Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating LLMs

Tunazzina Islam, Dan Goldwasser

Department of Computer Science

Purdue University, West Lafayette, IN 47907, USA



International AAAI Conference on Web and Social Media

June 23-26, 2025, Copenhagen, Denmark



Unsupervised Text Analysis

- Topic Modeling.
 - Shallow Themes.



- Topic Modeling.
 - Shallow Themes.
- Manual and qualitative coding (Hagen et al., 2022; Nguyen et al., 2021; Del Valle et al., 2020).



- Predefined set of labels, themes, and arguments (Islam et al., 2023; Islam & Goldwasser, 2022).
 - Fixed and established based on existing topics or theoretical frameworks, such as Moral Foundations Theory (MFT) (Haidt and Graham 2007).
 - Often fails to capture the nuances of messaging choices.



- Predefined set of labels, themes, and arguments (Islam et al., 2023; Islam & Goldwasser, 2022).
 - Fixed and established based on existing topics or theoretical frameworks, such as Moral Foundations Theory (MFT) (Haidt and Graham 2007).
 - Often fails to capture the nuances of messaging choices.



• Latent Theme Discovery (Pacheco et al., 2023; Pacheco et al., 2022b;a).

- Human-in-loop (Pacheco et al. 2022b;a).
 - Costly scalability.
 - Time consuming.



Machine-in-the-Loop Approach

- Human-in-loop (Pacheco et al. 2022b;a).
 - Costly scalability.
 - Time consuming.
- Machine-in-the-Loop: LLMs-in-the-Loop.
 - LLMs possess extensive domain insights.
 - Reasoning capabilities.
 - Accelerate the process of refinement.





- Candidate generation
 - (a) Clustering.
 - (b) Cluster coherency check: LLMs
 - (c) Cluster summery: LLMs
 - (d) Merging/redundancy check.



- Candidate generation
 - (a) Clustering.
 - (b) Cluster coherency check: LLMs
 - (c) Cluster summery: LLMs
 - (d) Merging/redundancy check.
- Human validation given the new set of clusters decide if - (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?



- Candidate generation
 - (a) Clustering.
 - (b) Cluster coherency check: LLMs
 - (c) Cluster summery: LLMs
 - (d) Merging/redundancy check.
- Human validation given the new set of clusters decide if - (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?
- Assignment check if a new text belongs to the cluster summary by few-shot prompting LLMs.



- Candidate generation
 - (a) Clustering.
- GPT-4 (b) Cluster coherency check: LLMs(c) Cluster summery: LLMs(d) Merging/redundancy check.
 - Human validation given the new set of clusters decide if (a) Should we merge? (b) Does the cluster summary seem ok? (c) Coherent?
 GPT-4
 - Assignment check if a new text belongs to the cluster summary by few-shot prompting LLMs.



Case Studies

- Climate campaigns.
 - **21k ads** *(Islam et al. 2023)*, January 2021 to January 2022.
 - Stance (e.g., *pro-energy*, *clean-energy*) and seed theme (e.g., *support climate policy*).



Case Studies

- Climate campaigns.
 - **21k ads** *(Islam et al. 2023)*, January 2021 to January 2022.
 - Stance (e.g., *pro-energy*, *clean-energy*) and seed theme (e.g., *support climate policy*).
- COVID-19 vaccine campaigns.
 - 9k ads (Islam and Goldwasser 2022), December 2020 to January 2022.
 - Moral foundation (e.g., *care/harm*) (*Haidt and Graham, 2007*) and **seed theme** (e.g., *vaccine equity*).





Results: Coverage

• SBERT embedding for theme assignment for each ad.

CASE	METHOD	NUM.	NUM. COVERED ADS									
STUDY		THEMES	THR < 0.6	$\mathrm{THR} < 0.5$	$\mathrm{THR} < 0.4$	THR < 0.3						
	Pre-existing	13	14652	9725	3731	558						
Climate	+After Iter1	20	18702	14583	8646	2944						
	+After Iter2	25	18988	15052	9079	3180						
	Pre-existing	15	7889	6426	3480	771						
COVID-19	+After Iter1	20	8852	7627	4737	1302						
	+After Iter2	23	9092	7898	5038	1590						

Results: Coverage

- SBERT embedding for theme assignment for each ad.
- Better coverage after two iterations of *machine-inthe-loop* approach.

CASE Study	Метнор	NUM. Themes	THR < 0.6	NUM. COUTHR < 0.5	VERED ADS THR < 0.4	THR < 0.3
Climate	Pre-existing +After Iter1	13 20	14652 18702	9725 14583	3731 8646	558 2944
	+After Iter2	25	18988	15052	9079	3180
	Pre-existing	15	7889	6426	3480	771
COVID-19	+After Iter1	20	8852	7627	4737	1302
	+After Iter2	23	9092	7898	5038	1590

Results: Mapping Quality

True Label

- Ground truth:
 - 1072 climate ads.
 - 565 COVID-19 ads.
- Mapping Quality w.r.t Human Judgements.

Case Study	Method	Acc. (%)	F1 (%)
Climate	SBERT Assign.	84.05	79.32
	LLMs Mapper	88.15	89.24
COVID-19	SBERT Assign.	41.42	44.83
	LLMs Mapper	85.49	81.74

								C	Cor	nfus	sio	n №	1at	rix	(%)									
Economy_pro - 98	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	- 100
ClimateSolution - 0	91	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	
Pragmatism - 17	8	75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Patriotism - 0	0	3	91	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
AgainstClimatePolicy - 0	0	3	3	88	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	00
Economy_clean - 0	3	0	2	2	77	0	2	3	0	6	0	0	0	2	0	0	2	2	0	0	0	0	0	0	- 80
HumanHealth - 0	0	0	0	0	0	90	2	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
FutureGeneration - 0	0	0	0	0	0	0	92	5	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
Environmental – 0	1	0	0	0	0	1	0	88	1	0	1	0	0	0	0	2	0	0	2	0	0	0	0	2	
Animals – 0	0	0	0	0	0	0	0	10	88	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	60
AltEnergy – 0	0	0	0	0	1	0	0	0	0	93	0	0	0	0	0	0	1	4	0	0	0	0	0	0	- 00
SupportClimatePolicy - 0	5	1	0	0	3	0	0	10	0	0	64	11	0	5	0	0	0	0	0	0	0	0	0	0	
PoliticalAffiliation - 0	0	0	0	0	0	0	0	2	0	0	0	85	0	12	0	0	0	0	0	0	0	0	0	0	
BidenGasPriceIncrease - 0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	
AgainstCorporateInterests - 0	0	0	0	0	0	0	0	0	0	0	5	2	0	92	0	0	0	0	0	0	0	0	0	0	- 40
GasTax - 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	40
Deforestation - 0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	82	12	0	0	0	4	0	0	0	
Carbon – 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	2	0	0	0	0	0	0	
CustomerBasedAltEnergy – 0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	81	0	0	0	0	0	0	
FoodSecurity - 0	3	0	0	0	0	0	5	3	0	0	0	0	0	0	0	0	5	0	84	0	0	0	0	0	- 20
EnergyAffordabilityandSustainabilityLegislation - 0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	89	0	0	0	0	20
EcofriendlyConsumerChoices - 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	
PlasticWasteandEnvironmentalImpact - 0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	89	0	0	
PromoteSustainableTransportation - 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	
— WaterManagementandSustainability - 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	- 0
2 2	Ľ	Ē	E	5 C	C.	÷	с С	a	<u>s</u>	ΣĒ	- S	с С	e G	ts	×	L L	ц	کل	ح	L L	S	ť	L L	ţ	-
	ıti	tis	tis	olio	lea	ealt	atic	ent	Шâ	jero,	olic	atic	eas	res	sTå	atic	rþ	erç	nri	atic	oice	рa	atic	ilia	
Ê	Solu	ma	trio	teP	~	Ή	ler	Ĕ	Ani	μ	tеР	filië	JC	hte	ß	esta	Ca	Шu	jec	islà	G	Ш	ť	na	
ou	ite(ag	Ъа	na	Ш	nal	Ger	D D	-	A	na	١Af	Gel	telr		ore		ĮĄ	Spc	e G	Jer	nta	spo	sta	
	Шa	đ		Ü	ouo	ЪГ	Le(nz.			Cli	ic.	Pric	Dra		Def		sec	ĕ	iťyl	'n	nei	ran	Sus	
	Ü			Jst	БС		ıtu	ш			τo	olit	as	гþа		_		Ba		lid	suc	DUC	еT	pu	
				gai			Ē				dd	а.	5uG	ů				ner		ina	Ň	۲ir	ab	nta	
				Ă							Su		ide	nst				tor		sta	Įb.	En	ain	nei	
													8	gai				Cus		Su	rier	pue	ust	ger	
														∢				0		anc	cof	te	eS	na	
																				itya	ш	Vas	not	Ăa	
																				lid		icV	ron	ter	
																				rda		ast	۵.	Wa	
																				ffo		Ч			
Confusion Mate		v •	(71	11	n	nt	0												JуА					
Confusion Man	12	Υ.	l		11	110	al	.e												erc					
																				ЕП					

Results: Mapping Quality

- Ground truth:
 - 1072 climate ads.
 - 565 COVID-19 ads.
- Mapping Quality w.r.t **Human** Judgements.

Case Study	Method	Acc. (%)	F1 (%)
Climate	SBERT Assign.	84.05	79.32
	LLMs Mapper	88.15	89.24
COVID-19	SBERT Assign.	41.42	44.83
	LLMs Mapper	85.49	81.74

									(Con	fus	ion	Mat	trix	(%)										100
GovDistrust -	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	100
GovTrust -	0	58	0	0	8	0	0	0	0	0	8	25	0	0	0	0	0	0	0	0	0	0	0	0		
VaccineRollout -	1	1	78	0	0	5	12	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0		
VaccineSymptom -	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
VaccineEquity -	0	0	0	0	94	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	_	80
VaccineStatus -	0	0	12	0	0	56	6	0	0	0	0	12	12	0	0	0	0	0	0	0	0	0	0	0		00
EncourageVaccination -	0	1	4	0	0	0	91	0	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	1		
VaccineMandate -	2	0	0	0	0	0	0	93	0	0	0	0	0	0	1	1	1	0	0	1	0	1	0	0		
VaccineReligion -	0	0	0	0	0	0	17	0	83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
VaccineEfficacy -	0	0	10	0	0	5	13	0	0	64	3	0	0	3	0	0	0	0	0	0	0	0	0	3	-	60
VaccineDevelopment -	0	0	0	0	8	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0	0	0	0	0		
CovidPlan -	0	0	0	0	0	0	10	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0		
VaccineMisinformation -	0	0	0	0	8	0	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0	0	0		
NaturalImmunity -	0	0	0	0	0	0	0	0	0	17	0	0	0	83	0	0	0	0	0	0	0	0	0	0		
Vote -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	-	40
VaccineRefusalNews -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0		
MaskMandate -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0		
VaccineBrewIncentives -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0		
CommunityServiceByCandidate -	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	0		
AdvocatingUnifiedLiberties -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	89	0	0	0	0	-	20
CovidEconomy -	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0		
UrgentPoliticalAdvocacy -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0		
AgainstSocialistPolicy -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0		
none -	33	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33		0
	GovDistrust -	GovTrust -	VaccineRollout -	VaccineSymptom -	VaccineEquity -	VaccineStatus -	EncourageVaccination -	VaccineMandate -	VaccineReligion -	VaccineEfficacy -	VaccineDevelopment -	CovidPlan -	VaccineMisinformation -	Naturallmmunity -	Vote -	VaccineRefusalNews -	MaskMandate -	VaccineBrewIncentives -	nunityServiceByCandidate -	AdvocatingUnifiedLiberties -	CovidEconomy -	UrgentPoliticalAdvocacy -	AgainstSocialistPolicy -	none -		0
fucion Matrix				x 7	T		1	0	1										Com							

Predicted Label

Confusion Matrix: COVID-19

True Label

Qualitative Analysis: Climate

- Correlation heatmap between identified **themes** and **stances** expressed in the ads (i.e., pro-energy or clean-energy).
- Stronger correlations with stances than the derived LDA and BERTopic topics.





Pro-Energy \rightarrow 'BidenGasPriceIncrease' & 'GasTax'

Clean-Energy \rightarrow 'Environmental' & 'AltEnergy'

Carbon -	-0.13	0.13	- 0.4
SupportClimatePolicy -	-0.13	0.13	
PoliticalAffiliation -	-0.13	0.13	
FoodSecurity -	-0.11	0.11	- 0.3
EcofriendlyConsumerChoices -	-0.11	0.11	
AgainstCorporateInterests -	-0.14	0.14	
BidenGasPriceIncrease -	0.33	-0.33	- 0.2
Deforestation -	-0.13	0.13	
Pragmatism -	0.17	-0.17	
PlasticWasteandEnvironmentalImpact -	-0.054	0.054	- 0.1
ClimateSolution -	0.27	-0.27	
WaterManagementandSustainability -	-0.06	0.06	
Environmental -	-0.16	0.16	- 0.0
EnergyAffordabilityandSustainabilityLegislation -	-0.054	0.054	
PromoteSustainableTransportation -	-0.063	0.063	
AgainstClimatePolicy -	0.29	-0.29	0.1
FutureGeneration -	-0.12	0.12	
AltEnergy -	-0.16	0.16	
Animals -	-0.12	0.12	0.2
Economy_clean -	-0.14	0.14	
Economy_pro -	0.41	-0.41	
CustomerBasedAltEnergy -	-0.15	0.15	0 3
HumanHealth -	-0.11	0.11	-0.5
Patriotism -	0.3	-0.3	
GasTax -	0.37	-0.37	- 0.4
	pro-energy	clean-energy	-0.4

Baseline: 10 LDA Topics Baseline: 30 LDA Topics Baseline: 15 BERTopic Topics

- 0.0

-0.1

- -0.2

- -0.3

- -0.4

Ours: After 2nd round of iteration

Qualitative Analysis: COVID-19

topic--1 - -0.023 0.043 -0.039 0.062 0.19

topic-8 - -0.024 -0.052 -0.084 -0.015 -0.08

- Correlation heatmap between identified themes and moral foundation expressed in the ads (i.e., liberty/ oppression, fairness/cheating).
- Stronger correlations with moral foundations than ٠ the derived LDA and BERTopic topics.

-0.058 -0.052 0.17 -0.015 -0.055 -0.031 0.019

0.019 -0.048 -0.051 -0.014 0.17 -0.039 -0.05

0.028 0.038 0.13 -0.008 0.046 -0.13 -0.029

0.092 -0.043 -0.025

-0.046

-0.024 -0.039

0.068 -0.035

-0.038 0.023 -0.017 -0.0095 -0.014 -0.027 0.1

-0.038 -0.034 0.02 -0.0095 0.063 -0.0034 -0.035

0.075 0.003 0.081 -0.012 -0.0044 -0.084 -0.000

099 -0.0044 0.063 -0.013 -0.072 -0.053 0.032

0.064 -0.034 -0.054 -0.0095 -0.052 0.068 -0.035

0.00056 0.014 0.047 -0.016 -0.088 0.032 -0.025

0.1 -0.046 -0.008 0.0012 -0.043

0.031 -0.022 -0.045 -0.016 -0.017 0.032 0.0085

0.0072 0.017 -0.023 0.17 -0.056 0.035 -0.037

0.061 -0.046 -0.075 0.13 0.013 0.018 -0.0079

-0.012 -0.038 -0.02 -0.07

- 0.3

- 0.2

- 0.1

- -0.1

-0.2

-0.3

- -0.4

topic-4 -

topic-6 -

topic-16 - -0.049 -0.043 0.018 -0.012 -0.0072 0.056

0.038 -0.0019 -0.008

-0.038 0.023 -0.017 -0.0095 -0.052

0.021 0.076 0.0049 -0.013 -0.046

0.038 -0.034 -0.017 -0.0095 0.024

topic-11 - -0.051 0.04 -0.017 -0.013 -0.013 0.03 -0.0062

topic-23 - -0.049 -0.043 -0.07 -0.012 0.023 0.075 -0.002

topic-2 - -0.018 0.073 0.002 -0.014 0.0075 -0.056 0.067

topic-19 - -0.047 0.005 -0.0059 -0.012 0.062 0.0056 -0.043

topic-12 - 0.002 0.012 -0.029 -0.011 0.042 -0.08 0.1

0.0045 -0.037 -0.026 -0.01

topic-1

topic-10 -

topic-3 -

topic-17

topic-24

topic-13 -

topic-6 -

topic-21

topic-20

topic-14 -

-0.1

-0.2

- -0.3

-0.4

tonic-4

topic-7 - -0.032

topic-5

topic-15 -

topic-8

topic-18 -

topic-0 - -0.032

-0.076 -0.0076 0.069 -0.019 0.099 -0.046 -0.042

0.025 0.06 0.014 0.11 -0.031 -0.0088 -0.019

.071 0.045 -0.07 -0.012 0.083 -0.057 -0.0025

0.038 -0.011 -0.0076 -0.024 0.036 -0.0096 -0.042

.055 -0.026 -0.0054 -0.022 0.058 -0.053 0.02

0046 -0.054 0.032 -0.015 -0.084 0.033 0.049

0.016 0.0015 -0.017 -0.018 0.12 -0.055 -0.034

0.077 0.026 -0.069 0.036 -0.14 0.16 -0.004

0.021 -0.011 -0.029 -0.014 -0.051 0.023

-0.02 -0.034 -0.055 0.0079

Fairness/Cheating \rightarrow 'vaccine equity' Liberty/Oppression \rightarrow 'vaccine mandate' & 'advocating unified liberties'



Baseline: 10 LDA Topics Baseline: 25 LDA Topics Baseline: 15 BERTopic Topics

- How Themes Differ based on Gender.
- Extract ads **targeted exclusively** at *males* and *females*.

- How Themes Differ based on Gender.
- Extract ads **targeted exclusively** at *males* and *females*.





- How Themes Differ based on Gender.
- Extract ads **targeted exclusively** at *males* and *females*.





Theme: FutureGeneration





- How Themes Differ based on Red vs. Blue States.
 - ► North Dakota (ND) vs. Vermont (VT)
- Different emphasis placed on the *entity* 'Community'.

- How Themes Differ based on Red vs. Blue States.
 - ► North Dakota (ND) vs. Vermont (VT)
- Different emphasis placed on the *entity* 'Community'.



- How Themes Differ based on **Red** vs. **Blue** States.
 - ► North Dakota (ND) vs. Vermont (VT)
- Different emphasis placed on the *entity* 'Community'.







- Event1: Gas Price Increase, Date: May 12, 2021.
- Theme Freq.: AltEnergy, ClimateSolution increased.



Event: 12 May, 2021 Gas prices rise over \$3/GAL for the first time since 2014

- Event1: Gas Price Increase, Date: May 12, 2021.
- Theme Freq.: AltEnergy, ClimateSolution increased.
- New Themes: AgainstCorporateInterests, BidenGasPriceIncrease.



- Event2: Federal COVID-19 vaccine mandate, Date: September 09, 2021.
- Theme Freq.: VaccineMandate increased.



Event: Biden announced COVID-19 vaccine mandates for federal workers on September 09, 2021

- Event2: Federal COVID-19 vaccine mandate, Date: September 09, 2021.
- Theme Freq.: VaccineMandate increased.
- New Themes: GovDistrust, VaccineMisinformation.



Event: 12 May, 2021 Gas prices rise over \$3/GAL for the first time since 2014



Findings:

- Sponsored content is **highly responsive** to real-world events
- Reflect and potentially shape public discourse on pivotal issues.

References

- 1. Carroll J Glynn and Michael E Huge. Public opinion. The international encyclopedia of communication, 2008.
- 2. Vincent Price. On the public aspects of opinion: Linking levels of analysis in public opinion research. Communication research, 15(6):659–679, 1988.
- 3. J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," Social Justice Research, vol. 20, no. 1, pp. 98–116, 2007.
- T. Islam, R. Zhang, and D. Goldwasser, "Analysis of climate campaigns on social media using bayesian model averaging," in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, ser. AIES '23, Montréal, QC, Canada: Association for Computing Machinery, 2023, pp. 15–25, isbn: 9798400702310.
- 5. Tunazzina Islam and Dan Goldwasser. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In 2022 IEEE International Conference on Big Data (BigData), pp. 585–595. IEEE, 2022.
- 6. Tunazzina Islam and Dan Goldwasser. Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating LLMs. arXiv preprint arXiv:2403.10707, 2024.
- 7. Mar'ıa Leonor Pacheco*, Tunazzina Islam*, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. A holistic framework for analyzing the covid-19 vaccine debate. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5821–5839, 2022a. (* equal contribution)
- 8. Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. Interactively uncovering latent arguments in social media platforms: A case study on the covid-19 vaccine debate. In Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pp. 94–111, 2022b.
- 9. Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. Interactive concept learning for uncovering latent themes in large text collections. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Findings of the Association for Computational Linguistics: ACL 2023, pp. 5059–5080, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.313. URL https://aclanthology.org/2023.findings-acl.313.

References

- 10. Loni Hagen, Ashley Fox, Heather O'Leary, DeAndre Dyson, Kimberly Walker, Cecile A Lengacher, and Raquel Hernandez. The role of influential actors in fostering the polarized covid-19 vaccine discourse on twitter: Mixed methods of machine learning and inductive coding. Jmir Infodemiology, 2(1):e34231, 2022.
- 11. Thu T. Nguyen, Shaniece Criss, Eli K. Michaels, Rebekah I. Cross, Jackson S. Michaels, Pallavi Dwivedi, Dina Huang, Erica Hsu, Krishay Mukhija, Leah H. Nguyen, Isha Yardi, Amani M. Allen, Quynh C. Nguyen, and Gilbert C. Gee. Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. SSM Population Health, 15:100922, 2021. ISSN 2352-8273. doi: https://doi.org/10.1016/j.ssmph.2021.100922. URL https://www.sciencedirect. com/science/article/pii/S235282732100197X.
- 12. Marc Esteve Del Valle, Rimmert Sijtsma, Hanne Stegeman, and Rosa Borge. Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. Javnost-The Public, 27(3):211–229, 2020.



Code and Dataset: <u>https://github.com/tunazislam/latent-themes-llms/tree/main</u> Slide: <u>https://tunazislam.github.io/files/LatentThemesLLMsICWSM25.pdf</u>



Tunazzina Islam, Ph.D.

Department of Computer Science, Purdue University, West Lafayette, IN. Email: islam32@purdue.edu







Backup Slides

Research Questions (RQs)

• **RQ1**: Can LLMs determine if two given texts, without prior knowledge of existing codes, are discussing the same topic?

Research Questions (RQs)

- **RQ1**: Can LLMs determine if two given texts, without prior knowledge of existing codes, are discussing the same topic?
- **RQ2**: If provided with a definition of a theme, can LLMs successfully categorize other texts under that specific theme?

Coherency Checking

1. Here are some key things to know about used motor oil:

200 million gallons of oil are dumped into U.S. waterways, sewers, and landfills by people changing their own motor oil.

& Used motor oil is considered hazardous waste in California.

Used motor oil is recyclable! Every 1 gallon of used oil recovered can produce 2.5 quarts of re-refined oil.

A Oil never wears out—it just gets dirty. Most cars can go more than 5,000 miles between oil changes.

2. Check out more interesting facts about used motor oil and ways to prevent oil pollution at https://bit.ly/3wFZue1.

The oil and gas industry in Eddy County is the economic driver for our state. I fully support the oil and gas industry and am proud of the environmental standards these great companies impose on themselves.

3. The EPA will hold three virtual listening sessions to take public input on the Agency's upcoming regulations for the Oil and Natural Gas Industry. What you cannot see at an oil & gas well can hurt you. Local Front Range residents are visiting oil and gas sites to learn how invisible methane gas and other health-harming pollution are threatening their communities.

4. Biodiesel (also known as biofuel) is a gallon-for-gallon substitute for petroleum-based fuels, which have a higher carbon intensity. By 2030, it's estimated that biodiesel will displace 529 million gallons of heating oil!

5. Biodiesel is nontoxic, biodegradable renewable liquid fuel. The biofuel that's blended with heating oil to create Bioheat fuel is generally produced by agricultural byproducts, including used cooling oil, animal fats, inedible corn oil, soybean oil and canola oil. This puts excess oil and fats to good use!

Supports motor oil

-Against oil & gas industry

-Supports biofuel

Coherency Checking: Incoherent Cluster

1. Here are some key things to know about used motor oil:

200 million gallons of oil are dumped into U.S. waterways, sewers, and landfills by people changing their own motor oil.

 ${\boldsymbol{\ast}}$ Used motor oil is considered hazardous waste in California.

Used motor oil is recyclable! Every 1 gallon of used oil recovered can produce 2.5 quarts of re-refined oil.

A Oil never wears out—it just gets dirty. Most cars can go more than 5,000 miles between oil changes.

2. Check out more interesting facts about used motor oil and ways to prevent oil pollution at https://bit.ly/3wFZue1.

The oil and gas industry in Eddy County is the economic driver for our state. I fully support the oil and gas industry and am proud of the environmental standards these great companies impose on themselves.

3. The EPA will hold three virtual listening sessions to take public input on the Agency's upcoming regulations for the Oil and Natural Gas Industry. What you cannot see at an oil & gas well can hurt you. Local Front Range residents are visiting oil and gas sites to learn how invisible methane gas and other health-harming pollution are threatening their communities.

4. Biodiesel (also known as biofuel) is a gallon-for-gallon substitute for petroleum-based fuels, which have a higher carbon intensity. By 2030, it's estimated that biodiesel will displace 529 million gallons of heating oil!

5. Biodiesel is nontoxic, biodegradable renewable liquid fuel. The biofuel that's blended with heating oil to create Bioheat fuel is generally produced by agricultural byproducts, including used cooling oil, animal fats, inedible corn oil, soybean oil and canola oil. This puts excess oil and fats to good use!

Supports motor oil

LLMs categorize this cluster as **incoherent** cluster.

-Against oil & gas industry

Supports biofuel

Merging Checking

Summary of cluster 1: #PlantingChallenge aims to plant trees in California to combat climate change, while another initiative promotes the planting of "super trees" with superior carbon storage capabilities. Oxford is considering a Tree Planting and Urban Forestry program to maintain its aging tree canopy, emphasizing the multiple benefits of trees. Criticism is directed at the Obama Presidential Center for cutting down over 300 mature trees, underlining the importance of preserving older trees for carbon storage. Additionally, a service offers to plant trees worldwide for a small monthly fee to help mitigate climate change and replenish forests.

Summary of cluster 18: The Tongass National Forest in Alaska, a crucial carbon sink and habitat, is at the center of environmental advocacy efforts. Accusations against Central Maine Power (CMP) suggest misleading claims about renewable energy development, while campaigns emphasize the Tongass's role in storing 44% of carbon absorbed by U.S. national forests and its significance for clean air, water, and climate. Calls to action focus on reinstating the Roadless Rule to protect the forest from logging and development, with the U.S. Forest Service planning to reinforce these protections. This highlights the forest's importance for salmon populations, local communities, and climate resilience, urging public involvement in conservation efforts.

•Embed the summaries using SBERT.

•Compute the **cosine similarity** between the embedding summaries of cluster pairs.

Merging Checking: Merged

Summary of cluster 1: #PlantingChallenge aims to plant trees in California to combat climate change, while another initiative promotes the planting of "super trees" with superior carbon storage capabilities. Oxford is considering a Tree Planting and Urban Forestry program to maintain its aging tree canopy, emphasizing the multiple benefits of trees. Criticism is directed at the Obama Presidential Center for cutting down over 300 mature trees, underlining the importance of preserving older trees for carbon storage. Additionally, a service offers to plant trees worldwide for a small monthly fee to help mitigate climate change and replenish forests.

Summary of cluster 18: The Tongass National Forest in Alaska, a crucial carbon sink and habitat, is at the center of environmental advocacy efforts. Accusations against Central Maine Power (CMP) suggest misleading claims about renewable energy development, while campaigns emphasize the Tongass's role in storing 44% of carbon absorbed by U.S. national forests and its significance for clean air, water, and climate. Calls to action focus on reinstating the Roadless Rule to protect the forest from logging and development, with the U.S. Forest Service planning to reinforce these protections. This highlights the forest's importance for salmon populations, local communities, and climate resilience, urging public involvement in conservation efforts.

Argument: Take initiatives to plant new trees and protect existing forests to save our planet.

Clusters are **merged** (cosine similarity ≥ 0.6).

Results: Coverage

• Climate

- Pre-existing themes:

► Coverage: 17.5% ads.

- After Iter1:

• Coverage: 40.5% ads.

- After Iter2:
 - Coverage: **42.5%** ads.

CASE	METHOD	NUM.	NUM. COVERED ADS									
STUDY		THEMES	THR < 0.6	THR < 0.5	THR < 0.4	THR < 0.3						
Climate	Pre-existing	13	14652	9725	3731	558						
	+After Iter1	20	18702	14583	8646	2944						
	+After Iter2	25	18988	15052	9079	3180						
COVID-19	Pre-existing	15	7889	6426	3480	771						
	+After Iter1	20	8852	7627	4737	1302						
	+After Iter2	23	9092	7898	5038	1590						

Results: Coverage

• Climate

- Pre-existing themes:

• Coverage: 17.5% ads.

- After Iter1:

• Coverage: 40.5% ads.

- After Iter2:

• Coverage: 42.5% ads.

• COVID-19

- Pre-existing themes:Coverage: 35.08% ads.
- After Iter1:

• Coverage: 47.75% ads.

- After Iter2:

• Coverage: **50.79%** ads.

	CASE Study	Метнор	NUM. Themes	THR < 0.6	NUM. CONTHR < 0.5	ERED ADS THR < 0.4	THR < 0.3
	Climate	Pre-existing +After Iter1 +After Iter2	13 20 25	14652 18702 18988	9725 14583 15052	3731 8646 9079	558 2944 3180
•	COVID-19	Pre-existing +After Iter1 +After Iter2	15 20 23	7889 8852 9092	6426 7627 7898	3480 4737 5038	771 1302 1590

Is Machine-in-the-Loop Helpful?



• More diverse distribution of themes!