higher-peace countries from their language

Lower-peace: government /control

Higher-peace: daily activities/diversity of words

- Capstone 2020
- NOW Dataset

Effective predictions of lower-peace and higher-peace countries from their language

Unable to identify word differences between lower-peace and higher-peace countries

2021

- CAPSTONE 2021
- LexisNexis Dataset

Effective predictions of lower-peace and higher-peace countries from their language

Weak similarities of language within lower-peace and higher-peace countries

Fall 2022

- Capstone 2022
- Scraped Twitter & Local Newspapers

Lower-peace: government /control

Higher-peace: daily activities/diversity of words

Less diversity in words in lower-peace than in higher-peace countries

2023

- Guzman, Liebovitch
- NOW Dataset

Less diversity in words in lower-peace than in higher-peace countries

in progress

- Liebovitch, Mannis, Stone, Zadrozna
- LexisNexis Dataset

in progress

Data Sources

NOW

News On the Web: https://www.english-corpora.org/now/
18 countries, 723,574 media articles, January 2010 – September 2020
from newspapers, magazines, technical journals, and media broadcast stations
e.g. AlterNet, **Austin American-Statesman**, Business Insider, Business Wire (press release), Chicago Tribune, **FOX43.com (PA)**, Jerusalem Post, Israel News, **KCCI Des Moines (IA)**, Kentwired.

Lexis-Nexis

through Elsevier-Columbia University partnership
20 countries, 2,000,000 media articles, 2010-2020
e.g. 24 Hours Toronto, AARP The Magazine, **ABC News** Transcripts, Accounting Today, **AutoWeek**, BBC Monitoring: International Reports, Bollywood Country, **Business of Fashion**, FDA Week, Hair, Internet World, Journal of European Studies, Marie Claire, News Bites, Pittsburgh Tribune Review, Pizza Marketplace, Gas Week, St. Louis Post-Dispatch (Missouri), Tech News, The Nation, World Oil.

Local Newspapers, Twitter

media: 16 countries, 600,000 articles
e.g. The Star (Canada), The Independent (UK), **News in English (Norway),** Sunday World (Ireland), Daily Finland, The Straits Times (Singapore), 9news (Australia), France24, **Times of India**, The Independent (Uganda), Tehran Times, **The Nigerian Voice**, Bulawayo24 (Zimbabwe), The Point (Gambia), The Libya Observer.
twitter: 16 countries, 800,000 tweets

Lessons in Machine Learning
preparing the data

1. **Pre-Processing:**

    remove junk: @@@@, <p>, “”
    remove “named entities” (proper nouns)
        ottawa, alberta, kenya, australia
    remove “stop words” (common words)
        the, of, be, in
    lemmatize
        work=works=worked=worklife

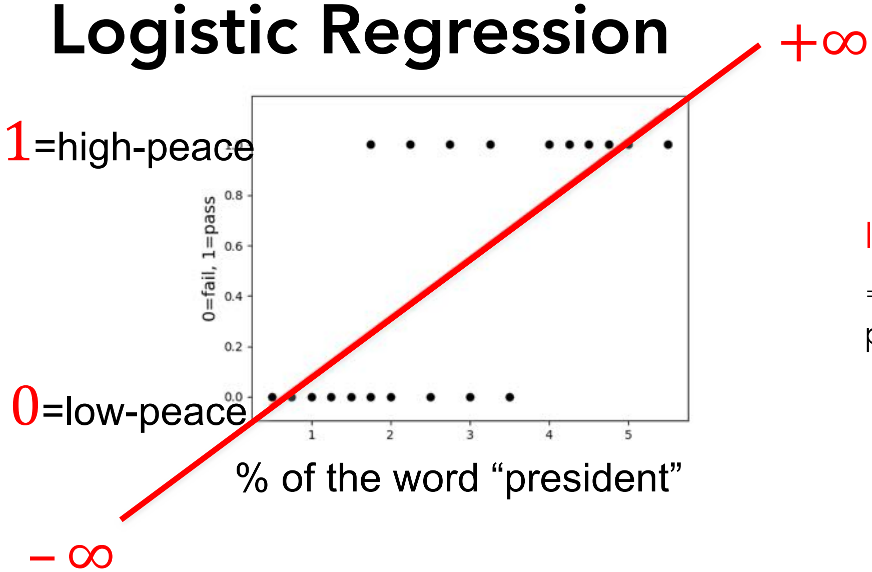
2. **James W Pennebaker**

    function words: can reveal social status, lying
    I, you, they
    to, of, for
    is, am, have

   TEDx talk: The Secret Life of Pronouns
   https://www.youtube.com/watch?v=PGsQwAu3PzU

Machine Learning Models

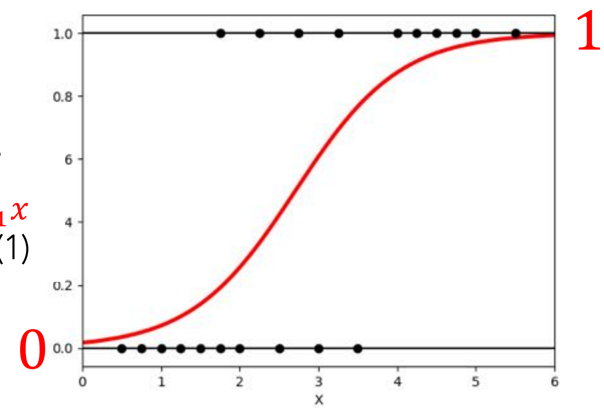
Logistic Regression



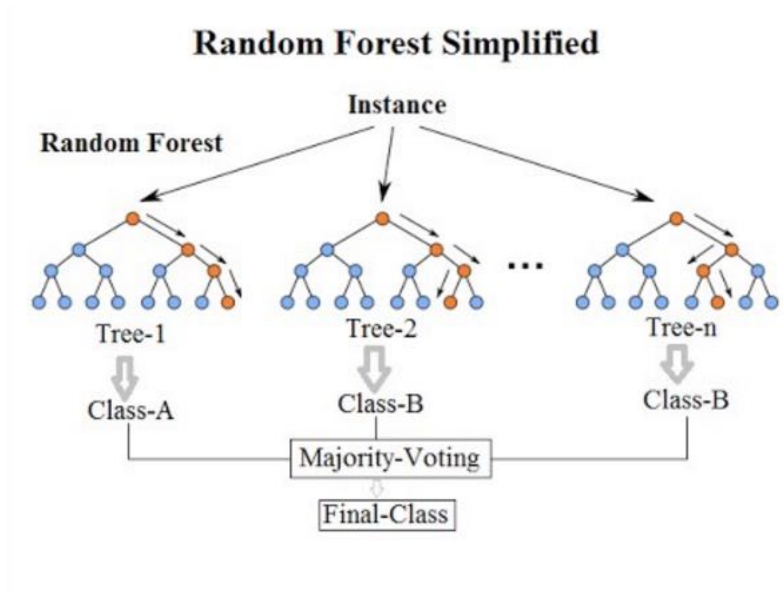
logit function $= log_e \frac{p(x)}{1-p(x)}$

= log(odds ratio) = $\beta_0 + \beta_1 x$

p(x) = probability of class (1)



Random Forest



tree = decision graph
for a subset of data
repeat with many different subsets
many trees = random forest

Other

ML: XGBoost, SVM, neural networks, K-means; NLP: NLTK, SpaCy, Word2Vec, BERT, XLNet

Data

From each country

computed the normalized word frequency in that country
most common 300 words

Combined data from countries

767 unique words across all 18 countries

Dataframe

18 rows (countries)
767 columns (word frequencies)

Models

3-class: low, intermediate, high peace countries, 18 x 767
2-class: low, high peace countries, 10 x 767

Supervised Learning: Train/Test

3-class: **80/20:** train/text split
cross-validation: 17 to predict 1, repeat for 18 separate trainings
uses the data more efficient, longer computation time
2-class: **cross-validation:** 9 to predict 1, repeat for 10 separate trainings
uses the data more efficient, longer computation time

Training Set

No agreement on Peace Indices!

	Country	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	Australia	1.41	1.53	7.3	24.59	0.93
2	Bangladesh	2.11	3.62	4.66	91.76	0.59
3	Canada	1.37	1.5	7.4	24.93	0.91
4	Ghana	1.8	2.95	4.74	68.88	0.58
5	Hong Kong			5.48		0.93
6	India	2.59	3.26	4.48	77.84	0.62
7	Ireland	1.43	1.38	7.01	23.58	0.93
8	Jamaica	2.12	2.53	5.55	64.99	0.72
9	Kenya	2.38	3.56	4.43	97.93	0.56
10	Malaysia	1.59	2.54	5.85	66.02	0.79
11	New Zealand	1.25	1.48	7.28	22.86	0.91
12	Nigeria	2.8	3.87	5.17	100.8	0.52
13	Philippines	2.5	3.28	5.24	84.73	0.7
14	Singapore	1.42	1.67	6.54	33.37	0.93
15	Sri Lanka	2.24	3.18	4.3	90.04	0.77
16	Tanzania	1.81	3.4	3.58	80.72	0.51
17	United Kingdom	1.84	1.55	6.9	34.08	0.92
18	United States	2.29	1.79	7.03	35.45	0.92

GPI=Global Peace Index
PPI=Positive Peace Index
WHI=World Happiness Index
FSI=Fragile States Index
HDI=Human Development Index

Lessons from Machine Learning

Unify the Training Data

3. **Data from different sources DISAGREES!**

    EASY: normalize the data from different sources

    HARD:

    - each index uses a different scale
        - convert all to one scale
    - algorithm
        - 3 out of 5 peaceful -> higher-peace country
        - 3 out of 5 not peaceful -> lower-peace country

L. Liebovitch, W. Powers, L. Shi, A. Chen-Carrel, P. Loustaunau, P. Coleman

To train the machine learning model, we need examples or countries that are Low, Middle, High peace.

GPI=Global Peace Index
PPI=Positive Peace Index
WHI=World Happiness Index
FSI=Fragile States Index
HDI=Human Development Index

Original Values

	Country	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	Australia	1.41	1.53	7.3	24.59	0.93
2	Bangladesh	2.11	3.62	4.66	91.76	0.59
3	Canada	1.37	1.5	7.4	24.93	0.91
4	Ghana	1.8	2.95	4.74	68.88	0.58
5	Hong Kong			5.48		0.93
6	India	2.59	3.26	4.48	77.84	0.62
7	Ireland	1.43	1.38	7.01	23.58	0.93
8	Jamaica	2.12	2.53	5.55	64.99	0.72
9	Kenya	2.38	3.56	4.43	97.93	0.56
10	Malaysia	1.59	2.54	5.85	66.02	0.79
11	New Zealand	1.25	1.48	7.28	22.86	0.91
12	Nigeria	2.8	3.87	5.17	100.8	0.52
13	Philippines	2.5	3.28	5.24	84.73	0.7
14	Singapore	1.42	1.67	6.54	33.37	0.93
15	Sri Lanka	2.24	3.18	4.3	90.04	0.77
16	Tanzania	1.81	3.4	3.58	80.72	0.51
17	United Kingdom	1.84	1.55	6.9	34.08	0.92
18	United States	2.29	1.79	7.03	35.45	0.92

Scaled 0 to 100

	Country	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	Australia	89.68	93.98	97.38	97.78	100
2	Bangladesh	44.52	10.04	28.27	11.55	19.05
3	Canada	92.26	95.18	100	97.34	95.24
4	Ghana	64.52	36.95	30.37	40.92	16.67
5	Hong Kong			49.74		100
6	India	13.55	24.50	23.56	29.42	26.19
7	Ireland	88.39	100	89.79	99.08	100
8	Jamaica	43.87	53.82	51.57	45.92	50.00
9	Kenya	27.10	12.45	22.25	3.63	11.90
10	Malaysia	78.06	53.41	59.42	44.60	66.67
11	New Zealand	100	95.98	96.86	100	95.24
12	Nigeria	0.00	0.00	41.62	0.00	2.38
13	Philippines	19.35	23.69	43.46	20.58	45.24
14	Singapore	89.03	88.35	77.49	86.51	100.00
15	Sri Lanka	36.13	27.71	18.85	13.76	61.90
16	Tanzania	63.87	18.88	0.00	25.73	0.00
17	United Kingdom	61.94	93.17	86.91	85.60	97.62
18	United States	32.90	83.53	90.31	83.84	97.62

Low, Middle, High peace

	Country	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	**Australia**	89.68	93.98	97.38	97.78	100
2	**Bangladesh**	44.52	10.04	28.27	11.55	19.05
3	**Canada**	92.26	95.18	100	97.34	95.24
4	Ghana	64.52	36.95	30.37	40.92	16.67
5	Hong Kong			49.74		100
6	India	13.55	24.50	23.56	29.42	26.19
7	**Ireland**	88.39	100	89.79	99.08	100
8	Jamaica	43.87	53.82	51.57	45.92	50.00
9	**Kenya**	27.10	12.45	22.25	3.63	11.90
10	Malaysia	78.06	53.41	59.42	44.60	66.67
11	**New Zealand**	100	95.98	96.86	100	95.24
12	**Nigeria**	0.00	0.00	41.62	0.00	2.38
13	Philippines	19.35	23.69	43.46	20.58	45.24
14	**Singapore**	89.03	88.35	77.49	86.51	100
15	Sri Lanka	36.13	27.71	18.85	13.76	61.90
16	**Tanzania**	63.87	18.88	0.00	25.73	0.00
17	**United Kingdom**	61.94	93.17	86.91	85.60	97.62
18	United States	32.90	83.53	90.31	83.84	97.62

L. Liebovitch, W. Powers, L. Shi, A. Chen-Carrel, P. Loustaunau, P. Coleman

	Country	Number of Articles	Number of Words	Per Cent Articles	Per cent Words
1	**Bangladesh**	15,245	1,183,478	2.11	2.05
2	**Kenya**	30,694	1,940,412	4.24	3.36
3	**Nigeria**	52,895	5,307,297	7.31	9.18
4	**Tanzania**	6,164	500,702	0.85	0.87
	total	**104,998**	**8,931,889**	**14.51**	**15.45**
1	Ghana	22,783	1,699,258	3.15	2.94
2	Hong Kong	2,301	253,881	0.32	0.44
3	India	76,555	5,294,277	10.58	9.16
4	Jamaica	33,401	2,580,436	4.62	4.46
5	Malaysia	30,394	2,210,748	4.2	3.82
6	Philippines	61,474	4,166,281	8.5	7.21
7	Sri Lanka	11,329	983,349	1.57	1.7
8	United States	67,406	5,975,558	9.32	10.33
	total	**305,643**	**23,163,788**	**42.24**	**40.06**
1	**Australia**	62,683	5,599,285	8.66	9.68
2	**Canada**	73,869	6,981,358	10.21	12.07
3	**Ireland**	60,190	4,293,895	8.32	7.43
4	**New Zealand**	56,483	4,417,416	7.81	7.64
5	**Singapore**	20,195	1,345,811	2.79	2.33
6	**United Kingdom**	39,513	3,085,992	5.46	5.34
	total	**312,933**	**25,723,757**	**43.25**	**44.49**

724,000 articles
58,000,000 words

TRAINING
3-class model
low peace
intermediate peace
high peace countries
18 x 767

80/20 split
cross-validation
17 to predict 1
18 separate trainings

Lessons from Machine Learning

Unify the Training Data

4. **What if DIFFERENT amounts of data from each feature?**

   EASY

   normalize word counts within each country

   HARD:

   "bootstrap"

   differential comparisons of small sets to part of large sets

Machine Learning Results

3-class model: low, intermediate, high peace

	Accuracy (TP+TN)/(FP+FN+TP+TN)	**Precision TP/(TP+FP)**	**Recall TP/(FN+TP)**	**F1 2(Precision x Recall)/(Precision+Recall)**
3-class, Random Guessing	0.356 ± 0.033	0.368 ± 0.034	0.356 ± 0.033	0.354 ± 0.033
Random Forest				
3-class, 80/20 train/test	**0.525 ± 0.040**	**0.420 ± 0.056**	**0.525 ± 0.040**	**0.437 ± 0.048**
3-class, 17 to predict one	**0.567 ± 0.015**	**0.500 ± 0.018**	**0.567 ± 0.015**	**0.525 ± 0.015**
2-class, 9 to predict one	0.960 ± 0.013	0.965 ± 0.012	0.960 ± 0.013	0.960 ± 0.013
Logistic Regression				
3-class, 80/20 train/test	**0.388 ± 0.053**	**0.238 ± 0.057**	**0.388 ± 0.053**	**0.272 ± 0.050**
3-class, 17 to predict one	**0.611 ***	**0.520 ***	**0.611 ***	**0.558 ***
2-class, 9 to predict one	1.000 *	1.000 *	1.000 *	1.000 *
*Since the logistic regression converges to the same values on each run, sem = 0 for these values.				

Low, Middle, High peace

	Country	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	**Australia**	89.68	93.98	97.38	97.78	100
2	**Bangladesh**	44.52	10.04	28.27	11.55	19.05
3	**Canada**	92.26	95.18	100	97.34	95.24
4	Ghana	64.52	36.95	30.37	40.92	16.67
5	Hong Kong			49.74		100
6	India	13.55	24.50	23.56	29.42	26.19
7	**Ireland**	88.39	100	89.79	99.08	100
8	Jamaica	43.87	53.82	51.57	45.92	50.00
9	**Kenya**	27.10	12.45	22.25	3.63	11.90
10	Malaysia	78.06	53.41	59.42	44.60	66.67
11	**New Zealand**	100	95.98	96.86	100	95.24
12	**Nigeria**	0.00	0.00	41.62	0.00	2.38
13	Philippines	19.35	23.69	43.46	20.58	45.24
14	**Singapore**	89.03	88.35	77.49	86.51	100
15	Sri Lanka	36.13	27.71	18.85	13.76	61.90
16	**Tanzania**	63.87	18.88	0.00	25.73	0.00
17	**United Kingdom**	61.94	93.17	86.91	85.60	97.62
18	United States	32.90	83.53	90.31	83.84	97.62

GPI=Global Peace Index
PPI=Positive Peace Index
WHI=World Happiness Index
FSI=Fragile States Index
HDI=Human Development Index

If the peace indices of the HUMANS disagree with each other
HOW can our machine learning model figure it out!

But, the HUMANS are pretty sure which countries are lower-peace and higher-peace.

SO,
we train our ML model on the extreme countries

L. Liebovitch, W. Powers, L. Shi, A. Chen-Carrel, P. Loustaunau, P. Coleman

	Country	Number of Articles	Number of Words	Per Cent Articles	Per cent Words
1	**Bangladesh**	15,245	1,183,478	3.65	3.41
2	**Kenya**	30,694	1,940,412	7.34	5.6
3	**Nigeria**	52,895	5,307,297	12.66	15.31
4	**Tanzania**	6,164	500,702	1.47	1.44
	total	**104,998**	**8,931,889**	**25.12**	**25.77**
1	**Australia**	62,683	5,599,285	15	16.16
2	**Canada**	73,869	6,981,358	17.67	20.14
3	**Ireland**	60,190	4,293,895	14.4	12.39
4	**New Zealand**	56,483	4,417,416	13.51	12.75
5	**Singapore**	20,195	1,345,811	4.83	3.88
6	**United Kingdom**	39,513	3,085,992	9.45	8.9
	total	**312,933**	**25,723,757**	**74.88**	**74.23**

417,931 articles
34,655,646 words

TRAINING
2-class model
low peace
high peace countries
10 x 767

80/20 split
cross-validation
9 to predict 1
10 separate trainings

Machine Learning Results

2-class model: low, high peace

	Accuracy (TP+TN)/(FP+FN+TP+TN)	**Precision TP/(TP+FP)**	**Recall TP/(FN+TP)**	**F1 2(Precision x Recall)/(Precision+Recall)**
3-class, Random Guessing	0.356 ± 0.033	0.368 ± 0.034	0.356 ± 0.033	0.354 ± 0.033
Random Forest				
3-class, 80/20 train/test	0.525 ± 0.040	0.420 ± 0.056	0.525 ± 0.040	0.437 ± 0.048
3-class, 17 to predict one	0.567 ± 0.015	0.500 ± 0.018	0.567 ± 0.015	0.525 ± 0.015
2-class, 9 to predict one	**0.960 ± 0.013**	**0.965 ± 0.012**	**0.960 ± 0.013**	**0.960 ± 0.013**
Logistic Regression				
3-class, 80/20 train/test	0.388 ± 0.053	0.238 ± 0.057	0.388 ± 0.053	0.272 ± 0.050
3-class, 17 to predict one	0.611 *	0.520 *	0.611 *	0.558 *
2-class, 9 to predict one	**1.000 ***	**1.000 ***	**1.000 ***	**1.000 ***
*Since the logistic regression converges to the same values on each run, sem = 0 for these values.				

Lessons from Machine Learning

Train on the Extremes

5. Sometimes it makes more sense to TRAIN the ML Model on the EXTREMES

Advanced Consortium on Cooperation, Conflict, and Complexity
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Higher-peace Lower-peace

Sustaining Peace Project
Liebovitch et al.

Higher-peace
Lower-peace

by word frequency

yellow=importance in making the classification

	High Peace	Count		High Peace	Count
1	**time**	4.10E+05	51	lead	1.16E+05
2	people	3.78E+05	52	public	1.15E+05
3	new	3.13E+05	53	number	1.15E+05
4	**work**	3.09E+05	54	**child**	1.14E+05
5	use	3.04E+05	55	school	1.13E+05
6	**like**	2.91E+05	56	set	1.13E+05
7	come	2.68E+05	57	woman	1.09E+05
8	need	2.09E+05	58	share	1.08E+05
9	**look**	2.04E+05	59	**run**	1.08E+05
10	include	2.01E+05	60	issue	1.06E+05
11	know	1.93E+05	61	**try**	1.06E+05
12	**want**	1.88E+05	62	lot	1.05E+05
13	**way**	1.86E+05	63	told	1.04E+05
14	company	1.84E+05	64	case	1.02E+05
15	**governme**	1.82E+05	65	comment	1.00E+05
16	**game**	1.73E+05	66	**state**	9.94E+04
17	**think**	1.71E+05	67	area	9.80E+04
18	**good**	1.68E+05	68	player	9.80E+04
19	**world**	1.68E+05	69	local	9.79E+04
20	**team**	1.65E+05	70	**found**	9.71E+04
21	high	1.64E+05	71	**add**	9.67E+04
22	**home**	1.60E+05	72	health	9.61E+04
23	right	1.55E+05	73	base	9.58E+04
24	change	1.52E+05	74	**site**	9.49E+04
25	help	1.52E+05	75	follow	9.48E+04
26	business	1.50E+05	76	police	9.47E+04
27	life	1.48E+05	77	plan	9.26E+04
28	day	1.47E+05	78	win	9.19E+04
29	**start**	1.44E+05	79	ask	9.02E+04
30	service	1.41E+05	80	**find**	8.88E+04
31	**thing**	1.41E+05	81	**story**	8.86E+04
32	family	1.38E+05	82	cost	8.85E+04
33	**place**	1.37E+05	83	increase	8.84E+04
34	**country**	1.36E+05	84	**member**	8.80E+04
35	**play**	1.35E+05	85	man	8.79E+04
36	**big**	1.34E+05	86	event	8.71E+04
37	report	1.33E+05	87	young	8.70E+04
38	market	1.31E+05	88	news	8.61E+04
39	**informatio**	1.26E+05	89	**national**	8.53E+04
40	**great**	1.26E+05	90	**open**	8.52E+04
41	**provide**	1.25E+05	91	minister	8.49E+04
42	point	1.24E+05	92	system	8.45E+04
43	support	1.22E+05	93	price	8.42E+04
44	**year**	1.22E+05	94	result	8.42E+04
45	city	1.22E+05	95	level	8.39E+04
46	best	1.21E+05	96	**believe**	8.34E+04
47	long	1.21E+05	97	continue	8.31E+04
48	communit	1.20E+05	98	**experienc**	8.21E+04
49	end	1.20E+05	99	expect	8.17E+04
50	group	1.17E+05	100	term	8.15E+04

	Low Peace	Count		Low Peace	Count
1	**state**	1.88E+05	51	power	4.07E+04
2	**governme**	1.61E+05	52	**security**	4.05E+04
3	people	1.37E+05	53	group	4.03E+04
4	**country**	1.35E+05	54	support	4.01E+04
5	**president**	9.33E+04	55	federal	3.96E+04
6	**time**	9.12E+04	56	day	3.94E+04
7	come	8.10E+04	57	**start**	3.92E+04
8	**work**	6.93E+04	58	local	3.92E+04
9	new	6.85E+04	59	**place**	3.86E+04
10	use	6.76E+04	60	**sector**	3.78E+04
11	**national**	6.73E+04	61	money	3.70E+04
12	need	6.41E+04	62	**provide**	3.70E+04
13	**like**	6.28E+04	63	help	3.57E+04
14	know	6.03E+04	64	nation	3.54E+04
15	police	5.60E+04	65	office	3.51E+04
16	service	5.37E+04	66	chief	3.48E+04
17	include	5.35E+04	67	ensure	3.48E+04
18	high	5.32E+04	68	**internatio**	3.42E+04
19	**party**	5.31E+04	69	number	3.41E+04
20	governor	5.31E+04	70	road	3.40E+04
21	public	5.31E+04	71	**act**	3.35E+04
22	issue	5.23E+04	72	told	3.33E+04
23	company	5.19E+04	73	set	3.32E+04
24	**world**	5.14E+04	74	health	3.32E+04
25	**member**	5.14E+04	75	communit	3.31E+04
26	**election**	5.02E+04	76	**thing**	3.30E+04
27	developm	5.01E+04	77	**order**	3.27E+04
28	report	4.96E+04	78	increase	3.17E+04
29	**court**	4.96E+04	79	**home**	3.17E+04
30	**want**	4.89E+04	80	**team**	3.17E+04
31	business	4.86E+04	81	**look**	3.16E+04
32	**bank**	4.74E+04	82	house	3.16E+04
33	**way**	4.74E+04	83	**director**	3.15E+04
34	**good**	4.73E+04	84	**according**	3.14E+04
35	area	4.70E+04	85	education	3.12E+04
36	life	4.63E+04	86	follow	3.10E+04
37	school	4.59E+04	87	**officer**	3.06E+04
38	case	4.45E+04	88	end	3.06E+04
39	**political**	4.38E+04	89	system	3.04E+04
40	lead	4.37E+04	90	**economic**	3.03E+04
41	**general**	4.35E+04	91	**university**	3.02E+04
42	market	4.28E+04	92	change	3.01E+04
43	**project**	4.27E+04	93	family	2.98E+04
44	minister	4.26E+04	94	man	2.96E+04
45	**add**	4.21E+04	95	point	2.95E+04
46	**law**	4.20E+04	96	industry	2.92E+04
47	right	4.16E+04	97	student	2.90E+04
48	woman	4.16E+04	98	**chairman**	2.88E+04
49	**child**	4.10E+04	99	oil	2.88E+04
50	leader	4.08E+04	100	fund	2.85E+04

L. Liebovitch,
W. Powers, L. Shi
A. Chen-Carrel,
P. Loustaunau,
P. Coleman

L. Liebovitch, W. Powers, L. Shi, A. Chen-Carrel, P. Loustaunau, P. Coleman

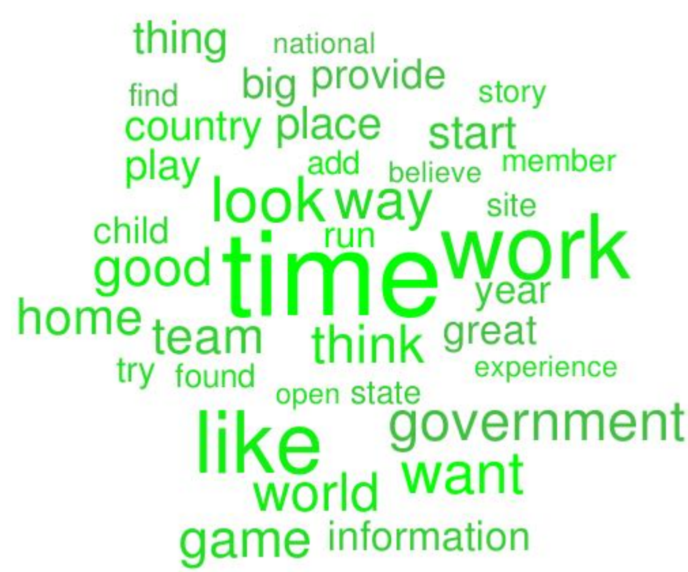
random forest

words of highest feature importance, scaled by their frequency of occurrence

Lower-peace

words: government, social control

Higher-peace



words: future, daily activities

Advanced Consortium on Cooperation, Conflict, and Complexity
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Sustaining Peace Project
Capstone 2022

Capstone 2022 Columbia MS Data Science Students

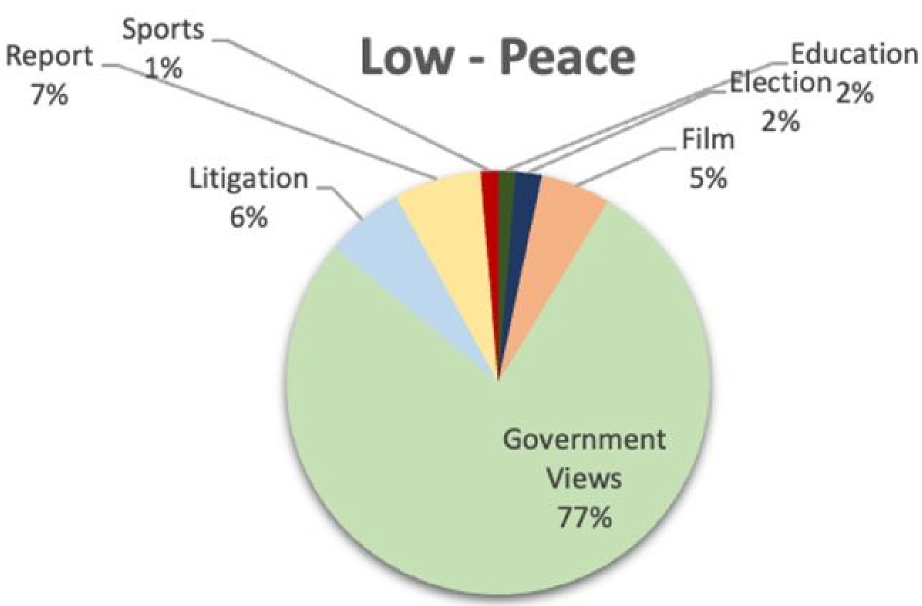
Yibo Chen, Hongou Liu, Ziheng Ru, Xinfu Su, Pinyi Yang, Yuwen Zhang

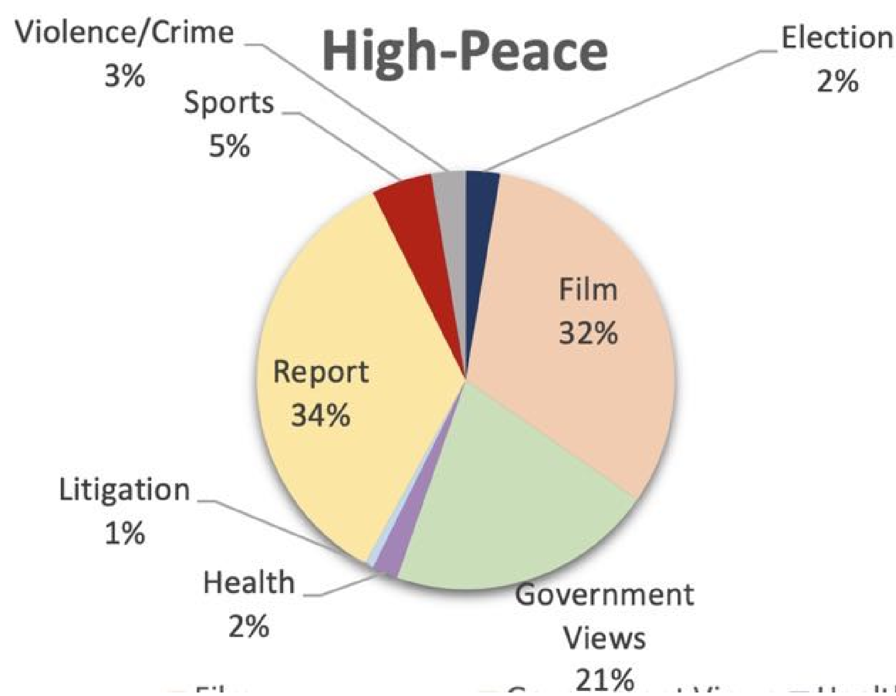
https://github.com/tthatyuwen/Peace-Speech-Project-Git

TF-IDF, DistilBERT, XLNet, k-means, logistic regression, XGBoost, SVM

Results

Identified clusters of word meanings in lower and higher peace countries.





TRAINED on EXTREMES
2-class
low, hi peace
recognizes BETWEENS as well!

Logistic Regression ML Peace Index
= 100*p

	Country	ML Index	AVG GPI	AVG PPI	AVG WHI	AVG FSI	AVG HDI
1	**Tanzania**	**6.22**	63.87	18.88	0.00	25.73	0.00
2	**Nigeria**	**6.30**	0.00	0.00	41.62	0.00	2.38
3	**Bangladesh**	**9.56**	44.52	10.04	28.27	11.55	19.05
4	Sri Lanka	**12.69**	36.13	27.71	18.85	13.76	61.90
5	Ghana	**13.61**	64.52	36.95	30.37	40.92	16.67
6	**Kenya**	**14.31**	27.10	12.45	22.25	3.63	11.90
7	Jamaica	**43.71**	43.87	53.82	51.57	45.92	50.00
8	Malaysia	**49.42**	78.06	53.41	59.42	44.60	66.67
9	Philippines	**53.78**	19.35	23.69	43.46	20.58	45.24
10	India	**56.45**	13.55	24.50	23.56	29.42	26.19
11	Hong Kong	**57.99**			49.74		100
12	**Singapore**	**90.38**	89.03	88.35	77.49	86.51	100
13	**New Zealand**	**92.50**	100	95.98	96.86	100	95.24
14	United States	**94.01**	32.90	83.53	90.31	83.84	97.62
15	**Canada**	**94.47**	92.26	95.18	100	97.34	95.24
16	**United Kingdom**	**94.47**	61.94	93.17	86.91	85.60	97.62
17	**Ireland**	**95.87**	88.39	100	89.79	99.08	100
18	**Australia**	**95.91**	89.68	93.98	97.38	97.78	100

L. Liebovitch, W. Powers, L. Shi A. Chen-Carrel, P. Loustaunau, P. Coleman

Results

- Machine learning tools accurately classify countries as lower or higher peace from their news and social media.
- There are more words of government control in lower-peace countries and more words of daily activities in higher-peace countries.
- Trained machine learning models can provide a quantitative level of peace from media data.

General Lessons

1. NLT Preprocessing is more difficult than you think, CHECK the results, did it work the way you expected?
2. Are you taking out the "common" words that may be the most important (Pennebaker)?
3. How will you handle inconsistent data from different sources?
4. How will you balance features with different amounts of data, when is this necessary?
5. Can using data from extreme cases improve the learning that can be applied to all cases?
6. Do you KNOW the mathematical ASSUMPTIONS assumed in the methods that you are using?
7. Did you check that all your methods are doing what you think they're doing?
8. Do you known enough about the subject DOMAIN to know that your methods:
   a) are appropriate and b) give you new understanding about the subject domain?