Information Extraction from Social Media: Tasks, Data, and Open-Source Tools

Shubhanshu Mishra[1*], NLP Researcher
Rezvaneh (Shadi) Rezapour[2], PhD Candidate
Jana Diesner[2], Associate Professor

[1] Twitter, Inc.

[2] University of Illinois at Urbana-Champaign (UIUC)

***Work presented here was done during my PhD at UIUC**

Content and views expressed in this tutorial are solely the responsibility of the presenters.

https://socialmediaie.github.io/tutorials/WWW2021/

ILLINOIS
School of Information Sciences

THE WEB CONFERENCE

Initial setup

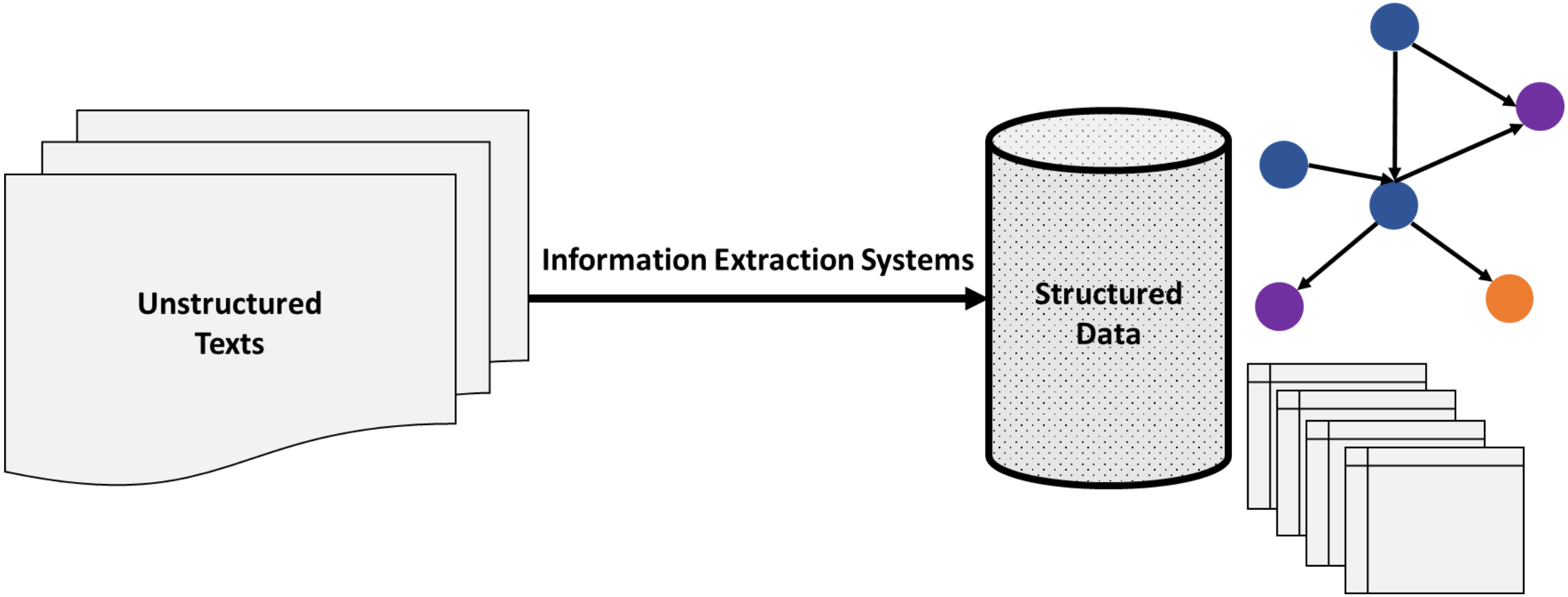
- Open google Colab notebook specified at: https://socialmediaie.github.io/tutorials/WWW2021/#software-setup
- On Colab click **Connect**
- Then on the Menu click **Runtime > Restart and run all**
- Meanwhile you can also follow the steps on the link above to install SocialMediaIE locally on your machine.
- If you face any issues with installation please report an issue at: https://github.com/socialmediaie/SocialMediaIE/issues

Agenda

- Introduction (30 mins) (Shubhanshu)
- Applications of Information Extraction(IE) (30 mins) (Shubhanshu and Shadi)
- Collecting and distributing social media data (20 mins)
- Break (10 mins)
- Hands on Practice (Shubhanshu)
  - Improving IE on social media data using machine learning (1 hr)
- Conclusion and future direction (20 mins)

Introduction

Information extraction https://shubhanshu.com/phd_thesis/



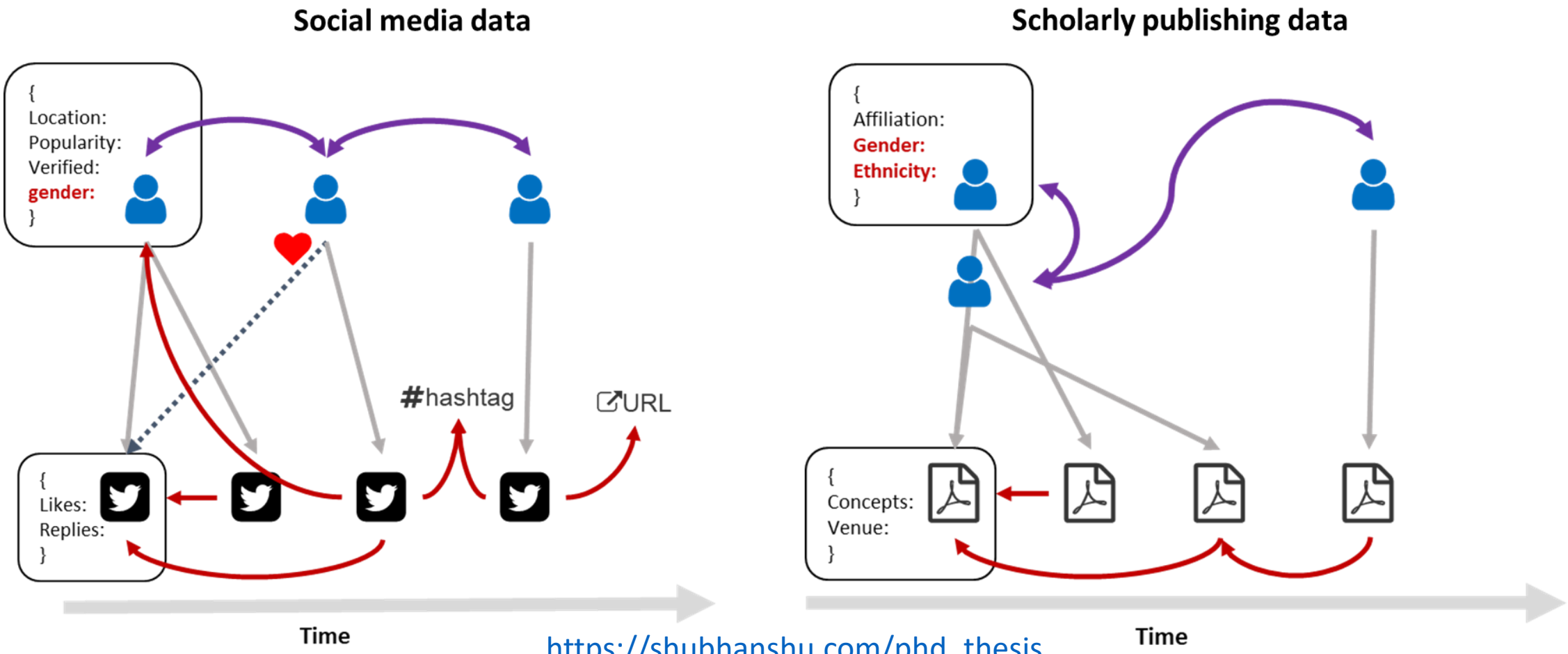
"Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources."
– (Sarawagi, 2008)

Digital Social Trace Data https://shubhanshu.com/phd_thesis/

Digital Social Trace Data (DSTD) are digital activity traces generated by individuals as part of a social interactions, such as interactions on social media websites like Twitter, Facebook; or in scientific publications.

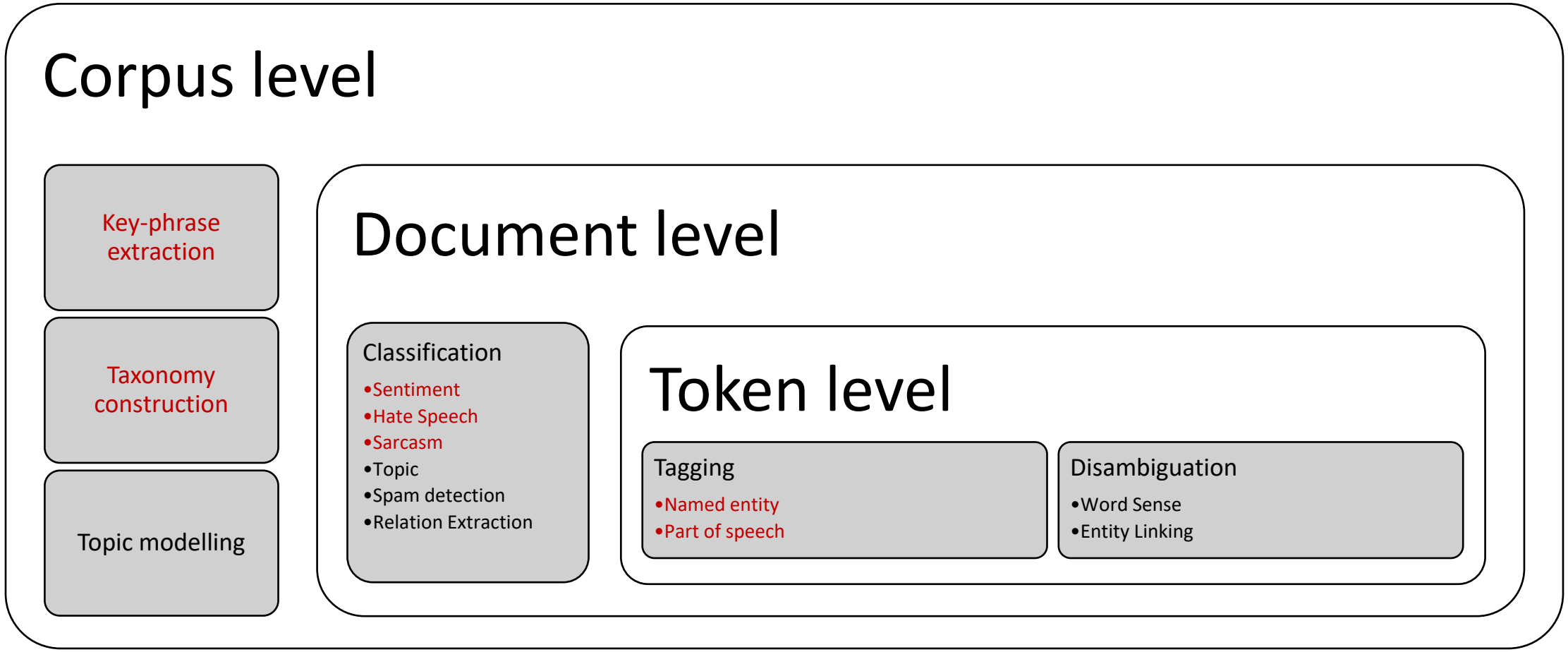
Inspired from Digital Trace Data (Howison et. al, 2011)

Digital Social Trace Data (DSTD)



https://shubhanshu.com/phd_thesis

Information extraction tasks https://shubhanshu.com/phd_thesis



Examples of information extraction for social media text

Text classification https://github.com/socialmediaie/SocialMediaIE

Input

I know this tweet is late but I just want to say I absolutely fucking hated this season of
@GameOfThrones
what a waste of time.

Predict

Output

abusive

founta

abusive	hateful	normal	spam
0.830	0.084	0.085	0.002

waseem

none	racism	sexism
0.970	0.002	0.027

sentiment

clarin

negative	neutral	positive
0.956	0.036	0.008

other

negative	neutral	positive
0.906	0.063	0.031

politics

negative	neutral	positive
0.917	0.048	0.035

semeval

negative	neutral	positive
0.966	0.030	0.004

uncertainity

sarcasm

not sarcasm	sarcasm
0.914	0.086

veridicality

definitely no	definitely yes	probably no	probably yes	uncertain
0.033	0.244	0.112	0.189	0.422

Sequence tagging https://github.com/socialmediaie/SocialMediaIE

Input

john oliver coined the term donal drumph as a joke on his show #LastWeekTonight

Predict

Output

tokens	john	oliver	coined	the	term	donal	drumph	as	a	joke	on	his	show	#LastWeekTonight
ud_pos	PROPN	PROPN	VERB	DET	NOUN	PROPN	PROPN	ADP	DET	NOUN	ADP	PRON	NOUN	X
ark_pos	^	^	V	D	N	^	^	P	D	N	P	D	N	#
ptb_pos	NNP	NNP	VBD	DT	NN	NNP	NNP	IN	DT	NN	IN	PRP$	NN	HT
multimodal_ner	PER					PER								
broad_ner	PER													
wnut17_ner	PERSON													
ritter_ner	PERSON													
yodie_ner	PERSON													
ritter_chunk	NP	VP	NP			NP		PP	NP		PP	NP		
ritter_ccg	NOUN.PERSON	VERB.COMMUNICATION		NOUN.COMMUNICATION						NOUN.COMMUNICATION			NOUN.COMMUNICATION	

Applications of information extraction

Index documents by entities

DocID	Entity	Entity type	WikiURL
1	Roger Federer	Person	URL1
2	Facebook	Organization	URL2
3	Katy Perry	Music Artist	URL3

Entity mention clustering

Washington is a great place.
I just visited **Washington**.

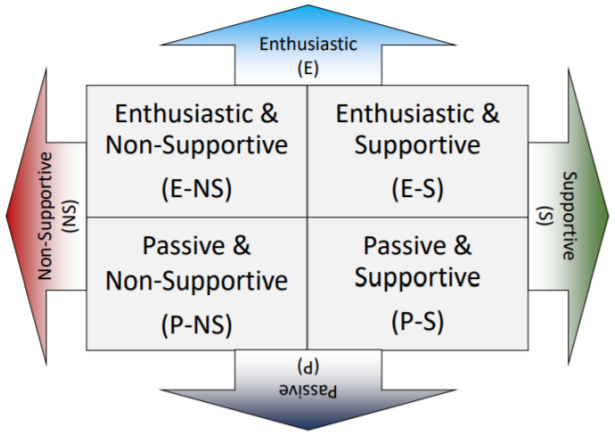
Washington was a great president.
Washington made some good changes to constitution.

Applications of Information extraction

Applications

- Indexing social media corpora in database
- Network construction from text corpora,
- Visualizing temporal trends in social media corpora using social communication temporal graphs,
- Aggregating text-based signals at user level, Improving text classification using user level attributes,
- Analyzing social debate using sentiment and political identity signals otherwise,
- Detecting and Prioritizing Needs during Crisis Events (e.g., COVID19),
- Mining and Analyzing Public Opinion Related to COVID-19, and
- Detecting COVID-19 Misinformation in Videos on YouTube

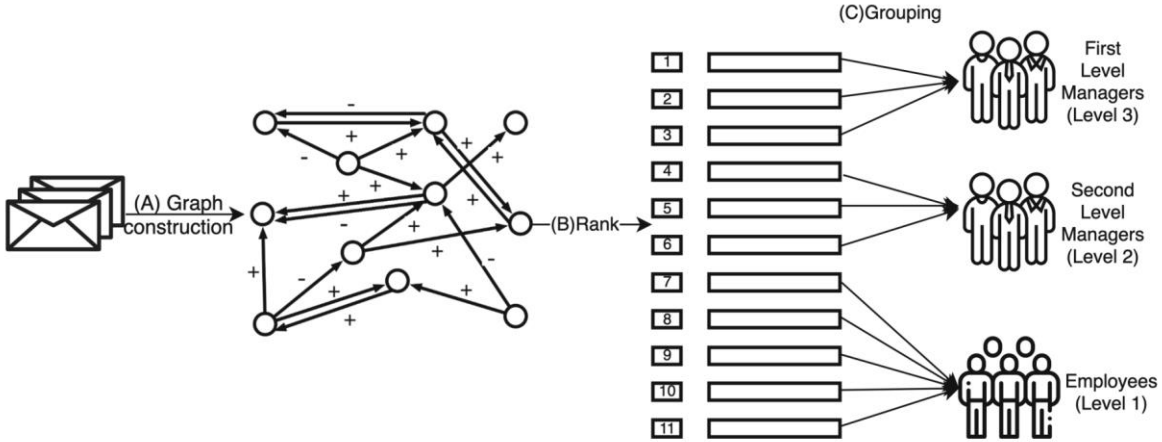
Network construction from text classification labels and identification of influential users



	CTE Account	PR	CB Account	PR	LGBT Account	PR
E/P	USR1	0.191	USR2	0.050	free_equal	0.033
	Sports_Brain	0.191	USR4	0.050	UN_Women	0.030
	USR3	0.041	USR5	0.043	USR_FilmExpert	0.030
S/NS	USR6	0.186	USR2	0.062	free_equal	0.044
	USR12	0.068	USR4	0.062	HRC	0.033
	NFL	0.066	USR5	0.054	USR_FilmExpert	0.028
All	USR7	0.021	USR8	0.009	HRC	0.024
	NFL	0.015	USR9	0.008	Tedofficialpage	0.010
	frontlinepbs	0.009	USR10	0.008	USR11	0.010

Table 9: Top 3 nodes in the mention network based on different PageRank algorithms (PR=PageRank score). In the *All* row, ranking and scores are based on overall PageRank. Accounts of individuals were replaced with USR to protect privacy.

Using signed networks in Email Corpora



- Mishra, Shubhanshu, and Jana Diesner. "Capturing signals of enthusiasm and support towards social issues from twitter." Proceedings of the 5th International Workshop on Social Media World Sensors. 2019.
- Jiang, Lan, Ly Dinh, Rezvaneh Rezapour, and Jana Diesner. "Which Group Do You Belong To? Sentiment-Based PageRank to Measure Formal and Informal Influence of Nodes in Networks." In International Conference on Complex Networks and Their Applications, pp. 623-636. Springer, Cham, 2020.

Identifying trending topics and events

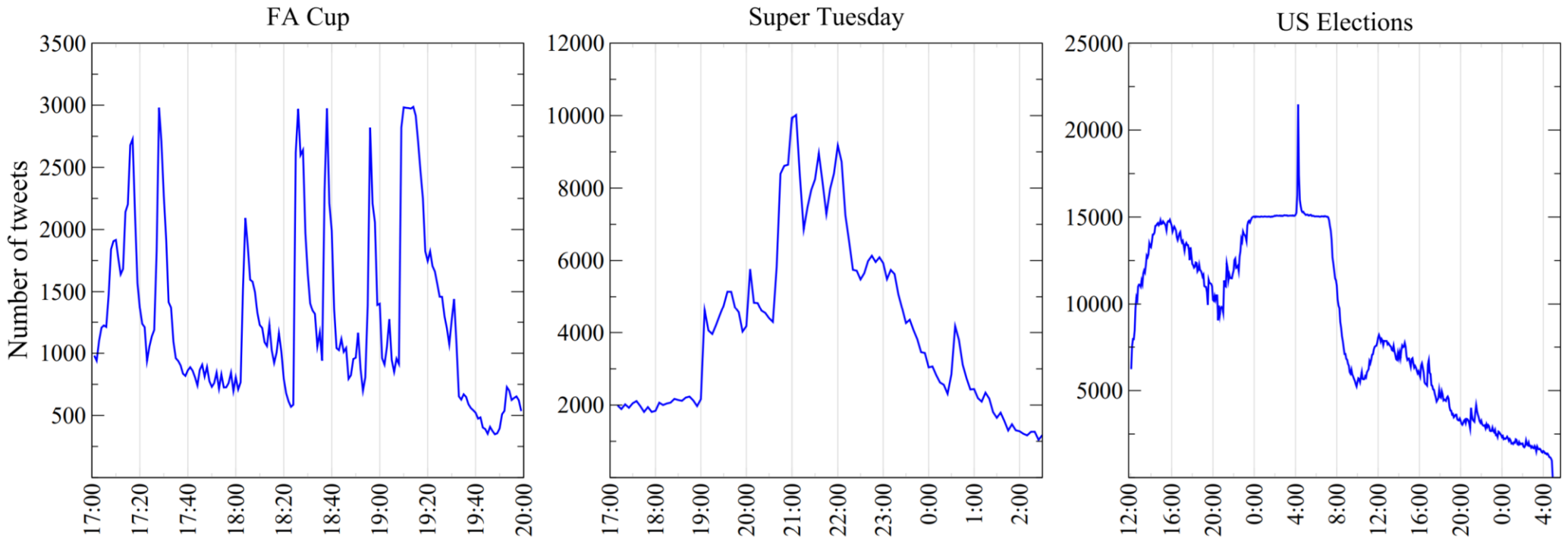
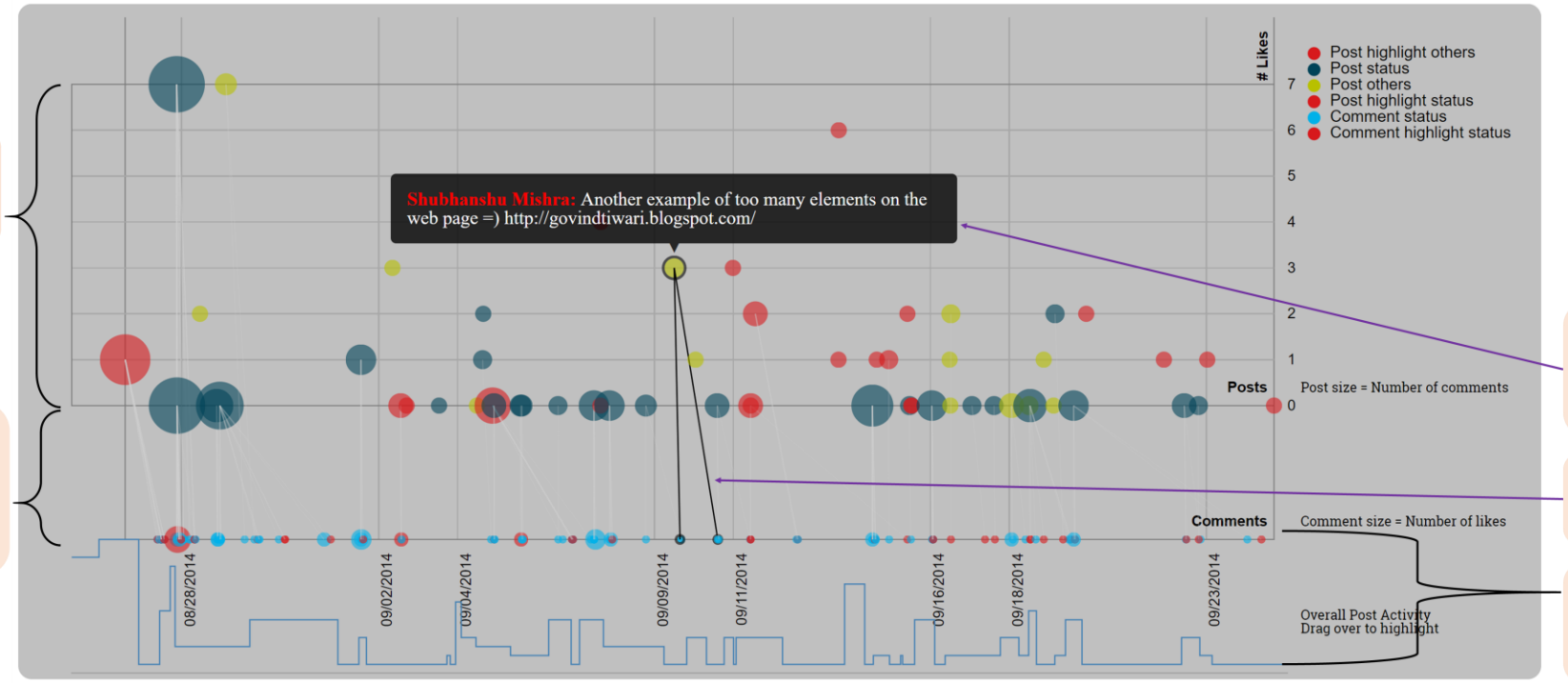


Fig. 2. Twitter activity during events. For the FA Cup, the peaks correspond to start and end of the match and the goals. For the two political collections, the peaks correspond to the main result announcements.

Aiello, Luca Maria, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. "Sensing trending topics in Twitter." IEEE Transactions on Multimedia 15, no. 6 (2013): 1268-1282.

Visualizing temporal trends in data

https://shubhanshu.com/social-comm-temporal-graph/



Core components
e.g. a user in a feed or a specific post

Child components
e.g. associated posts by a user or comments to a post

Tool tips
Provide detail on demand about each point.

Links between components
Core communication is liked to its children

Activity timeline
This quantifies the temporal activity measurement

Lexicon-based Approach

Utilizes a lexicon to describe or extract information from a textual content, e.g., lexicon-based sentiment analysis to analyze polarity of text

- What to consider first:
  - How is the lexicon created
  - Scope:
    - Using MPQA lexicon to study hashtags in Tweets 
- Domain Adaptation
  - Fine-tuning of the lexicon to represent the data
- Evaluation of the results
  - Error analysis, hand annotation, close-reading,..

Sentiment Analysis, Presidential Election, and Candidates' Ranking

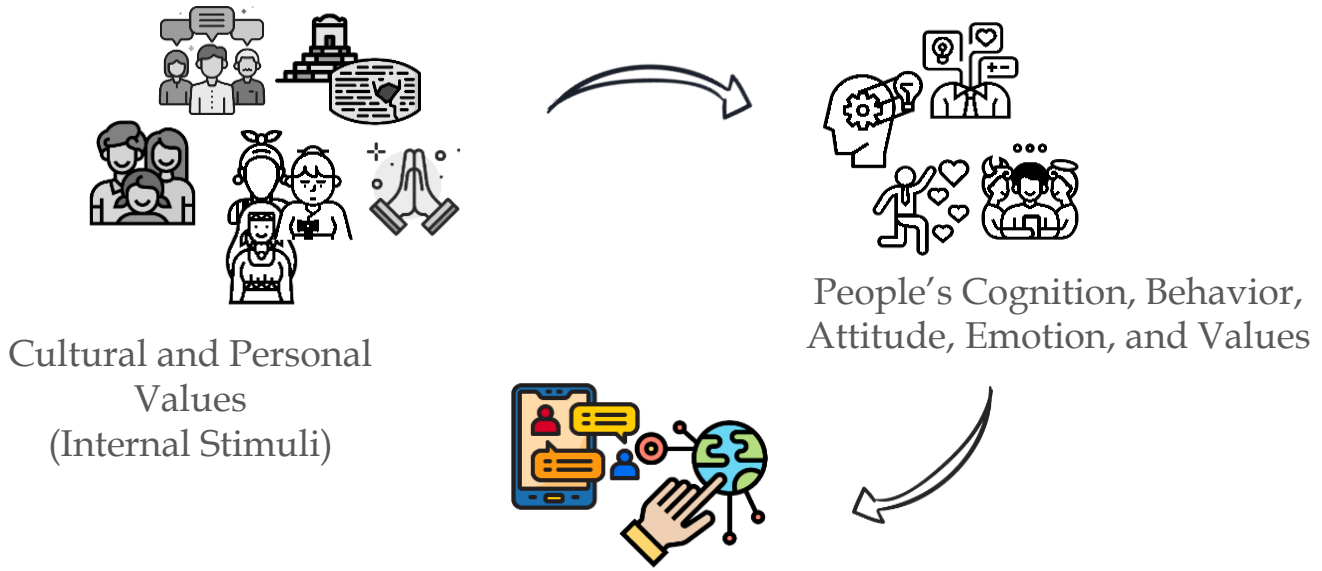
- Aim:
  - Test whether incorporating prevalent hashtags from a given dataset into a sentiment lexicon improves sentiment prediction accuracy
- Method:
  - Used hashtag-enhanced lexicon-based sentiment analysis to analyze tweets that mention the US Presidential candidates to find the correlation between the candidates' likeability in tweets with the actual voting outcomes in the New York State Presidential Primary election
  - Domain adapted the MPQA lexicon:
    - Extracted and annotated top hashtags and added them to the MPQA lexicon

Rezapour, R., Wang, L., Abdar, O., & Diesner, J. (2017). **Identifying the overlap between election result and candidates' ranking based on hashtag-enhanced, lexicon-based sentiment analysis.** In *2017 IEEE 11th International Conference on Semantic Computing (ICSC).* (pp. 93-96).

Using moral foundations to analyze social effects

- Motivation:

  “A language is not just words. It’s a culture, a tradition, a unification of a community, a whole history that creates what a community is. It’s all embodied in a language.” (Noam Chomsky)



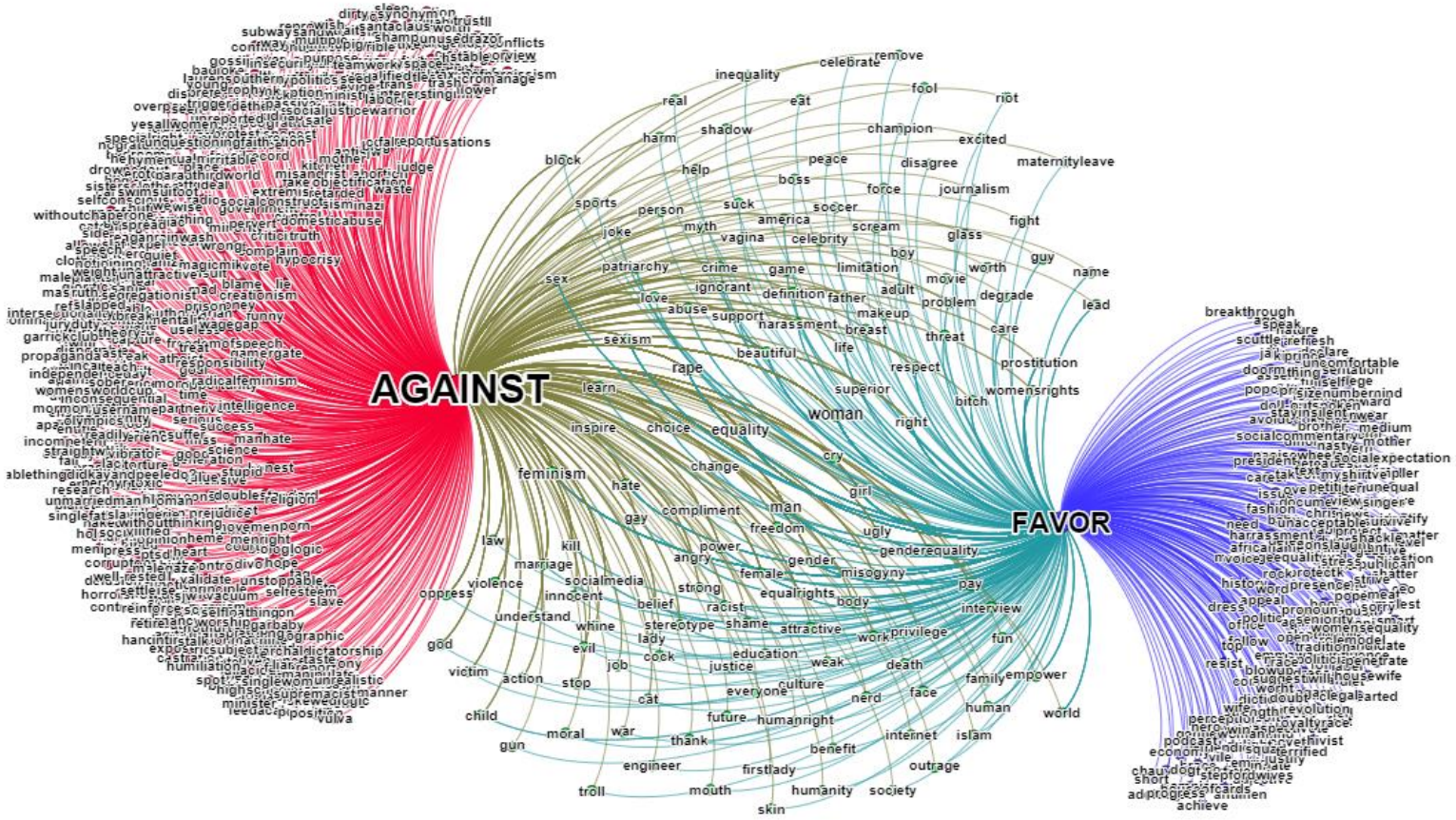
Using moral foundations analysis in analyzing social effects (contd.)

- Method:
  - Use Moral Foundations Dictionary (MFD) to extract words with moral weights and use them as features in prediction models
- Limitations with MFD:
  - Number of entries is small and might not capture (all) variations of terms indicative of morality in text data.
  - Entries are not syntactically disambiguated, which can limit the results, e.g., by capturing false positives.
    - Safe (noun) -> does not signal morality
    - Safe (adjective) -> represents care-virtue
- Enhanced MFD:
  - Used wordnet to get synonym, antonym and hypernym of the words and extensively pruned the lexicon

Rezapour, R., Shah, S. H., & Diesner, J. (2019). **Enhancing the measurement of social effects by capturing morality.** In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (WASSA). Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

Rezapour, Rezvaneh; Diesner, Jana (2019): **Expanded Morality Lexicon.** University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-3805242_V1.1

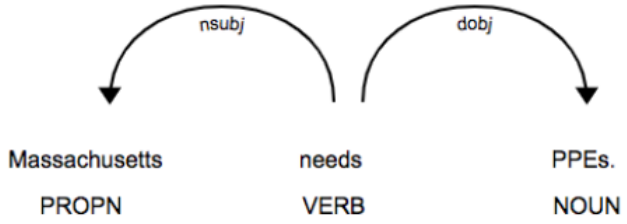
Analyzing Tweets to Examine Cross-Cutting Exposure in Social Media



Rezapour, R., Park, J., Diesner, J. (2020). Detecting Characteristics of Cross-cutting Language Networks on Social Media. In International Sunbelt Social Network Conference, Paris, France.

Detecting and Prioritizing Needs during Crisis Events (i.e., COVID19)

- Method:
  - Created a list of needed resources ranked by priority
    - Extracted phrases and terms closest to the terms “needs” and “supplies”
  - Extracted sentences that specify who-needs-what resources
    - Identified sentences where who is the subject and what is the direct object
    - Selected sentences where the left child of need in the dependency parse tree is a nominal subject (nsubj), and the right child is a direct object (dobj)



Sarol, M. J., Dinh, L., Rezapour, R., Chin, C. L., Yang, P., & Diesner, J. (2020, November). An Empirical Methodology for Detecting and Prioritizing Needs during Crisis Events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 4102-4107).

More on COVID 19 Crisis

- Hate speech detection (Hardage et al. 2020)
- Misinformation related to COVID 19 (Hossain et al. 2020)
- Symptom detection using social media data (Santosh et al. 2020)
- Impact of COVID 19 on language diversity (Dunn et al. 2020)
- Quantifying the effects of COVID 19 on mental health (Biester at al. 2020)

Collecting and distributing social media data

Use of Social Media Data for Research

- Publicly available online data provides a unique source of rich input for analyzing and studying people, their behavior, and feelings
- Availability of different tools from domains such as NLP and ML made it easier for everyone to perform various types of data analysis
- Things to consider before using any data:
  - How the data is it collected
  - Is the data reusable for your research
  - Is the data representative enough
  - Does the data or method answer your research question
  - How generalizable is the findings?

Publicly available social media data

- Many researchers make annotated social media data publicly available **for academic research**.
- Good place for benchmarking or evaluating your models.
- Many datasets available for text classification.
- Few for information extraction via sequence tagging (but still enough)
- Varied annotation practices and data scope:
- We have curated a large collection of social media corpuses from academic research at: https://socialmediaie.github.io/MetaCorpus/

Tagging data

https://shubhanshu.com/phd_thesis

Super sense tagging

data	split	labels	sequences	vocab	tokens
Ritter	train	40	551	3174	10652
	dev	37	118	1014	2242
	test	40	118	1011	2291
Johannsen2014	test	37	200	1249	3064

Part of speech tagging

data	split	labels	sequences	vocab	tokens
Owoputi	train	25	1547	6572	22326
	dev	23	327	2036	4823
	test	23	500	2754	7152
TwitIE	dev	43	269	1229	2998
	test	45	632	3539	12196
Ritter	train	45	632	3539	12196
	dev	38	71	695	1362
	test	42	84	735	1627
Tweetbankv2	dev	17	710	3271	11759
	train	17	1639	5632	24753
	test	17	1201	4699	19095
DiMSUM2016	train	17	4799	9113	73826
	test	17	1000	4010	16500
Foster	test	12	250	1068	2841
lowlands	test	12	1318	4805	19794

Named entity recognition

data	split	labels	sequences	vocab	tokens
YODIE	train	13	396	2554	7905
	test	13	397	2578	8032
Ritter	train	10	1900	7695	36936
	dev	10	240	1731	4612
	test	10	254	1776	4921
WNUT2016	train	10	2394	9068	46469
	test	10	3850	16012	61908
	dev	10	1000	5563	16261
WNUT2017	train	6	3394	12840	62730
	dev	6	1009	3538	15733
	test	6	1287	5759	23394
NEEL2016	train	7	2588	9731	51669
	dev	7	88	762	1647
	test	7	2663	9894	47488
Finin	train	3	10000	19663	172188
	test	3	5369	13027	97525
Hege	test	3	1545	4552	20664
BROAD	train	3	5605	19523	90060
	dev	3	933	5312	15169
	test	3	2802	11772	45159
MultiModal	train	4	4000	20221	64439
	dev	4	1000	6832	16178
	test	4	3257	17381	52822
MSM2013	train	4	2815	8514	51521
	test	4	1450	5701	29089

Chunking

data	split	boundaries	labels	labels	sequences	vocab	tokens
Ritter	train	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP, CONJP]	9	551	3158	10584
	dev	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP]	8	118	994	2317
	test	[I, B, O]	[ADJP, PP, INTJ, ADVP, PRT, NP, SBAR, VP]	8	119	988	2310

Classification data

https://shubhanshu.com/phd_thesis

data	split	tokens	tweets	vocab
Airline	dev	20079	981	3273
	test	50777	2452	5630
	train	182040	8825	11697
Clarin	dev	80672	4934	15387
	test	205126	12334	31373
	train	732743	44399	84279
GOP	dev	16339	803	3610
	test	41226	2006	6541
	train	148358	7221	14342
Healthcare	dev	15797	724	3304
	test	16022	717	3471
	train	14923	690	3511
Obama	dev	3472	209	1118
	test	8816	522	2043
	train	31074	1877	4349
SemEval	dev	105108	4583	14468
	test	528234	23103	43812
	train	281468	12245	29673

Sentiment classification

data	split	tokens	tweets	vocab
Founta	dev	102534	4663	22529
	test	256569	11657	44540
	train	922028	41961	118349
WaseemSRW	dev	25588	1464	5907
	test	64893	3659	10646
	train	234550	13172	23042

Abusive content identification

data	split	tokens	tweets	vocab
Riloff	dev	2126	145	1002
	test	5576	362	1986
	train	19652	1301	5090
Swamy	dev	1597	73	738
	test	3909	183	1259
	train	14026	655	2921

Uncertainty indicator classification

Collecting new social media data

- **Twarc** is a good tool to collect Twitter data: https://twarc-project.readthedocs.io/en/latest/
- It requires that you have a Twitter Developer API key - https://developer.twitter.com/en/apps
- It also allows you to also hydrate tweet IDs to tweet json using the API
- Often a file with one tweet ID per line can be hydrated as:

```
twarc hydrate ids.txt > data.jsonl
twarc search blacklivesmatter > tweets.jsonl
twarc followers jack > users.jsonl
twarc users ids.txt > users.jsonl
```

Methods for Extracting Information from Social Media Data

Machine learning approaches

Rule or Lexicon-based approaches

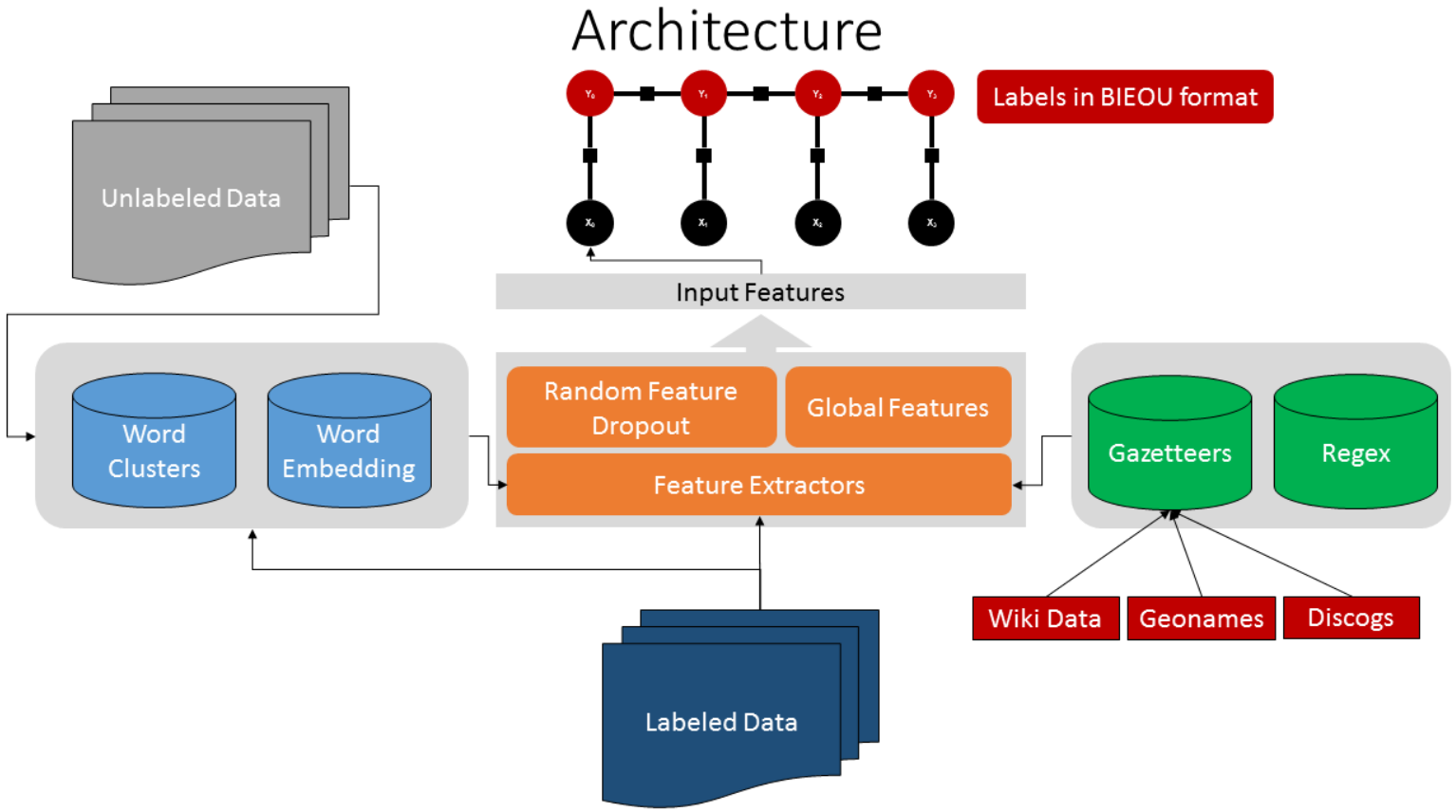
Network analysis

GUI tool for using IE to extract networks from text data

- ConText tool: http://context.ischool.illinois.edu/
- Bread and butter techniques for text analysis and extracting relational data from text data
- Convert text into network data

Rule based Twitter NER

Mishra & Diesner (2016). https://github.com/napsternxg/TwitterNER



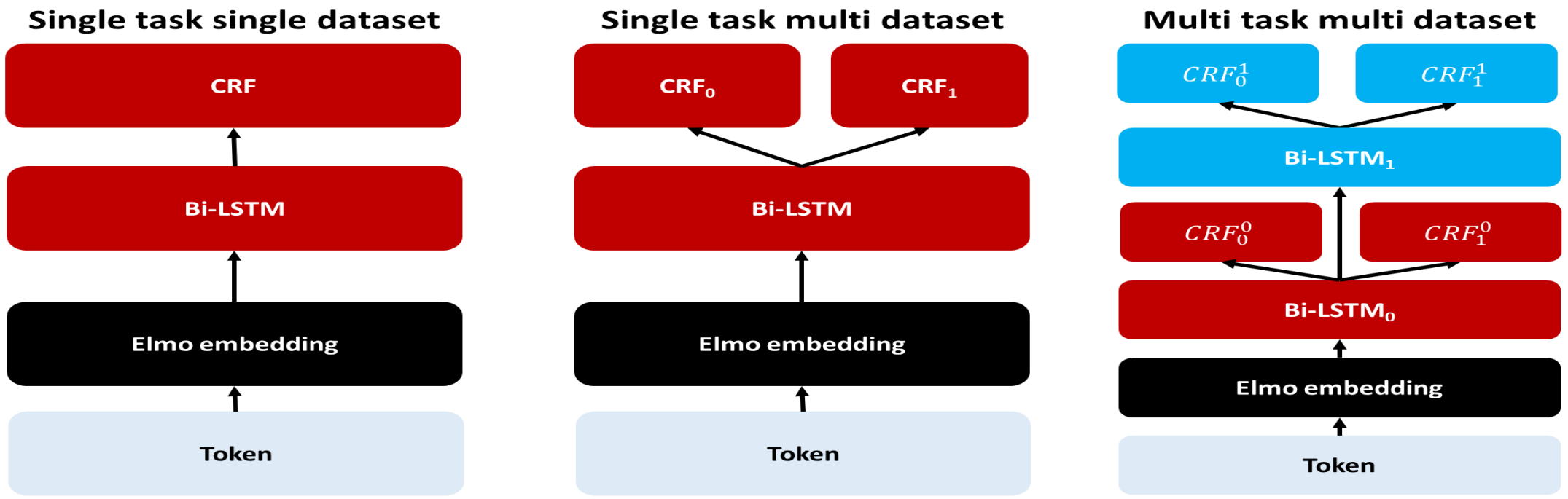
Mishra, Shubhanshu, & Diesner, Jana (2016). Semi-supervised Named Entity Recognition in noisy-text. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (pp. 203–212). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclweb.org/anthology/papers/W/W16/W16-3927/

Evaluating Twitter NER (F1-score) Mishra & Diesner (2016).

Rank	1	2	3	4	5	6	7	8	9	10	TD	TDT_E
10-types	52.4	46.2	44.8	40.1	39.0	37.2	**37.0**	36.2	29.8	19.3	**46.4**	**47.3**
No-types	65.9	63.2	60.2	59.1	55.2	51.4	**47.8**	46.7	44.3	40.7	**57.3**	**59.0**
company	57.2	46.9	43.8	31.3	38.9	34.5	**25.8**	42.6	24.3	10.2	42.1	46.2
facility	42.4	31.6	36.1	36.5	20.3	30.4	**37.0**	40.5	26.3	26.1	37.5	34.8
geo-loc	72.6	68.4	63.3	61.1	61.1	57.0	**64.7**	60.9	47.4	37.0	70.1	71.0
movie	10.9	5.1	4.6	15.8	2.9	0.0	**4.0**	5.0	0.0	5.4	0.0	0.0
musicartist	9.5	8.5	7.0	17.4	5.7	37.2	**1.8**	0.0	2.8	0.0	7.6	5.8
other	31.7	27.1	29.2	26.3	21.1	22.5	**16.2**	13.0	22.6	8.4	31.7	32.4
person	59.0	51.8	52.8	48.8	52.0	42.6	**40.5**	52.3	34.1	20.6	51.3	52.2
product	20.1	11.5	18.3	3.8	10.0	7.3	**5.7**	15.4	6.3	0.8	10.0	9.3
sportsteam	52.4	34.2	38.5	18.5	34.6	15.9	**9.1**	19.7	11.0	0.0	31.3	32.0
tvshow	5.9	0.0	4.7	5.4	7.3	9.8	**4.8**	0.0	5.1	0.0	5.7	5.7
Rank	1	2	3	4	5	6	**7**	8	9	10	~2	~2

Multi-task-multi-dataset learning

Mishra 2019, HT' 19



(A) **(B)** **(C)**

S - Single

MD – Multi-dataset
MTS – Multi task Shared

MTL – Multi task Stacked (Layered)

Shubhanshu Mishra. 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19). ACM, New York, NY, USA, 283-284. DOI: https://doi.org/10.1145/3342220.3344929

Evaluating MTL models Mishra 2019, HT' 19

Part of speech tagging (overall accuracy)

Data	Our best	SOTA	Diff %
DiMSUM2016	86.77	82.49	5%
Owoputi	91.76	88.89	3%
TwitIE	91.62	89.37	3%
Ritter	92.01	90	2%
Tweetbankv2	92.44	93.3	-1%
Foster	69.34	90.4	-23%
lowlands	68.1	89.37	-24%

Super sense tagging (micro f1)

Data	Our best	SOTA	Diff %
Ritter	59.16	57.14	3.5%
Johannsen2014	42.38	42.42	-0.1%

Chunking (micro f1)

Data	Our best	SOTA	Diff %
Ritter	88.92	None	NA

Named entity recognition (micro f1)

Data	Our best	SOTA	Diff %
BROAD	77.40	None	NA
YODIE	65.39	None	NA
Finin	56.42	32.43	74.0%
MSM2013	80.46	58.72	37.0%
Ritter	86.04	82.6	4.2%
MultiModal	73.39	70.69	3.8%
Hege	89.45	86.9	2.9%
WNUT2016	53.16	52.41	1.4%
WNUT2017	49.86	49.49	0.8%

Shubhanshu Mishra. 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19). ACM, New York, NY, USA, 283-284. DOI: https://doi.org/10.1145/3342220.3344929

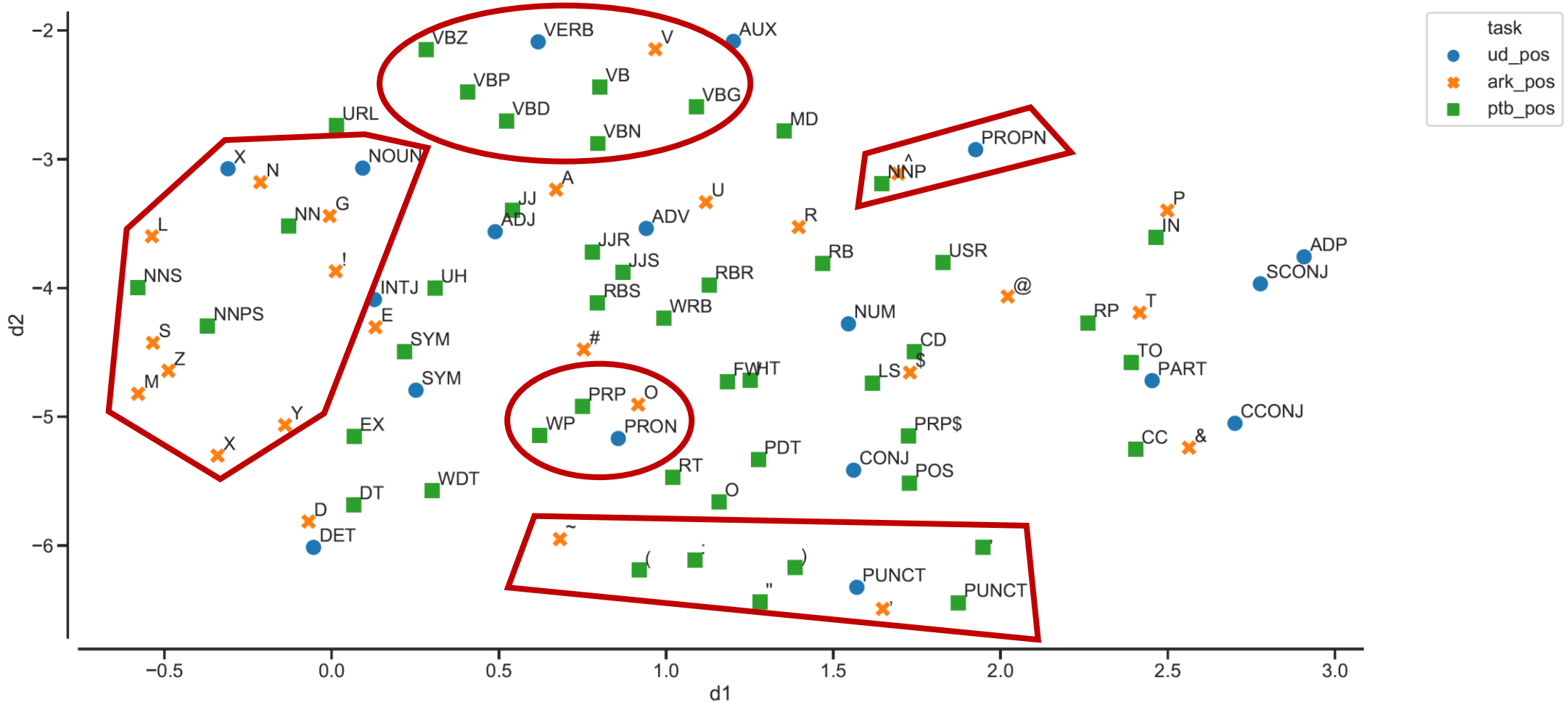
Training Mishra 2019, HT' 19

- Sample mini-batches from a task/data
- Compute loss for the mini-batch
- Individual loss is the log loss for conditional random field
- Update the model except the Elmo module
- During an epoch go through all tasks and datasets
- Train for a max number of epochs
- Use early stopping to stop training

- Models trained on single datasets have prefix **S**
- Models trained on all datasets of same task have prefix **MD**
- Models trained on all datasets have prefix **MTS** for multitask models with **shared module**, and **MTL** for **stacked modules**
- Models with LR=1e-3 and no L2 regularization have suffix **"*"**
- Models trained without NEEL2016 have suffix **"#"**

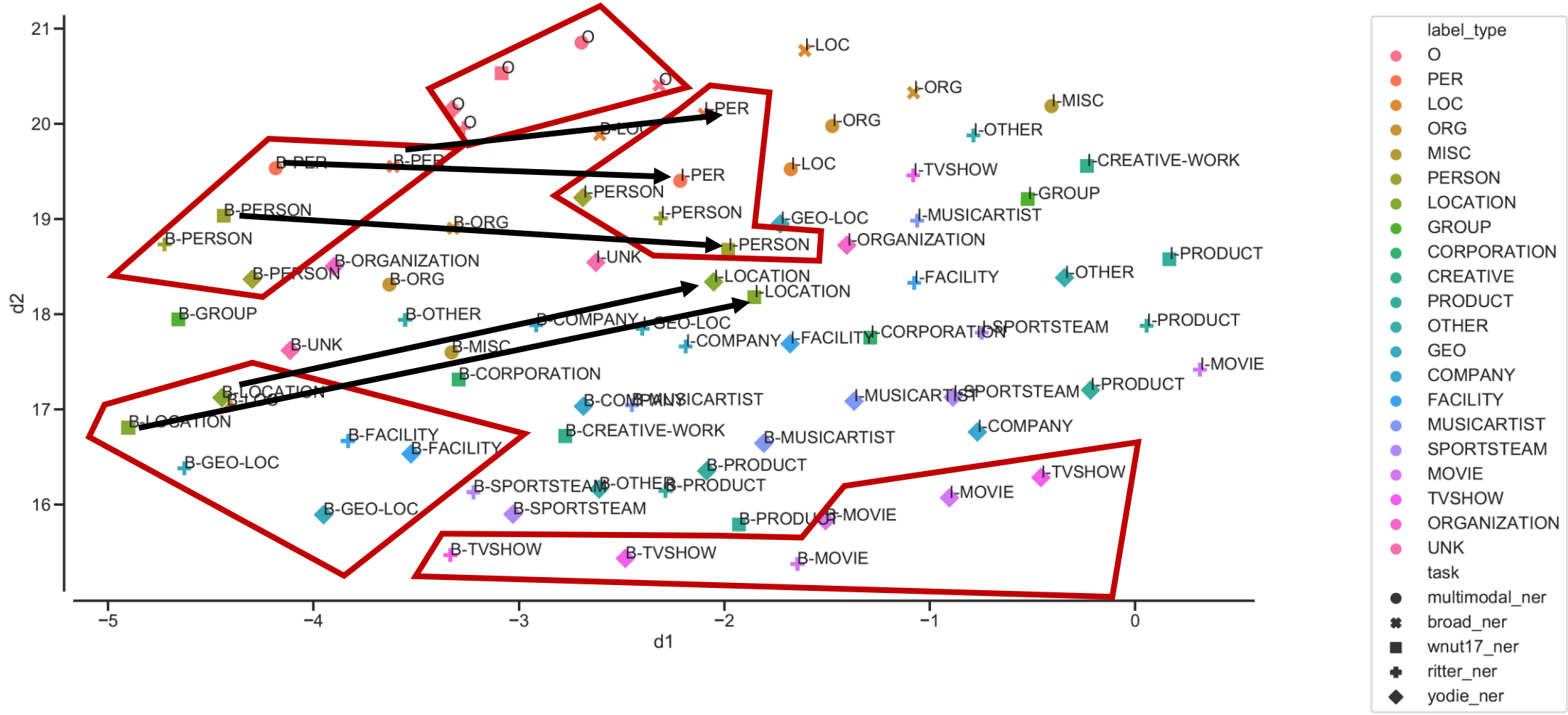
Label embeddings (POS)

- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets



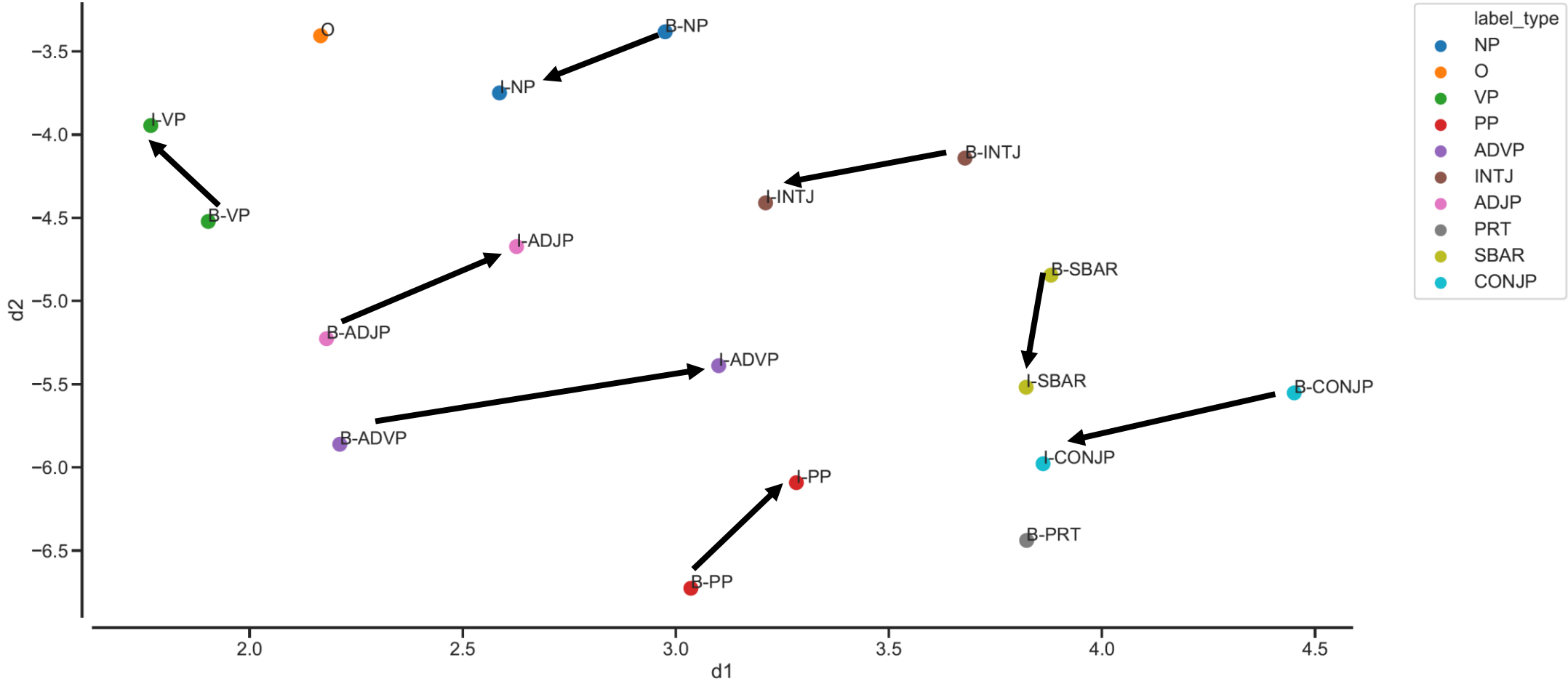
Label embeddings (NER)

- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets

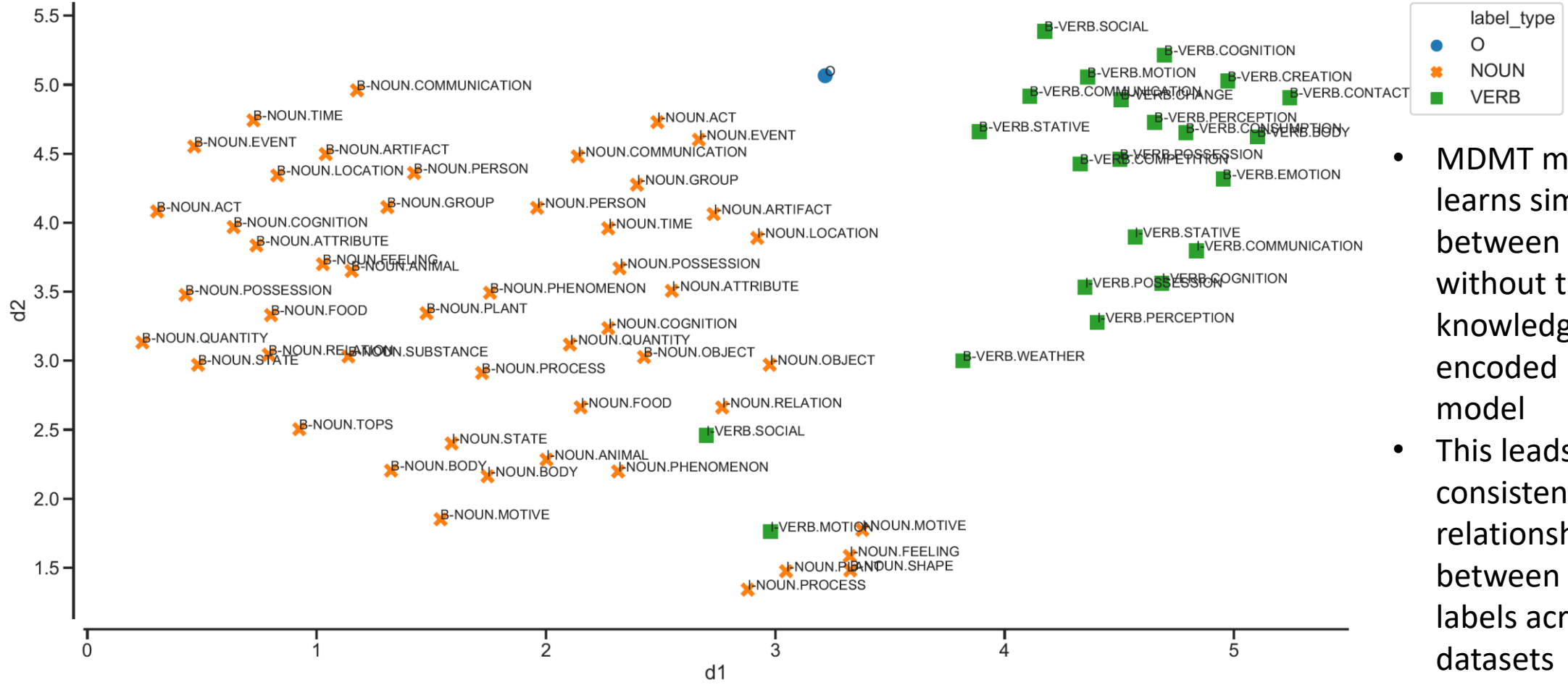


Label embeddings (chunking)

- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets

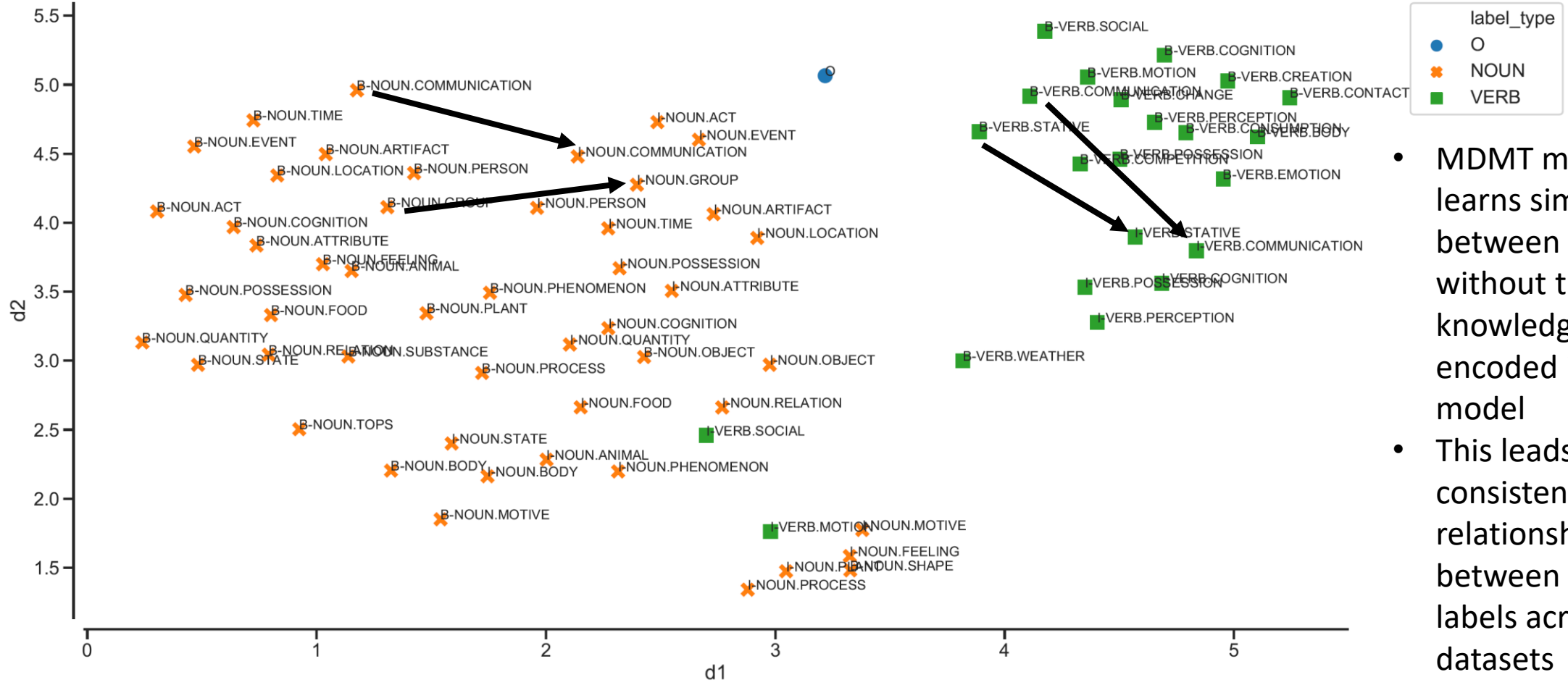


Label embeddings (super-sense tagging)



- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets

Label embeddings (super-sense tagging)



- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets

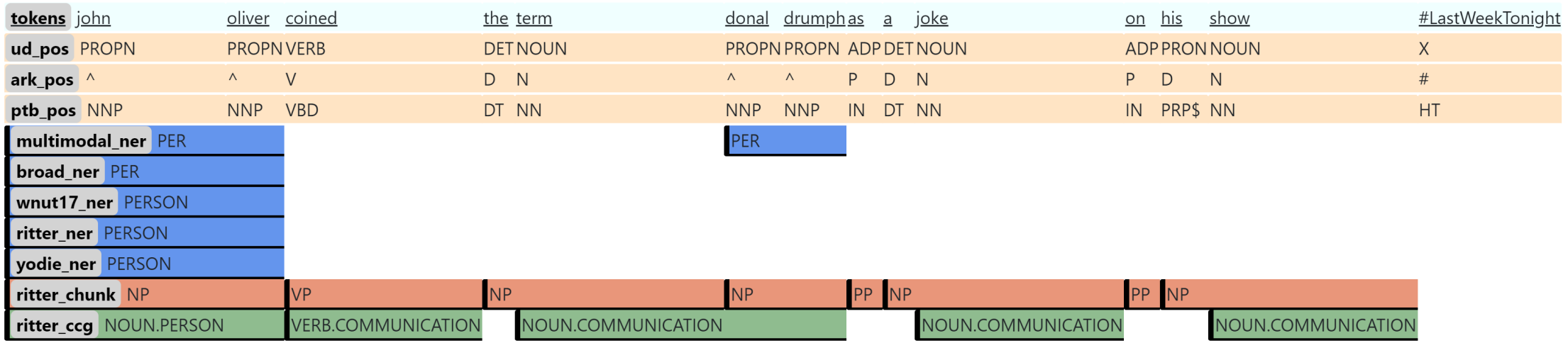
Web based UI https://github.com/socialmediaie/SocialMediaIE

Input

john oliver coined the term donal drumph as a joke on his show #LastWeekTonight

Predict

Output



Multi-task-multi-dataset learning - classification

data	split	tokens	tweets	vocab
Airline	dev	20079	981	3273
	test	50777	2452	5630
	train	182040	8825	11697
Clarin	dev	80672	4934	15387
	test	205126	12334	31373
	train	732743	44399	84279
GOP	dev	16339	803	3610
	test	41226	2006	6541
	train	148358	7221	14342
Healthcare	dev	15797	724	3304
	test	16022	717	3471
	train	14923	690	3511
Obama	dev	3472	209	1118
	test	8816	522	2043
	train	31074	1877	4349
SemEval	dev	105108	4583	14468
	test	528234	23103	43812
	train	281468	12245	29673

Sentiment classification

data	split	tokens	tweets	vocab
Founta	dev	102534	4663	22529
	test	256569	11657	44540
	train	922028	41961	118349
WaseemSRW	dev	25588	1464	5907
	test	64893	3659	10646
	train	234550	13172	23042

Abusive content identification

data	split	tokens	tweets	vocab
Riloff	dev	2126	145	1002
	test	5576	362	1986
	train	19652	1301	5090
Swamy	dev	1597	73	738
	test	3909	183	1259
	train	14026	655	2921

Uncertainty indicator classification

https://github.com/socialmediaie/SocialMediaIE

Sentiment classification results https://github.com/socialmediaie/SocialMediaIE

file	Airline		Clarin		GOP		Healthcare		Obama		SemEval	
model	**r**	**v**	**r**	**v**	**r**	**v**	**r**	**v**	**r**	**v**	**r**	**v**
S bilstm	8	80.46	8	65.71	5	67.05	6	63.88	9	59.0	9	65.57
MD bilstm	9	79.77	9	65.28	8	65.95	9	60.95	8	59.6	6	67.05
MTS bilstm	11	63.21	10	47.37	10	56.78	10	60.25	11	38.9	11	40.43
MTL bilstm	10	63.70	11	47.00	11	45.21	11	59.69	10	44.6	10	49.92
S bilstm *	6	81.69	3	**67.71**	3	**67.55**	3	**65.97**	1	**62.6**	7	66.47
MD bilstm *	5	81.85	7	66.23	7	66.50	4	64.85	3	**61.7**	3	**68.98**
MTS bilstm *	7	81.65	6	66.55	4	67.45	2	**66.81**	7	60.3	1	**69.52**
MTL bilstm *	2	**82.22**	4	67.60	2	**68.10**	1	**67.09**	6	61.3	2	**69.10**
S cnn *	3	**82.10**	1	**68.18**	1	**68.89**	8	62.34	1	**62.6**	8	66.19
MD cnn *	1	**82.54**	5	67.01	6	66.65	7	63.18	5	61.5	4	68.04
MTS cnn *	4	82.06	2	**67.72**	9	64.81	5	64.57	3	**61.7**	5	67.63

https://github.com/socialmediaie/SocialMediaIE

Abusive content identification

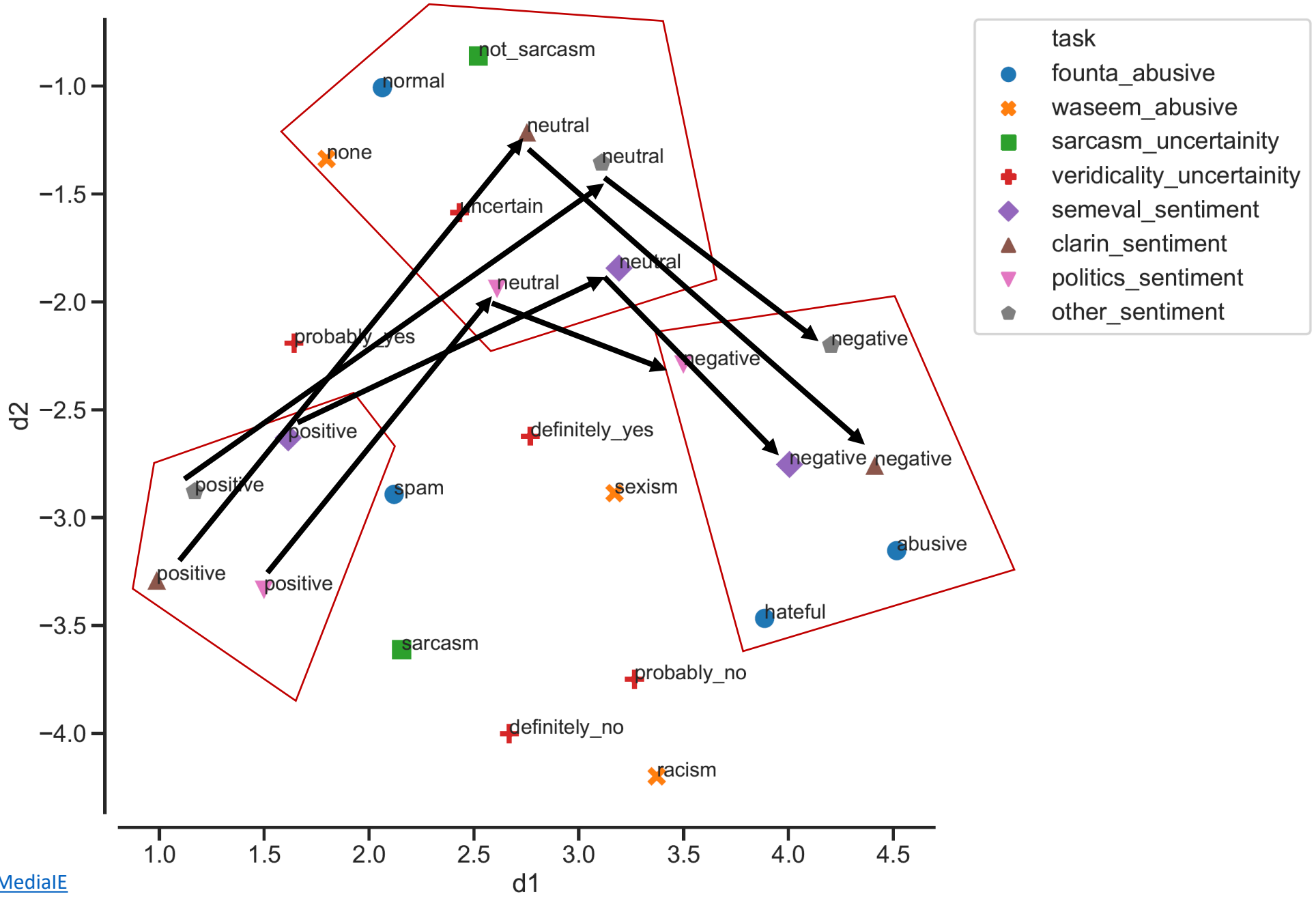
file	Founta		WaseemSRW	
model	r	v	r	v
S bilstm	8	79.33	8	81.72
MD bilstm	9	79.03	9	81.31
MTS bilstm	11	61.48	11	68.57
MTL bilstm	10	69.26	10	70.13
S bilstm *	1	**80.6**	3	**82.95**
MD bilstm *	2	**80.35**	2	**83.22**
MTS bilstm *	6	80.11	7	81.99
MTL bilstm *	4	80.23	5	82.78
S cnn *	3	**80.25**	4	82.89
MD cnn *	5	80.18	1	**84.42**
MTS cnn *	7	79.92	6	82.67

Uncertainty indicators

file	Riloff		Swamy	
model	r	v	r	v
S bilstm	6	81.22	5	38.80
MD bilstm	9	79.28	1	**39.34**
MTS bilstm	10	58.84	10	27.87
MTL bilstm	11	58.01	11	23.50
S bilstm *	3	**83.43**	1	**39.34**
MD bilstm *	7	80.94	1	**39.34**
MTS bilstm *	5	82.60	6	38.25
MTL bilstm *	2	**83.98**	1	**39.34**
S cnn *	1	**85.64**	7	35.52
MD cnn *	4	83.15	8	32.79
MTS cnn *	8	80.11	9	31.15

Label embeddings

- MDMT model learns similarity between labels without this knowledge being encoded in the model
- This leads to consistent relationship between similar labels across datasets



https://github.com/socialmediaie/SocialMediaIE

Web based UI https://github.com/socialmediaie/SocialMediaIE

Input

I know this tweet is late but I just want to say I absolutely fucking hated this season of
@GameOfThrones
what a waste of time.

Predict

Output

abusive

founta

abusive	hateful	normal	spam
0.830	0.084	0.085	0.002

waseem

none	racism	sexism
0.970	0.002	0.027

sentiment

clarin

negative	neutral	positive
0.956	0.036	0.008

other

negative	neutral	positive
0.906	0.063	0.031

politics

negative	neutral	positive
0.917	0.048	0.035

semeval

negative	neutral	positive
0.966	0.030	0.004

uncertainity

sarcasm

not sarcasm	sarcasm
0.914	0.086

veridicality

definitely no	definitely yes	probably no	probably yes	uncertain
0.033	0.244	0.112	0.189	0.422

Incremental learning of text classifiers with human-in-the-loop

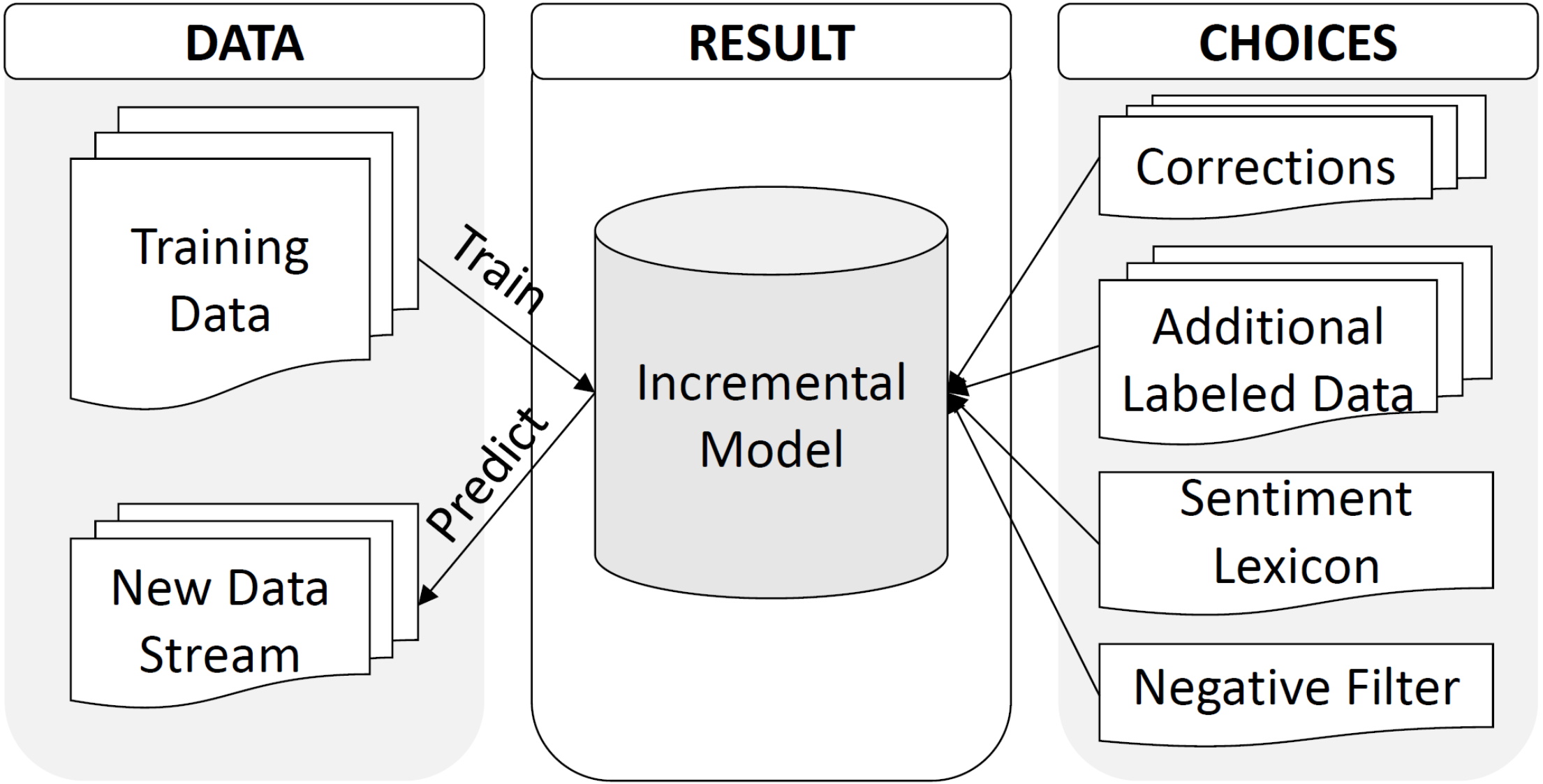
- Given a large unlabeled corpus, can we label it efficiently using fewer human annotations?
- Can existing models be updated efficiently to work with new data?
- Proposal:
  - Use active learning for data labeling
  - Use incremental learning algorithms for model updates
- Highly application to social media data:
  - Streaming data
  - Model should adapt to new data

Mishra, Shubhanshu, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. "Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 323–25. New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022.

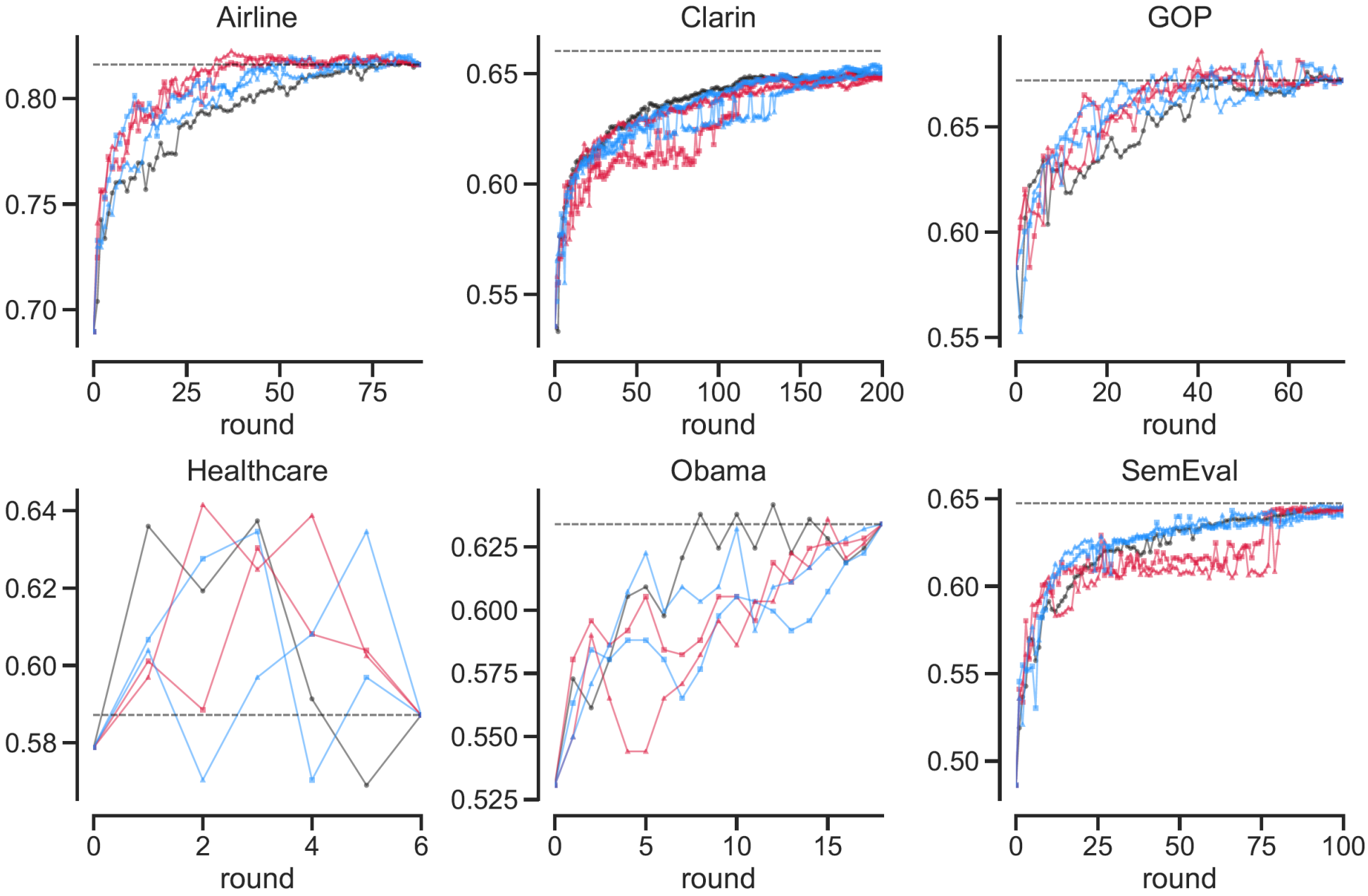
Active Learning

1. Given a model and unlabeled data
2. Select samples from the unlabeled data to be annotated, based on selection criterion
3. Update model with collected labeled examples
4. Repeat steps 2 to 3 till desired accuracy is reached or data exhausted

Mishra et al. (2015)



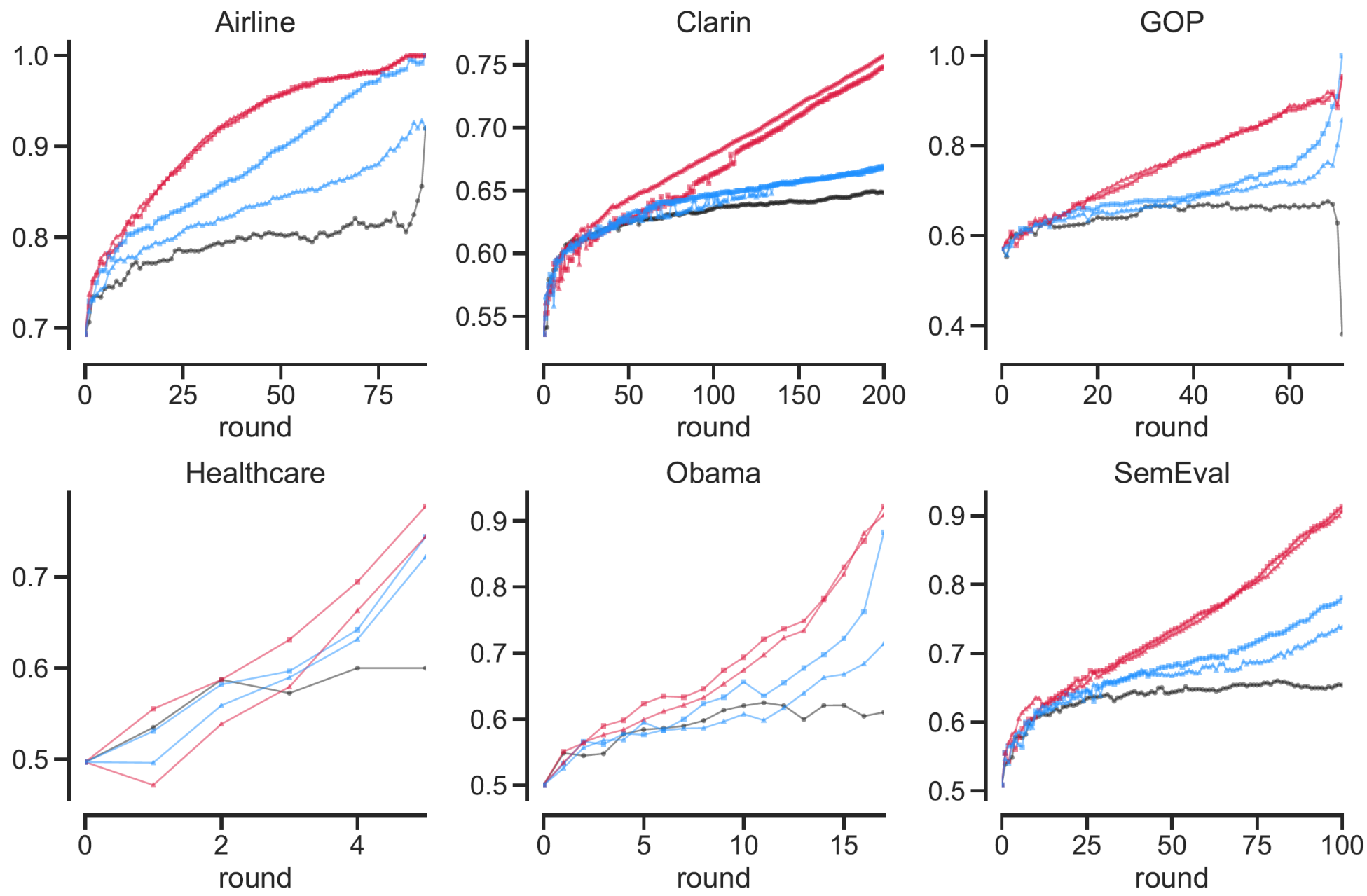
Mishra et al. (2015)



- Each round query 100 samples
- Classifier is logistic regression with unigram and lexicon features
- Max rounds is 100 (except Clarin)

Data ordered alphabetically and X and Y axes are not shared.

https://github.com/socialmediaie/SocialMediaIE



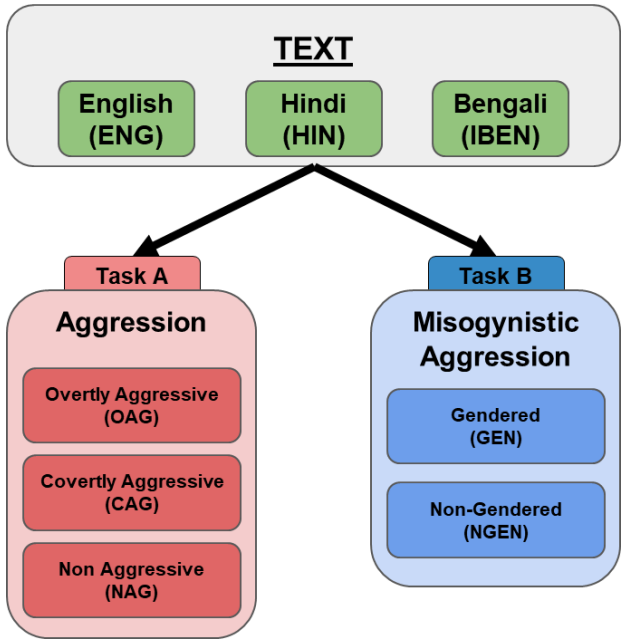
- Evaluate only on the data not used for training
- Top strategy queries efficiently and can help in labeling full data more quickly.

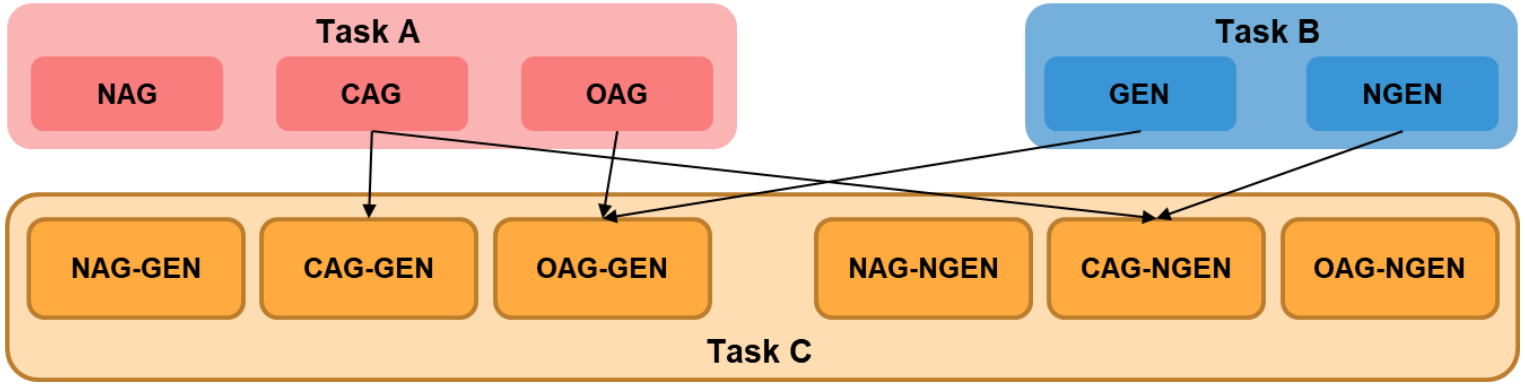
Data ordered alphabetically and X and Y axes are not shared.

https://github.com/socialmediaie/SocialMediaIE

Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020

https://github.com/socialmediaie/TRAC2020





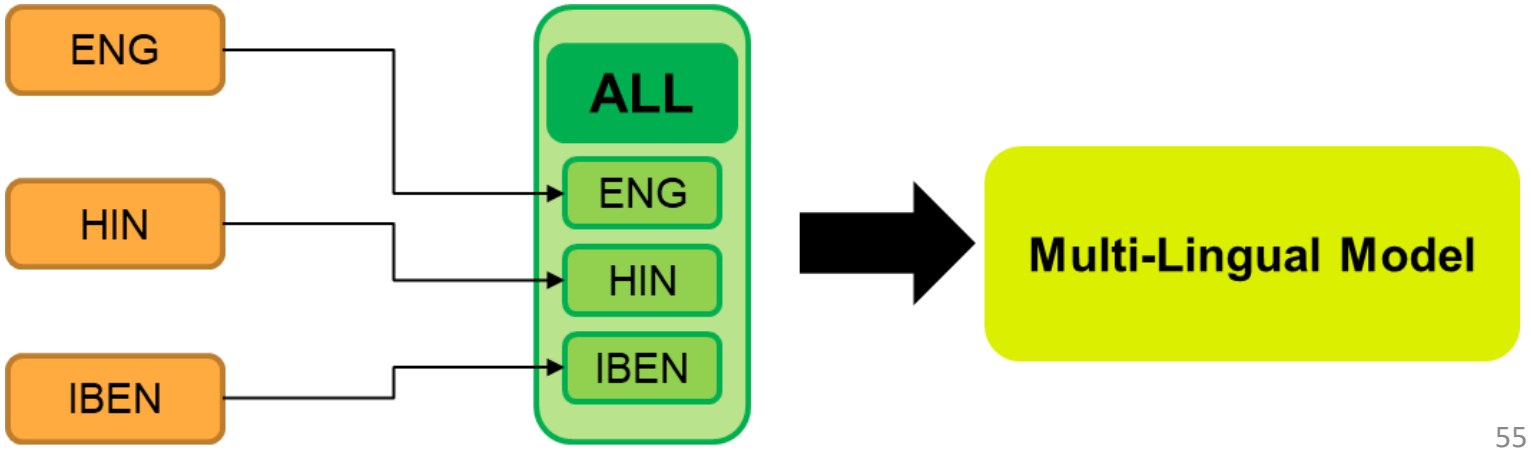
$$P(\text{NAG}) = P(\text{NAG-GEN}) + P(\text{NAG-NGEN})$$

2nd in 1/6 sub-tasks: ENG A

3rd in 3/6 sub-tasks: HIN A, B, and IBEN B

4th in 1/6 sub-tasks: ENG B

Computationally faster and cheaper inference cost.



Multilingual learning for hate speech detection

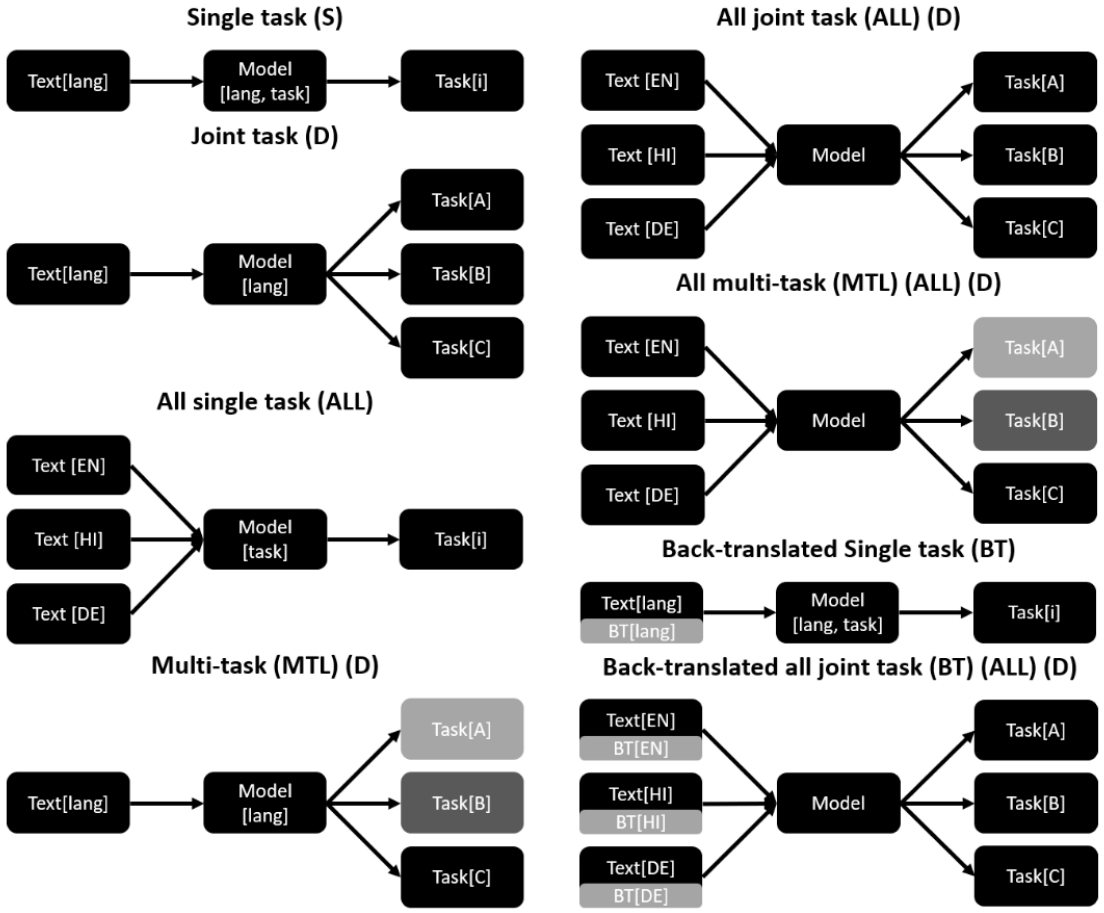
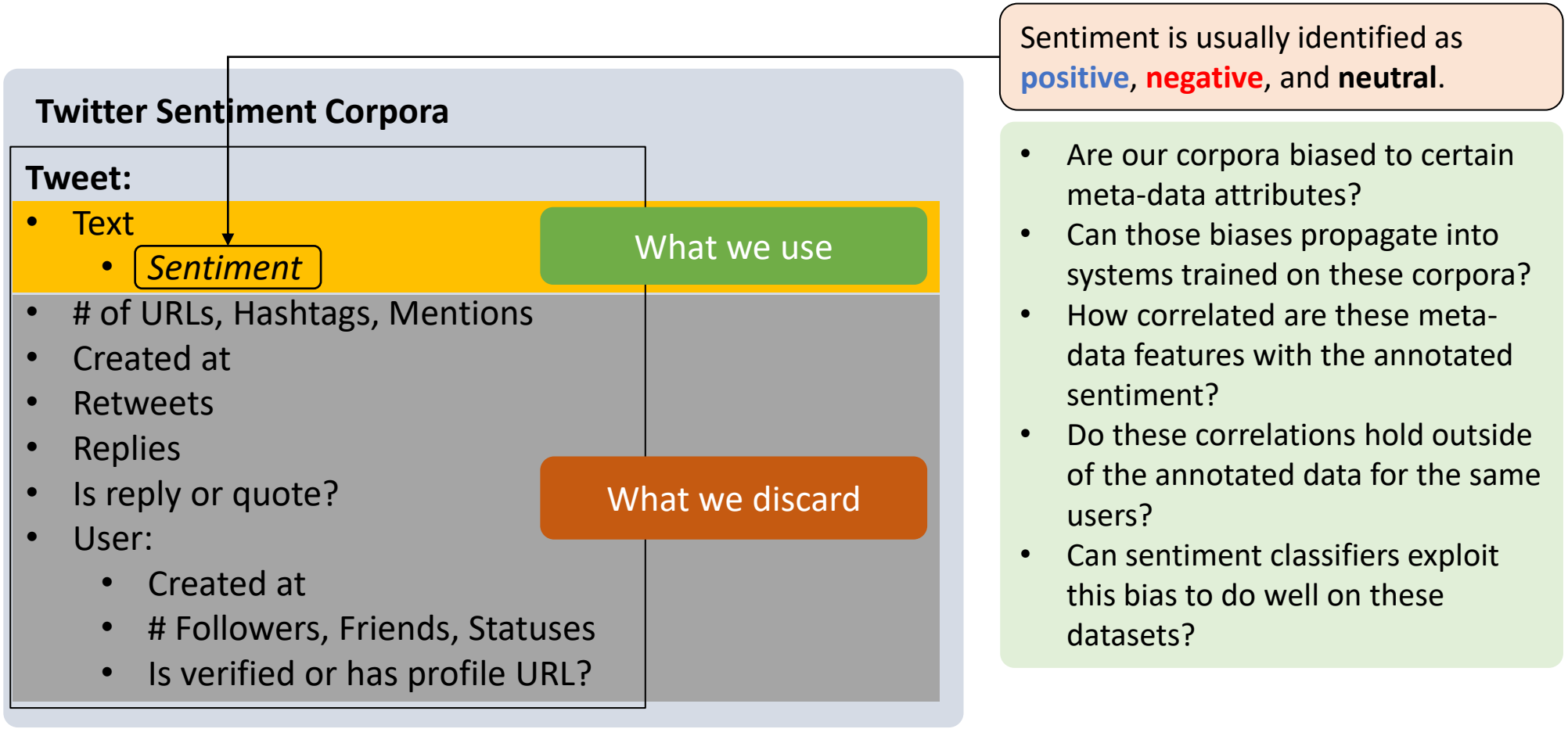


Fig. 2: An overview of various model architectures we used. Shaded task boxes represent that we first compute a marginal representation of labels only belonging to that task before computing the loss.

Mishra, S., Prasad, S. & Mishra, S. Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media. SN COMPUT. SCI. 2, 72 (2021). https://doi.org/10.1007/s42979-021-00455-5

Code: https://github.com/socialmediaie/MTML_HateSpeech

Improving sentiment classification using user and tweet metadata



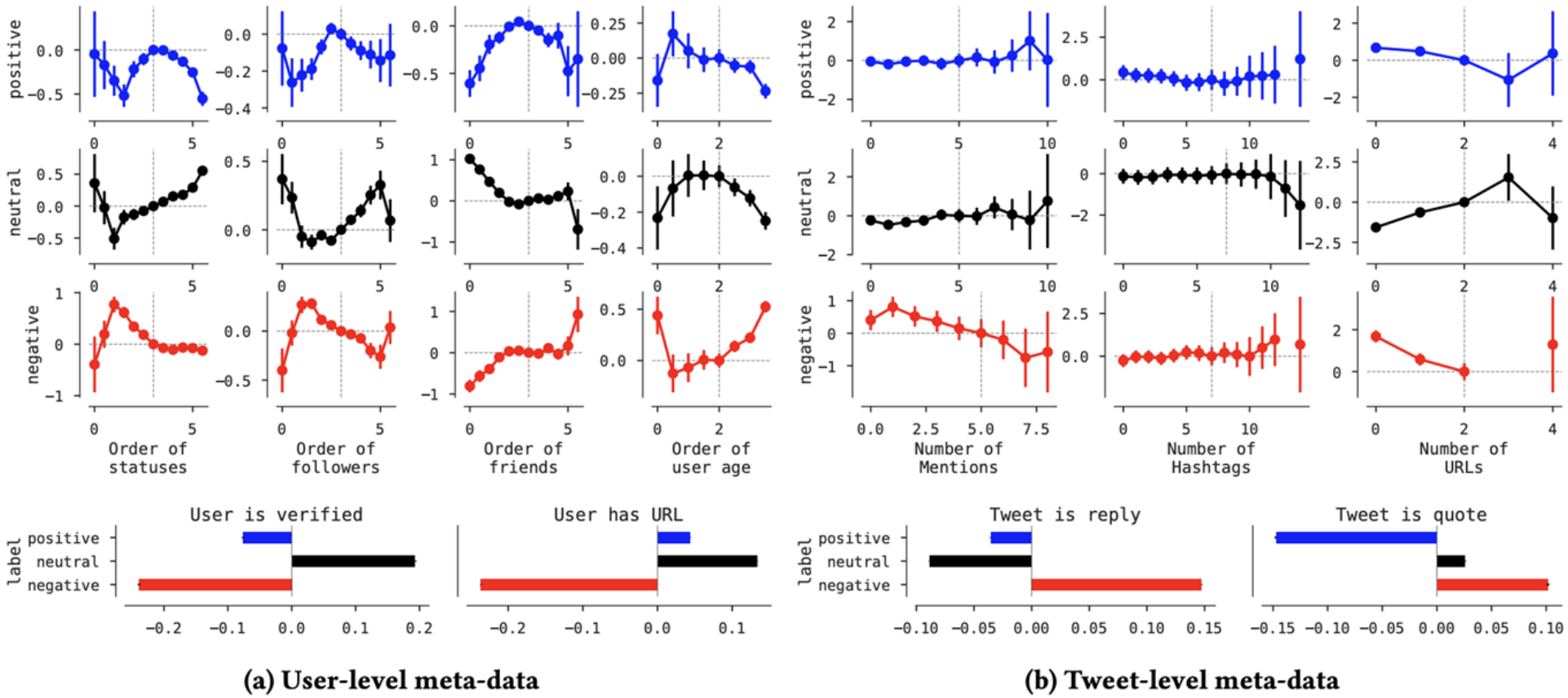
Mishra, S., & Diesner, J. (2018, July 3). Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. Proceedings of the 29th on Hypertext and Social Media. HT '18: 29th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3209542.3209562

Types of metadata and what they quantify

Quantification	User metadata
Activity level	# Statuses
Social Interest of the user	# Friends
Social status	# Followers
Account age	# days since account creation to posted tweet
Profile authenticity	Presence of URL on the profile or if the profile is verified

Quantification	Tweet metadata
Topical variety	# hashtags
Reference to sources	# URLs
Reference to network	# user mentions
Part of conversation	Is reply
Reference to conversation	Is quote

User metadata v/s Sentiment



(a) User-level meta-data

(b) Tweet-level meta-data

Figure 3: Meta-data features vs. sentiment classes. Y-axis in top plots and X-axis in bottom plots, is log-odds ratio, with respect to point at dashed lines.

Using metadata features can improve sentiment classification

Dataset	Model	Acc.	P	R	F1	KLD
Airline	meta	63.9	61.1	36.8	32.8	0.663
	text	80.0	78.3	69.0	72.4	0.026
	joint	80.3	76.6	72.0	**74.0**	0.005
Clarin	meta	45.7	42.1	40.9	37.8	0.238
	text	64.1	64.5	62.2	62.9	0.012
	joint	64.1	64.0	63.0	**63.4**	0.000
GOP	meta	59.9	54.3	37.5	33.6	0.776
	text	66.4	63.7	51.4	53.6	0.111
	joint	65.6	59.9	56.5	**57.8**	0.006
Healthcare	meta	56.7	36.8	39.4	35.1	0.717
	text	64.2	71.3	49.5	51.0	0.233
	joint	65.6	61.6	58.3	**59.5**	0.007
Obama	meta	39.3	37.0	35.1	32.0	0.282
	text	61.5	64.8	59.7	60.9	0.030
	joint	62.3	63.2	61.6	**62.2**	0.002
SemEval	meta	47.0	31.0	36.2	33.0	0.845
	text	65.5	64.1	58.0	59.5	0.032
	joint	65.6	62.7	60.5	**61.4**	0.001

Boost in F1 is mostly due to better recall. Precision is lower.

MESC might be helping with tweets with high OOV rates, where text classifiers don't do well.

Hands on session using SocialMediaIE

Links to install instructions and google colaboratory notebook at:

https://socialmediaie.github.io/tutorials/WWW2021

List of social media IE tools

- SocialMediaIE - https://github.com/socialmediaie/SocialMediaIE
- TwitterNER - https://github.com/socialmediaie/TwitterNER (more lightweight NER focused on English tweets)
- Social Communication Temporal Graph - https://github.com/napsternxg/social-comm-temporal-graph/ (visualizing temporal networks)
- ConText - https://github.com/uiuc-ischool-scanr/ConText (generate networks from text data)
- SAIL - https://github.com/uiuc-ischool-scanr/SAIL (active learning for text classification, python version coming soon at https://github.com/socialmediaie/)

Using SocialMediaIE for IE from text

- Notebook link: https://colab.research.google.com/drive/1ptfxPMGBsvsSRzoas7_qD2D_m1WrbbNp?usp=sharing
- Use one multi-task model to extract POS, named entities, chunks, and super-sense tags from text efficiently
- Use one multi-task model to label sentiment, abusive content, and uncertainty (sarcasm and veridicality) from text efficiently
- Copy the model output JSON to our UI interface https://codepen.io/napsternxg/full/YzwRqEb to see visual representation of the labels
- Try on your own text data
- Try to run SocialMediaIE on your local machine

Other models for multi-task learning

- Hierarchical labels or multi-label settings
  - Mishra, S., Prasad, S., & Mishra, S. (2020). Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 120–125). Marseille, France: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.trac-1.19. Code: https://github.com/socialmediaie/TRAC2020
  - Mishra, S., & Mishra, S. (2019). 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages. In *FIRE (Working Notes)* (pp. 208-213). Retrieved from http://ceur-ws.org/Vol-2517/T3-4.pdf. Code: https://github.com/socialmediaie/HASOC2019

Visualize temporal network of social media data in your browser

- Social Communication Temporal Graph: https://shubhanshu.com/social-comm-temporal-graph/
- Recent tweet comparison – Compare user-tweet network on tweets about 2 search queries
- Recent Tweet Sentiments – Compare user and tweet level sentiment on tweets about a single search query
- Wikipedia Revisions – Compare Wikipedia edit activity across 2 pages and identify common users

Thank you

- Questions
- Tweet to us at:
  - Shubhanshu Mishra - @TheShubhanshu
  - Rezvaneh (Shadi) Rezapour - @shadi_rezapour
  - Jana Diesner - @janadiesner @DiesnerLab
- All material presented here can be found at: https://socialmediaie.github.io/tutorials/WWW2021/
- If you have questions or feature requests about any of the tools open an issue on github e.g. for SocialMediaIE at: https://github.com/socialmediaie/SocialMediaIE/issues

References

- Diesner, J. (2015) Small decisions with big impact on data analytics. *Big Data & Society, special issue on Assumptions of Sociality, 2*(2). doi: 10.1177/2053951715617185
- Diesner, J. (2013). From Texts to Networks: Detecting and managing the impact of methodological choices for extracting network data from text data. *Kuenstliche Intelligenz Journal (Artificial Intelligence), 27*(1), 75-78.
- Mishra, Shubhanshu (2019): Trained models for multi-task multi-dataset learning for text classification as well as sequence tagging in tweets. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-1094364_V1
- Mishra, Shubhanshu (2019): Trained models for multi-task multi-dataset learning for text classification in tweets. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-1917934_V1
- Mishra, Shubhanshu (2019): Trained models for multi-task multi-dataset learning for sequence prediction in tweets. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-0934773_V1
- Shubhanshu Mishra. 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19). ACM*, New York, NY, USA, 283-284. DOI: https://doi.org/10.1145/3342220.3344929

References

- Mishra, Shubhanshu, & Diesner, Jana (2016). Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (pp. 203–212). Osaka, Japan: The COLING 2016 Organizing Committee.* Retrieved from https://aclweb.org/anthology/papers/W/W16/W16-3927/
- Mishra, Shubhanshu, Diesner, Jana, Byrne, Jason, & Surbeck, Elizabeth (2015). Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15 (pp. 323–325)*. New York, New York, USA: ACM Press. https://doi.org/10.1145/2700171.2791022
- Rezapour, Rezvaneh., Wang, Lufan., Abdar, Omid., & Diesner, Jana. (2017). Identifying the Overlap Between Election Result and Candidates' Ranking based on Hashtag-enhanced, Lexicon-based Sentiment Analysis. *Proceedings of IEEE 11th International Conference on Semantic Computing (ICSC), (pp. 93-96),* San Diego, CA. doi: 10.1109/ICSC.2017.92
- Rezapour, Rezvaneh., Shah, Saumil., & Diesner, Jana. (2019) Enhancing the Measurement of Social Effects by Capturing Morality. Proceedings of *the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, MN. [pdf]
- Sarol, M. Janina., Dinh, Ly., Rezapour, Rezvaneh., Chin, Chieh-Li., Yang, P.ingjing, & Diesner, Jana. (2020, November). An Empirical Methodology for Detecting and Prioritizing Needs during Crisis Events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 4102-4107). [pdf]

References

- Santosh, R., Schwartz, H. A., Eichstaedt, J. C., Ungar, L. H., & Guntuku, S. C. (2020). Detecting Emerging Symptoms of COVID-19 using Context-based Twitter Embeddings. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. [pdf]
- Dunn, J., Coupe, T., & Adams, B. (2021). Measuring linguistic diversity during covid-19. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 1–10 Online*. [pdf]
- Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., & Singh, S. (2020, December). COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. [pdf]
- Hardage, D., & Najafirad, P. (2020). Hate and Toxic Speech Detection in the Context of Covid-19 Pandemic using XAI: Ongoing Applied Research. [pdf]
- Biester, L., Matton, K., Rajendran, J., Provost, E. M., & Mihalcea, R. Quantifying the Effects of COVID-19 on Mental Health Support Forums. [pdf]