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

Shubhanshu Mishra* (Senior ML Researcher, Twitter, Inc)
Rezvaneh (Shadi) Rezapour (Assistant Professor, Drexel University, College of Computing and Informatics)
Jana Diesner, (Associate Professor, University of Illinois at Urbana-Champaign)

***Some of the work presented here was done during my PhD at UIUC**
Work done at twitter will be marked with Twitter logo.
Content and views expressed in this tutorial are solely the responsibility of the presenters.

https://socialmediaie.github.io/tutorials/CIKM2022/

QnA Page: https://slido.com with #3287167

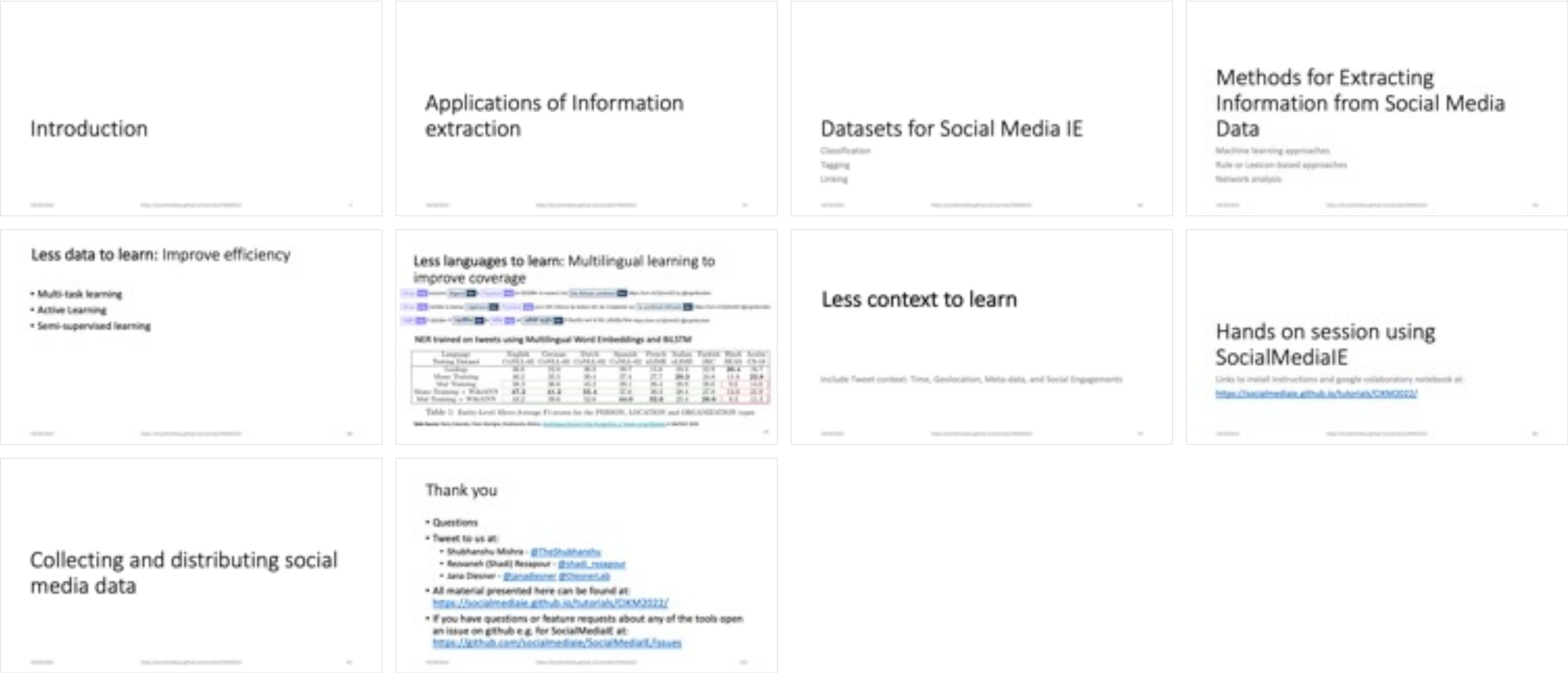
ILLINOIS
School of Information Sciences

DREXEL UNIVERSITY
College of
Computing & Informatics

Agenda

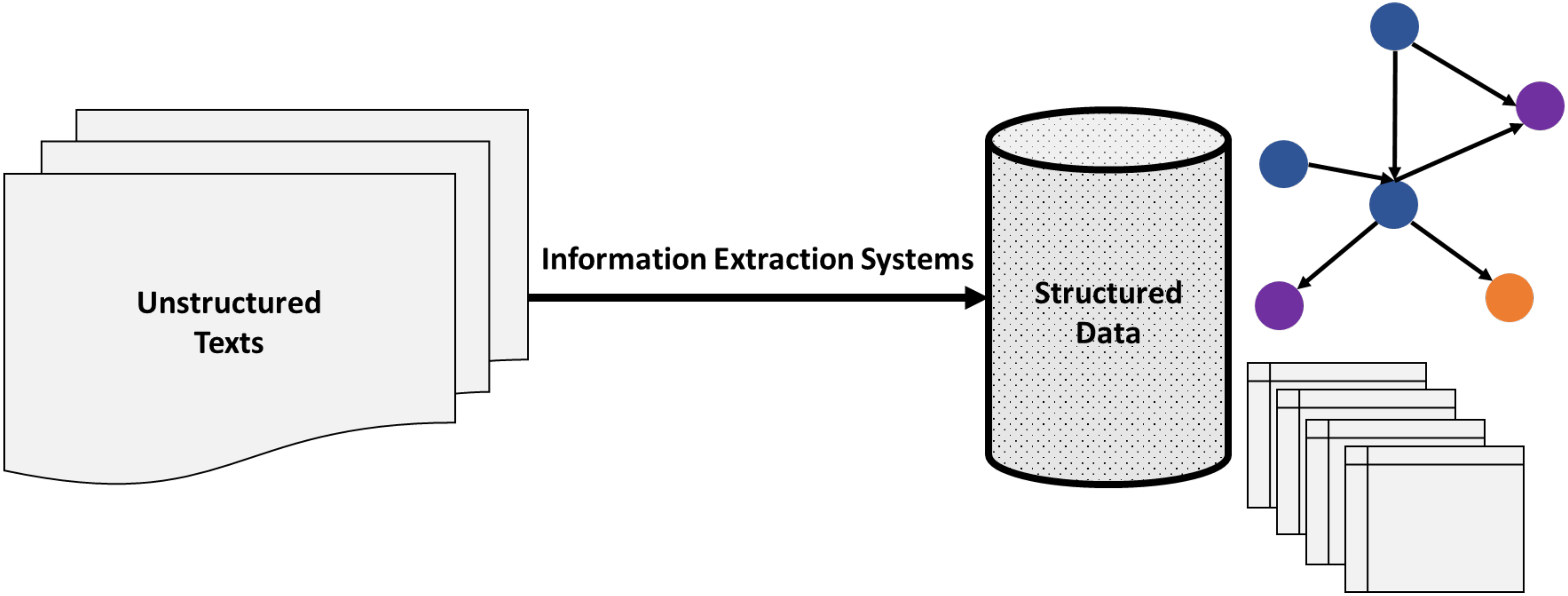
- 09:30 AM CST - Setup and Introduction (1 hr) – Shubh
- 10:30 AM CST - Applications of information extraction (30 mins) - PART 1 - Shubh or Shadi
- 11:00 AM CST - Applications of information extraction (30 mins) - PART 2 - Shubh or Shadi
- **11:30 AM CST - Lunch Break (1.5 hrs)**
- 01:00 PM CST - Datasets
- 01:30 PM CST – Hands On: Improving IE on social media data via ML (1 hr) - PART 1 – Shubh
- **02:30 PM CST – Break (30 mins)**
- 03:00 PM CST - Hands On: Improving IE on social media data via ML (1 hr) - PART 2 - Shubh
- 04:00 PM CST - Collecting and distributing social media data (25) - Shubh and Jana
- 04:45 PM CST - Conclusion and future directions (5 mins) - Shubh, Shadi, and Jana

Overview


Introduction
Applications of Information extraction
Datasets for Social Media IE
Methods for Extracting Information from Social Media Data
Less data to learn: Improve efficiency
• Multi-task learning
• Active Learning
• Semi-supervised learning
Less languages to learn: Multilingual learning to improve coverage
Less context to learn
Hands on session using SocialMediaIE
Collecting and distributing social media data
Thank you
• Questions
• Tweet to us at:
https://socialmediaie.github.io/tutorials/CIKM2022/
https://github.com/socialmediaie/SocialMediaIE/issues

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)

Types of Text based Media

Chapter 1

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

1813 - Pride and Prejudice, by Jane Austen

India vs West Indies | In 1000th ODI, facile win for India against Windies

Amol Karhadkar

AHMEDABAD FEBRUARY 10, 2022 07:15 IST
UPDATED: FEBRUARY 10, 2022 07:15 IST

Chahal, Washington and skipper Rohit ensure a victory in historic 1000th ODI for India

Washington Sundar returned to international cricket in style, Yuzvendra Chahal proved his worth with his wristspin and Rohit Sharma marked his first hit as full-time ODI with a quickfire fifty to ensure a perfect outing during India's 1000th ODI on Sunday.

Once Washington and Chahal broke the backbone of West Indies middle order on a helpful Narendra Modi Stadium strip, despite Jason Holder playing a trademark innings in the latter half, West Indies could manage only 176 before being bowled out in the 44th over.

2022 - The Hindu

```
Vulphere @ Libera.Chat / #archlinux - HexChat
rver  Settings  Window  Help
a.org/show_bug.cgi?id=1749908 | Help out testing the AUR https://lists.archlinux.org/pipermail/a
                       again.
[11:11:13]  Namarrgon  sanchex: are you running iwd and nm at the same time?
[11:12:14]    sanchex  I am running nm, I don't know if iwd is also running
[11:12:35]  Namarrgon  did you configure nm to use iwd as the backend instead of wpa_supplicant?
[11:13:07]    sanchex  No
[11:13:11]  Namarrgon  then why is iwd running?
[11:13:36]          *  julia (~quassel@user/julia) has joined
[11:15:58]          *  DeepDayze has quit (Quit: Leaving)
[11:17:02]    sanchex  good question
[11:17:45]  Namarrgon  how did you install arch?
[11:18:08]  Namarrgon  you're the third one with this issue today
[11:18:23]          *  gehidore is curious too
[11:18:54]          *  cabo40 (~cabo40@189.217.81.59) has joined
```

2021 - Internet Relay Chat - Wikipedia

- *Work on farm Fri. Burning piles of brush WindyFire got out of control. Thank God for good naber He help get undr control Pants-BurnLegWound.* [redacted])
- *Boom! Ya ur website suxx bro*
  [redacted]
- *...dats why pluto is pluto it can neva b a star*
  [redacted]
- *michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.*
  [redacted]

2013 - Social Media/ Eisenstein NAACL-HLT

http client info

[redacted]@aero.iitkgp.ernet.in
Tue, 21 Mar 1995 01:33:55 -0500

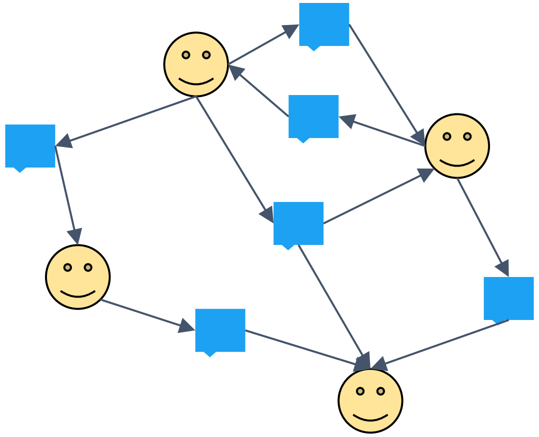
- **Messages sorted by:** [date][thread][subject][author]
- **Next message:** cyn@prism.nmt.edu: "Need help!"
- **Previous message:** jremick@u.washington.edu: "Where I am in here"

I have a running version of lynx here. I am unable to retrieve html documents. should I have a http daemon running
on my machine? Could you direct me to some FAQ on http programs and daemons
Thanks.
[redacted]

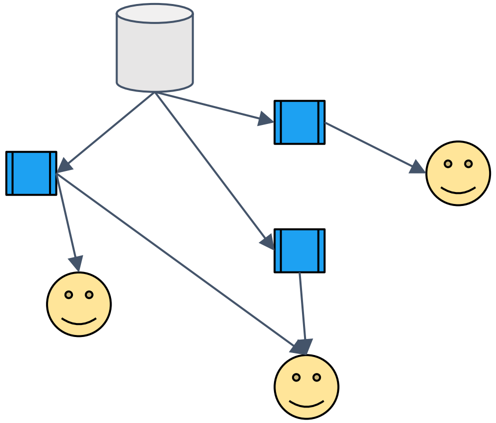
- **Next message:** [redacted]: "Need help!"
- **Previous message:** [redacted] "Where I am in here"

1995 - Usenet

Social Media v/s Traditional Media



Social Media



Traditional Media

"Many social media outlets **differ from traditional media** (e.g., print magazines and newspapers, TV, and radio broadcasting) in many ways, including **quality, reach, frequency, usability, relevancy, and permanence**. Additionally, social media outlets operate in a **dialogic transmission system, i.e., many sources to many receivers**, while **traditional media outlets operate under a monologic transmission model (i.e., one source to many receivers)**."

"For instance, a newspaper is delivered to many subscribers and a radio station broadcasts the same programs to an entire city."

"**User-generated content**—such as **text posts or comments**, digital photos or videos, and data generated through all online interactions — is the lifeblood of social media."

"Social media **helps the development of online social networks** by connecting a user's profile with those of other individuals or groups."

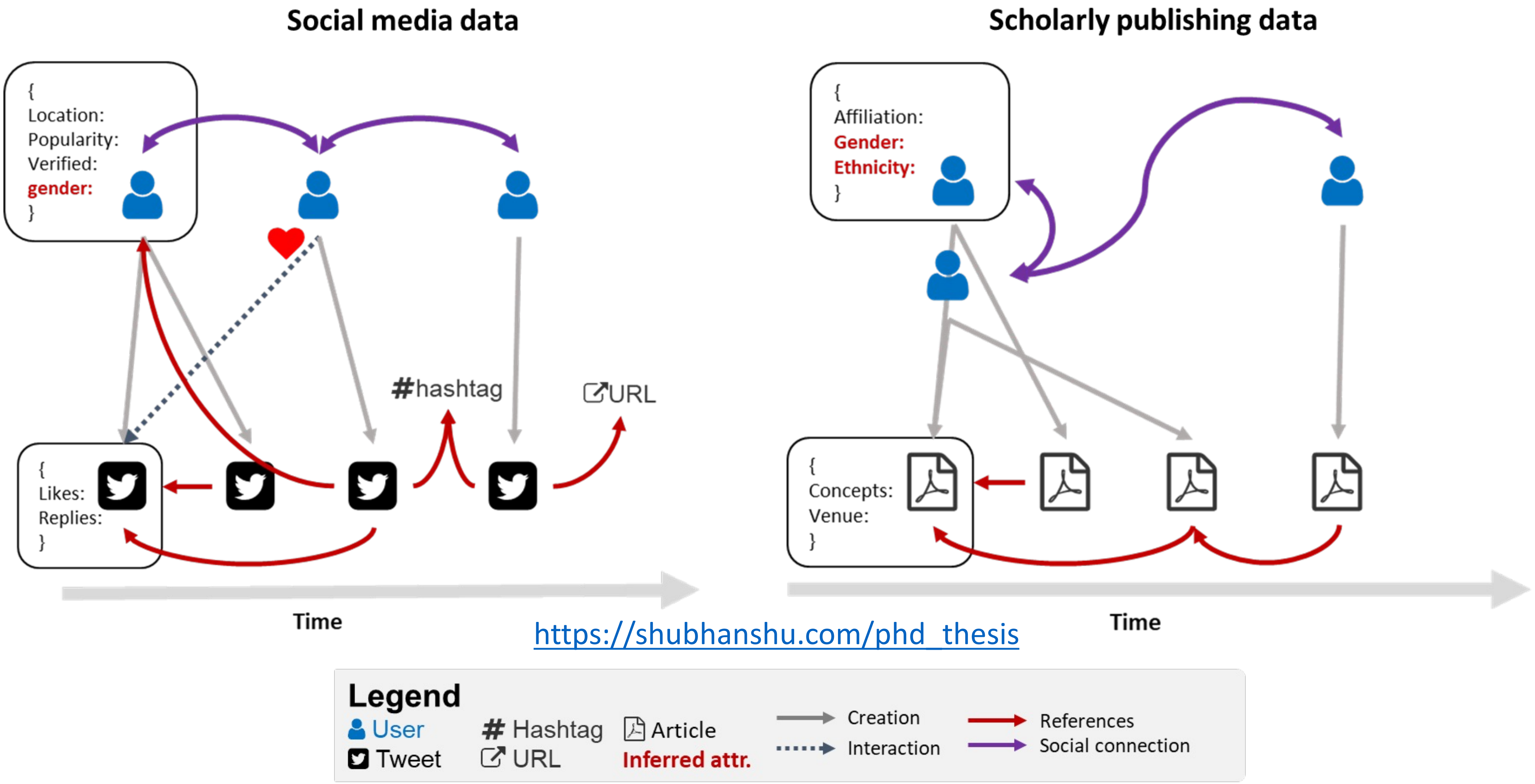
Source: [Social media - Wikipedia]

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

Corpus level

- Key-phrase extraction
- Taxonomy construction
- Topic modelling

Document level

Classification

- Sentiment
- Hate Speech
- Sarcasm
- Topic
- Spam detection
- Relation Extraction

Token level

Tagging

- Named entity
- Part of speech

Disambiguation

- Word Sense
- Entity Linking

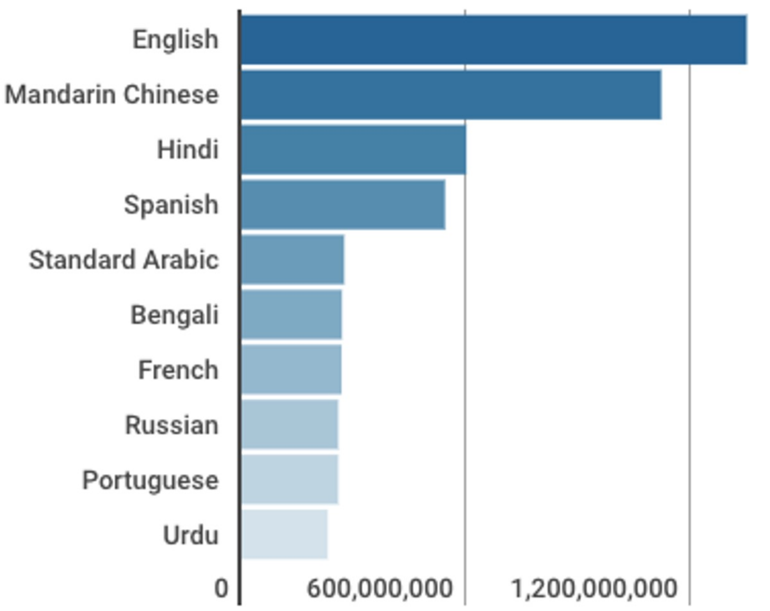
Why social media data is challenging?

Social Media text often has a inherent structure, which provides context, e.g.

- user mentions
- hashtags
- comment threads
- less formally written language
- lot of unseen words
- typos, etc.

Language Diversity

Top 10 most spoken languages, 2021



Ethnologue

Source: https://www.ethnologue.com/guides/ethnologue200

Languages		Regions	Participation			Active editors					Edits		Usage	Content
Code ⇒ Project Main Page	Language ⇒ Wikipedia article		Speakers in millions (log scale) (?) Editors per million speakers (5+ edits)	Prim.+Sec. Speakers M=millions k=thousands	Editors (5+) per million speakers	Months since 3 or more active editors	5+ edits p/month (3m avg)	100+ edits p/month (3m avg)	Admins	Bots	Bot edits	Human edits by unreg. users	Views per hour	Article count
Σ	All languages	AF AS EU NA SA OC CL W												
en	English	AF AS EU NA OC		1121 M	27		30684	3445	1274	312	9%	31%	4,858,539	5,779,516
ceb	Cebuano	AS		20 M	1		26	2	4	60	99%	19%	1,311	5,379,752
sv	Swedish	EU		10 M	64		641	101	66	40	57%	20%	53,206	3,761,531
de	German	EU		132 M	41		5395	900	198	374	10%	20%	726,852	2,254,737
fr	French	AF AS EU NA OC SA		285 M	17		4864	790	161	107	19%	21%	461,591	2,069,464
nl	Dutch	EU SA		28 M	42		1185	214	45	269	38%	19%	97,322	1,953,504
ru	Russian	AS EU		264 M	12		3188	518	87	84	17%	25%	634,782	1,518,909
es	Spanish	AF AS EU NA SA		513 M	8		4135	544	71	36	17%	37%	417,439	1,496,759
it	Italian	EU		68 M	35		2355	398	109	173	29%	32%	270,709	1,489,914
pl	Polish	EU		43 M	29		1256	237	106	68	34%	19%	185,774	1,313,943

Source: https://stats.wikimedia.org/EN/Sitemap.htm#comparisons

I am Japanese.

Source: https://tatoeba.org/eng/sentences/show/657403

Translations

- Ich bin Japaner.
- Ich bin Japanerin.
- Είμαι Γιαπωνέζα.
- Mi estas japanino.
- Mi estas japana.
- Olen japanilainen.
- Mä oon japanilainen.
- Je suis Japonais.
- אני יפני.
- אני יפנית.
- मैं जापानी हूं।
- Japán vagyok.
- Sono giapponese.
- Io sono giapponese.
- 私は日本人です。

NER performance difference

Named entity recognition performance over the evaluation partition of the Ritter dataset (best score in bold).

	Per-entity F1				Overall		
System	Location	Misc	Org	Person	P	R	F1
ANNIE	40.23	0.00	16.00	24.81	36.14	16.29	22.46
DBpedia Spotlight	46.06	6.99	19.44	48.55	34.70	28.35	31.20
Lupedia	41.07	13.91	18.92	25.00	38.85	18.62	25.17
NERD-ML	**61.94**	23.73	**32.73**	**71.28**	52.31	**50.69**	**51.49**
Stanford	60.49	**25.24**	28.57	63.22	**59.00**	32.00	41.00
Stanford-Twitter	60.87	25.00	26.97	64.00	54.39	44.83	49.15
TextRazor	36.99	12.50	19.33	70.07	36.33	38.84	37.54
Zemanta	44.04	12.05	10.00	35.77	34.94	20.07	25.49

Source: Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. Information Processing & Management, 51(2), 32–49.
https://doi.org/10.1016/j.ipm.2014.10.006

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

	negative	neutral	positive
clarin	0.956	0.036	0.008
other	0.906	0.063	0.031
politics	0.917	0.048	0.035
semeval	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	

Entity Linking

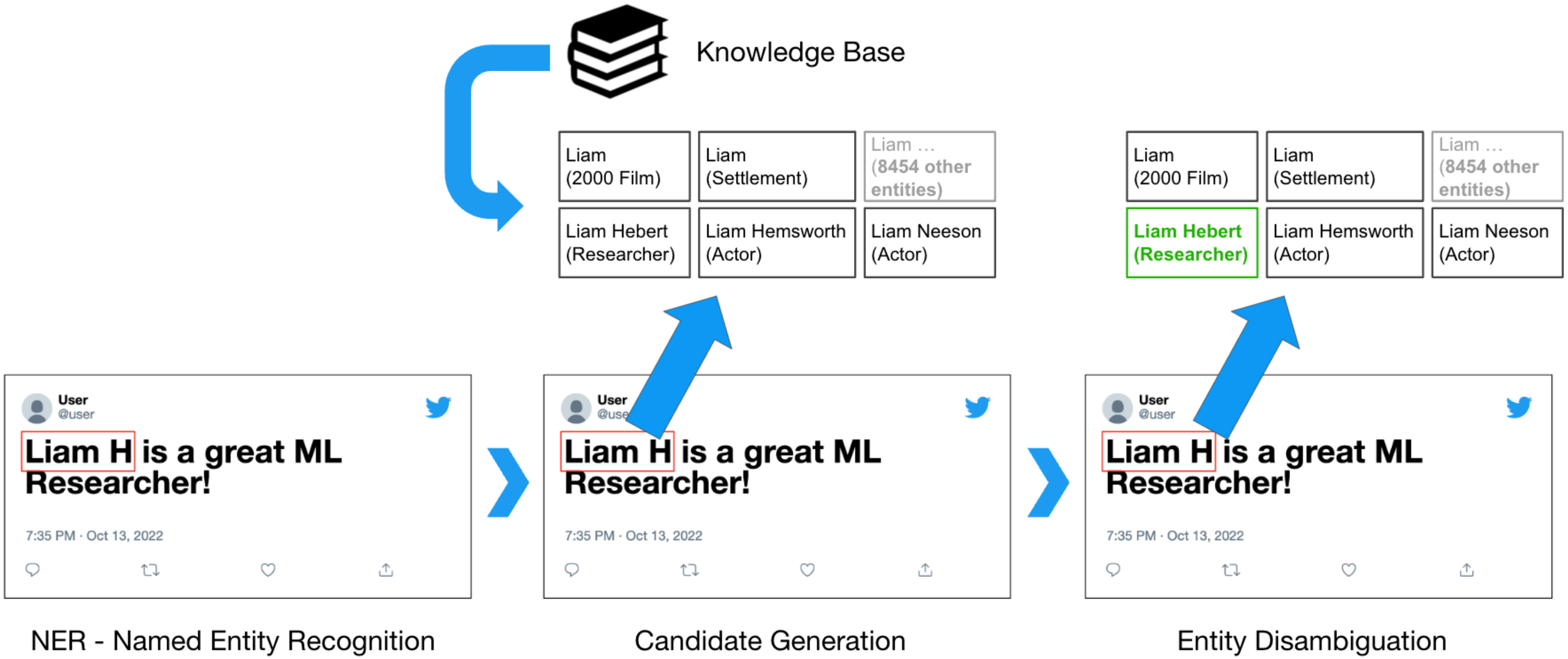


Figure Source: https://www.youtube.com/watch?v=Ug-d2tK3PDQ

Liam Hebert, Raheleh Makki, Shubhanshu Mishra, Hamidreza Saghir, Anusha Kamath, and Yuval Merhav. 2022. Robust Candidate Generation for Entity Linking on Short Social Media Texts. In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), pages 83–89, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Application of NER: Trends

Sonic The Hedgeblog
@Sonic_Hedgeblog

The Dreamcast was launched 20 years ago today, and the US release of 'Sonic Adventure'! Special DLC was available to celebrate the launch of the system. Touching some of them brings up this message. ift.tt/2PXJoMA

RPG Site
@RPGSite

Happy 20th North American birthday to the Dreamcast, which first hit NA on this day in 1999 - the famed 9/9/99. The machine launched with games including Sonic Adventure, Power Stone, House of the Dead 2 and Ready 2 Rumble Boxing.

2 · Trending
Dreamcast
46.8K people are Tweeting about this

Identifying trending topics and events

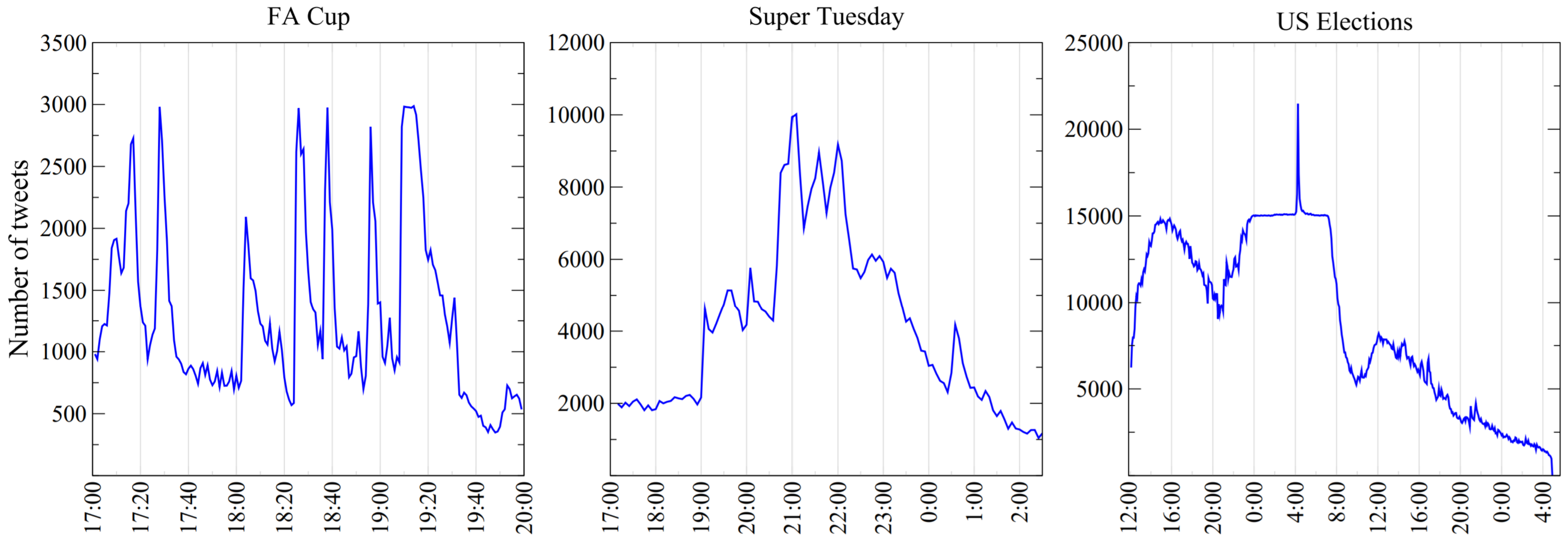
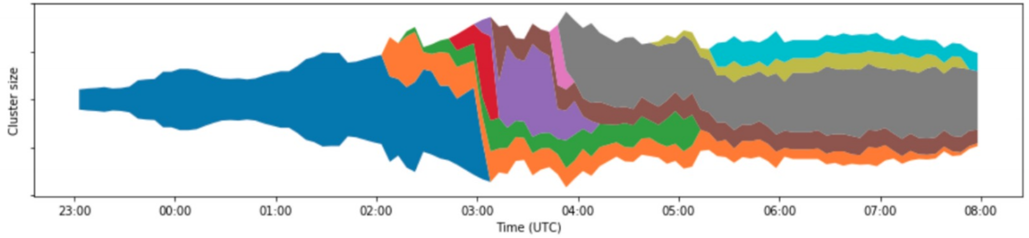


Fig. 2. Twitter activity during events. For the FA Cup, the peaks correspond to start and end of the match and the goals. For the two political collections, the peaks correspond to the main result announcements.

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.

Application of NER: Events Detection



Title	Top entities
General conversation	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Lady Gaga, Sandra Oh, Spider-Man: Into the Spider-Verse, Gaga
Hosts' opening speech	Andy Samberg, Black Panther, Sandra Oh, #blackpanther, Jim Carrey, Michael B. Jordan
Green Book	Green Book, Mahershala Ali, Regina King, #greenbook
Christian Bale receives the best actor in comedy or musical award for "Vice"	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Christian Bale, Sandra Oh, Lady Gaga, Darren Criss, Vice
General conversation	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Lady Gaga, Jeff Bridges, Darren Criss
Christian Bale thanks Satan in his acceptance speech	Christian Bale, The 76th Annual Golden Globe Awards 2019, Vice, Mitch McConnell, Satan
General conversation	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Sandra Oh, Alfonso Cuarón, Rami Malek, Roma, Olivia Colman
Rami Malek receives the best actor in a drama award for "Bohemian Rhapsody"	The 76th Annual Golden Globe Awards 2019, #goldenglobes, Rami Malek, Bohemian Rhapsody, Lady Gaga, Sandra Oh
Glenn Close receives the best actress in drama award for "The Wife"	Glenn Close, Taylor Swift, Lady Gaga, best actress, Glenn, Bradley Cooper
Green Book	Green Book, Mahershala Ali, Regina King, #greenbook

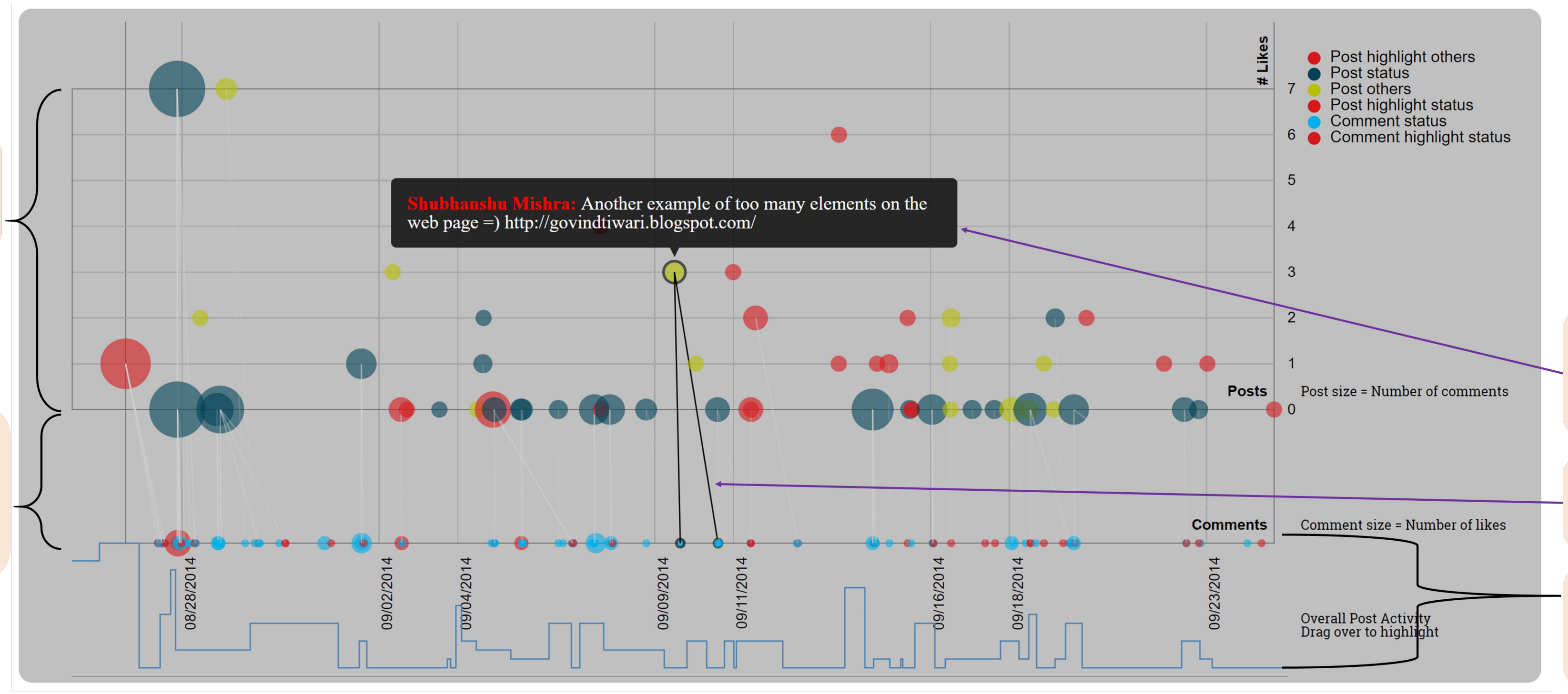
[Fedoryszak et al., 2019]

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

Application of NER: User Interest

Last Engagements

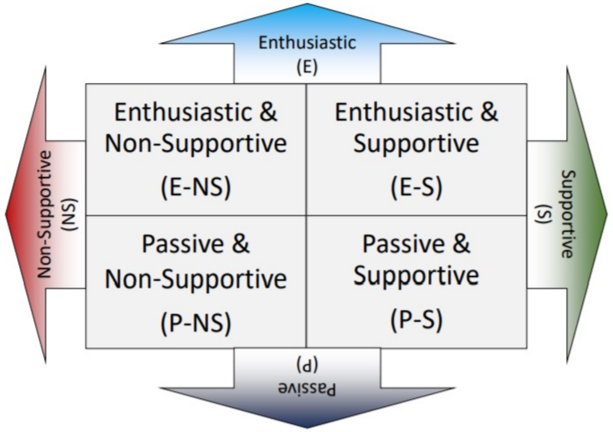
Twitter (9), India (9), US (7), Pilani (7), NASA (3),

Linkedin (3), Stanford CoreNLP (2)

BITS Pilani (1)

Person
Location
Organization
Product
Other

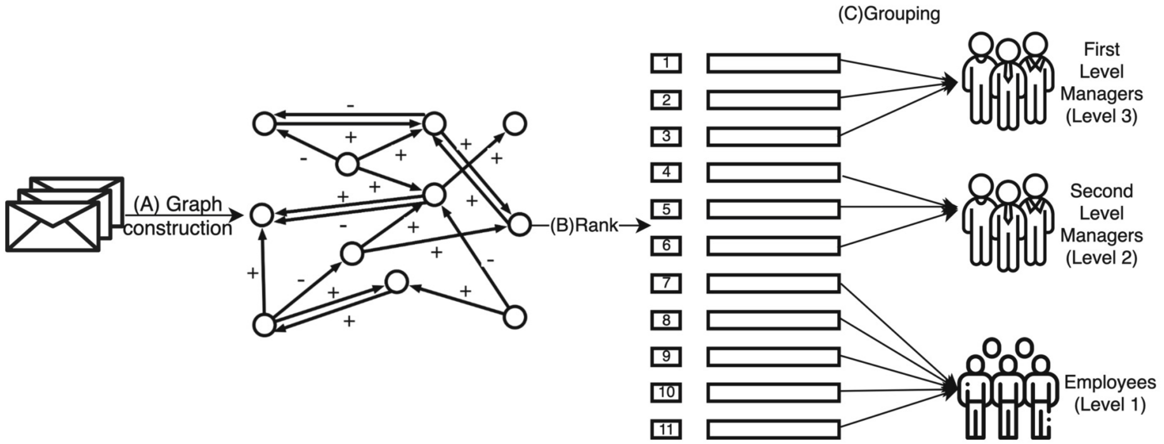
Network construction from text classification labels and identification of influential users



	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.

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

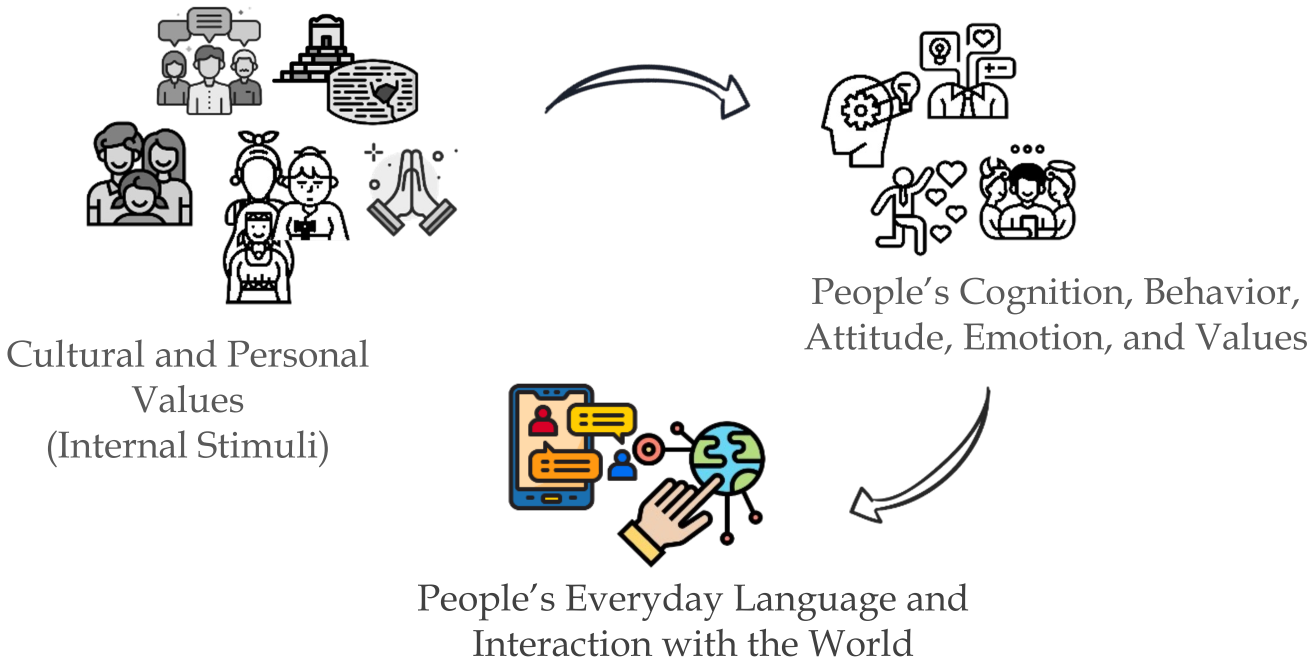
- Aim:
  - Test whether incorporating prevalent hashtags from a given dataset into a sentiment lexicon improves sentiment prediction accuracy
- Method:
  - Used hashtag-enhanced lexicon-based sentiment analysis to analyze tweets that mention the US Presidential candidates to find the correlation between 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)

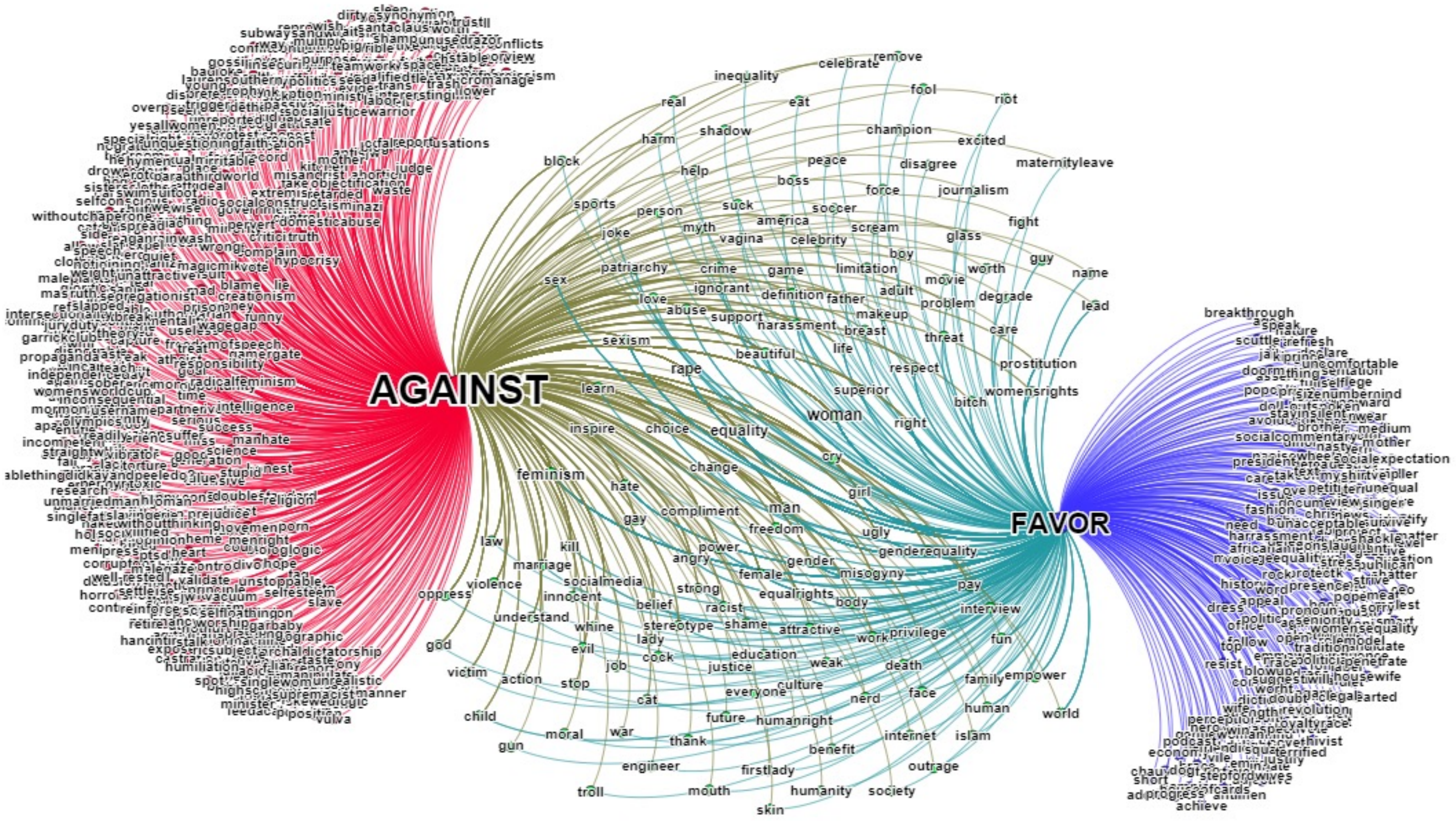


Using moral foundations analysis in analyzing social effects (contd.)

- 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, R., Dinh, L., & Diesner, J. (2021, August). **Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics.** In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 177-188).
Rezapour, Rezvaneh; Diesner, Jana (2019): **Expanded Morality Lexicon**. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-3805242_V1.1

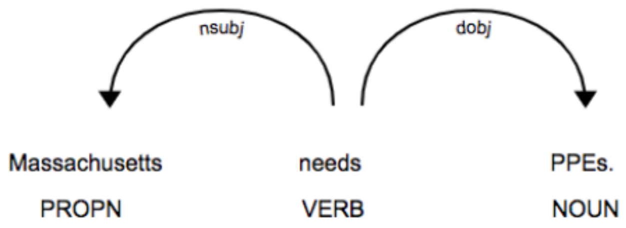
Analyzing tweets to examine cross-cutting exposure in social media



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)

Datasets for Social Media IE

Classification

Tagging

Linking

Classification benchmarks

- Twitter Sentiment Benchmarks - https://zenodo.org/record/2866988
- Social Media IE (explained next) – https://doi.org/10.5281/zenodo.5867160
- TweetEval - https://github.com/cardiffnlp/tweeteval
- SemEval 2017 - https://alt.qcri.org/semeval2017/task4/

Social Media IE: Classification data

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://doi.org/10.5281/zenodo.5867160 and https://shubhanshu.com/phd_thesis/

Tagging benchmarks

- Broad Twitter Corpus - https://github.com/GateNLP/broad_twitter_corpus
- Temporal Twitter Corpus - https://zenodo.org/record/3899040
- WNUT 2016 and 2017 shared tasks on Twitter NER - https://noisy-text.github.io/2017/emerging-rare-entities.html
- TweetNERD - End to End Entity Linking Benchmark for Tweets (largest benchmark for Tweets) - https://zenodo.org/record/6617192

Social Media IE: Tagging data

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

https://doi.org/10.5281/zenodo.5867160 and https://shubhanshu.com/phd_thesis/

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

TweetNERD – NER + Disambiguation

Largest Tweet Benchmark (To appear at Neurips 2022 Datasets and Benchmarks Track)

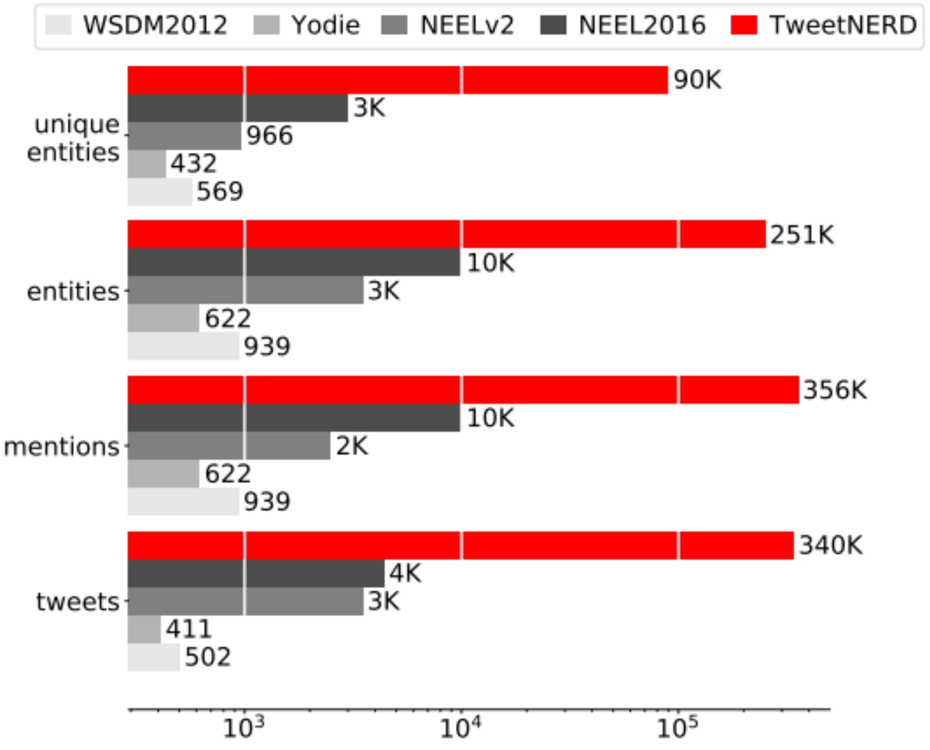
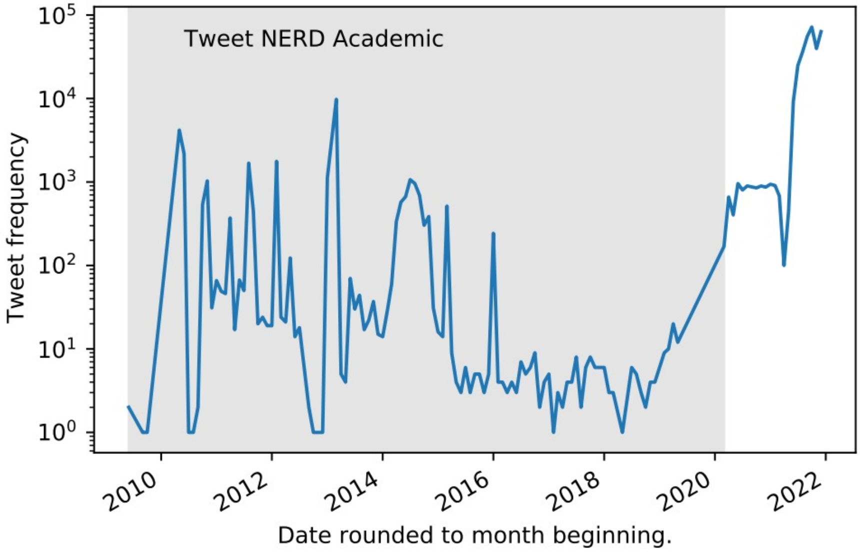


Figure 1: Comparison with existing Tweet entity linking datasets



Model	OOD	Academic
Spacy	0.377	0.454
StanzaNLP	0.421	0.503
SocialMediaIE	0.153	0.245
BERTweet WNUT17	0.278	0.46
TwitterNER	0.424	0.522
AllenNLP	0.454	0.552

(a) NER `strong_mention_match` F1 scores.

Model	`entity_match`		`strong_all_match`	
	OOD	Academic	OOD	Academic
GENRE	0.469	0.636	0.39	0.624
REL	0.463	0.614	0.387	0.56
Lookup	0.621	0.645	0.584	0.617

(b) Entity Linking given true spans (EL) F1 scores.

Model	`entity_match`		`strong_all_match`	
	OOD	Academic	OOD	Academic
DBpedia	0.292	0.399	0.231	0.347
NLAI	0.522	0.568	0.313	0.494
TAGME	0.402	0.431	0.293	0.381
REL	0.344	0.484	0.27	0.444
GENRE[4]	0.307	0.458	0.223	0.379

(c) End to End Entity Linking (End2End) F1 scores.

Shubhanshu Mishra, Aman Saini, Raheleh Makki, Sneha Mehta, Aria Haghighi, Ali Mollahosseini: "TweetNERD -- End to End Entity Linking Benchmark for Tweets", 2022; arXiv:2210.08129.

SocialMediaIE – MetaCorpus

A catalogue of ~500 social media datasets - https://github.com/socialmediaie/MetaCorpus

- Catalogue of datasets used in academic papers and benchmarks
- Around 500 social media datasets from Twitter, Facebook, Reddit, YouTube, Weibo, Gab, etc.
- Tasks comprise of: Classification, Tagging, NER, Entity Linking, Fact Checking, Paraphrasing, Machine Translation, Rumor detection, Conversation modeling, Retrieval and Sentence Similarity, Summarization
- Multi-lingual datasets
- Contribute your own datasets by sending a Pull Request at: https://github.com/socialmediaie/MetaCorpus

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

Key challenges for improving IE performance

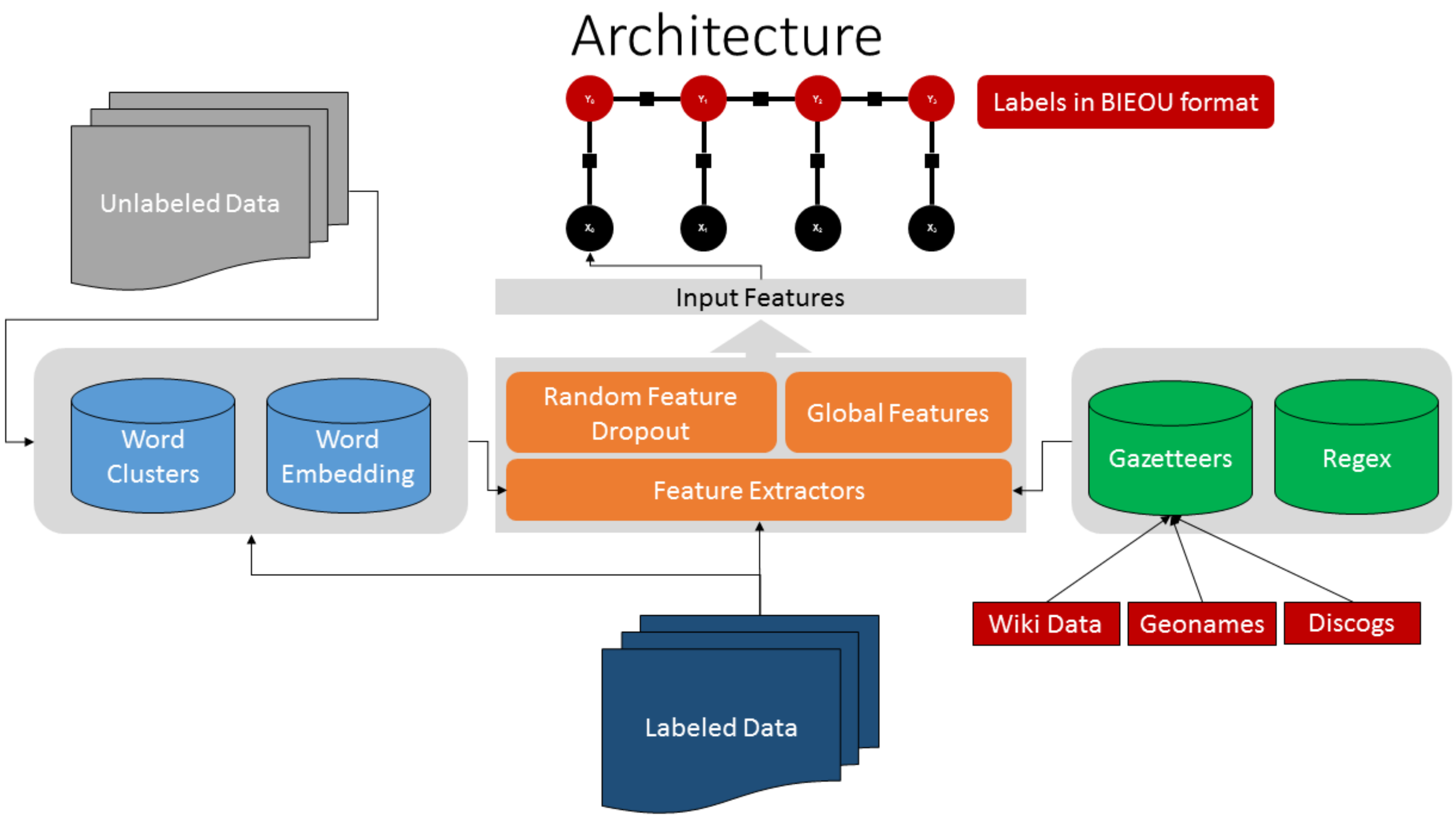
Challenge	Solution
Less data to learn	Multi-task learning, active learning, semi-supervised, or distantly supervised learning
Less languages to learn	Cross lingual alignment, Multilingual Knowledge bases
Less context to learn	Social and Graphical context of the tweet

Less data to learn: Improve efficiency

- Multi-task learning
- Active Learning
- Semi-supervised learning

Rule based Twitter NER

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

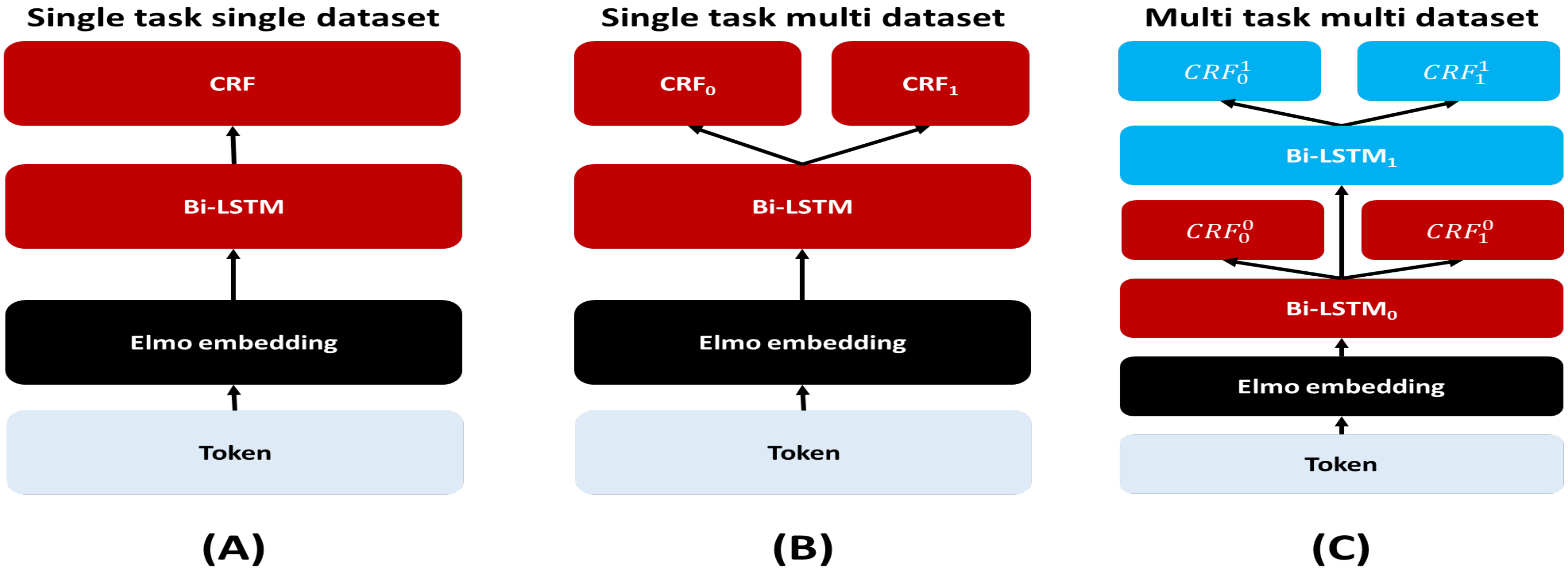


Mishra, Shubhanshu, & Diesner, Jana (2016). Semi-supervised Named Entity Recognition in noisy-text. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT) (pp. 203–212). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclweb.org/anthology/papers/W/W16/W16-3927/

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



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

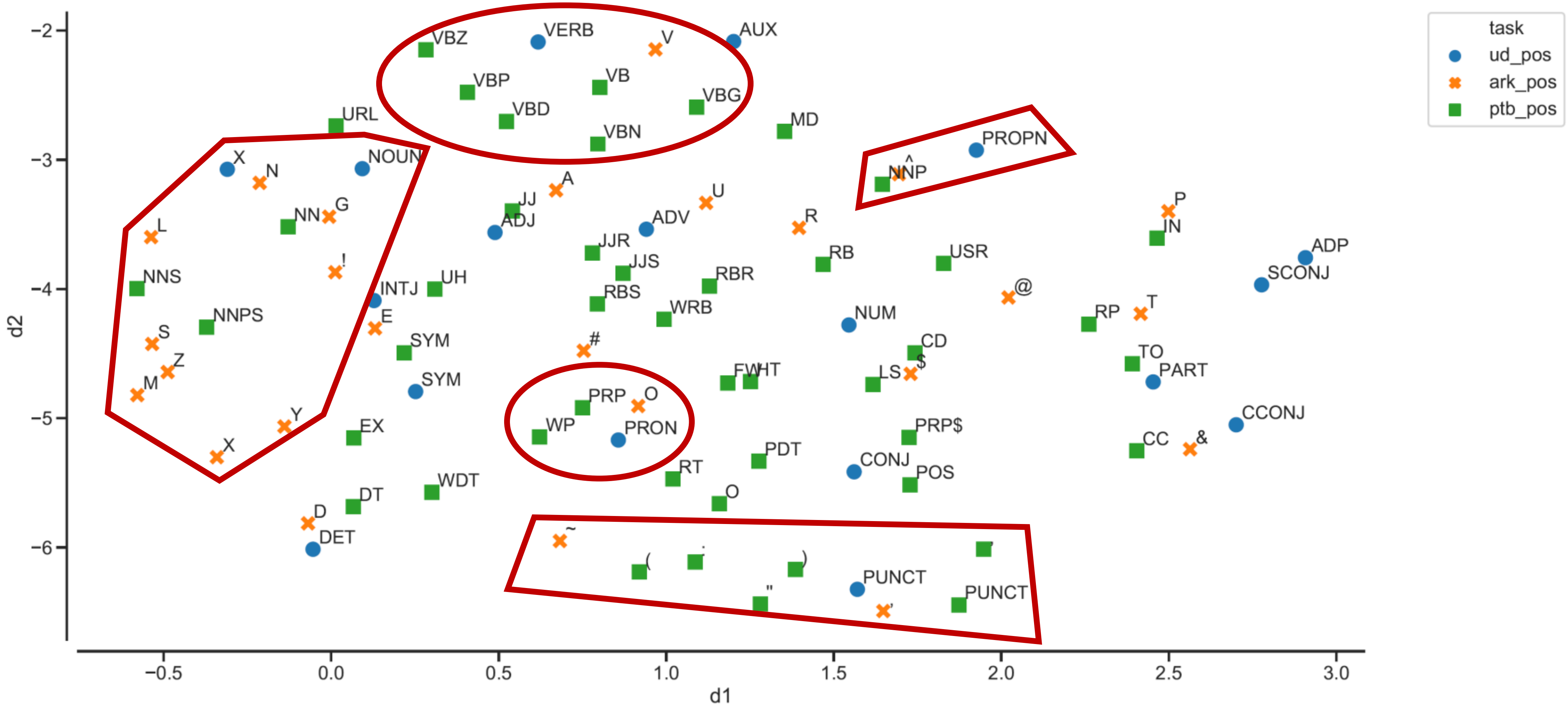
Training Mishra 2019, HT' 19

- 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 **"#"**

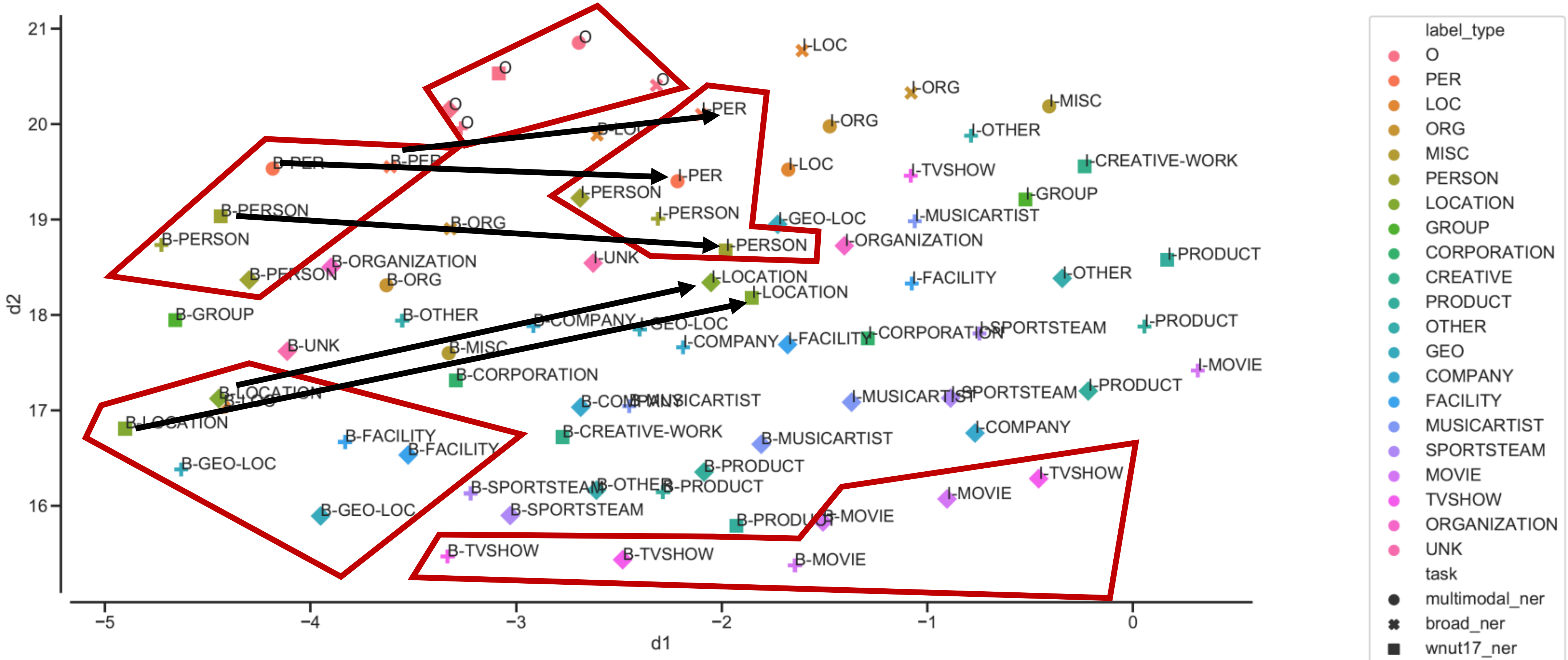
Label embeddings (POS)

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



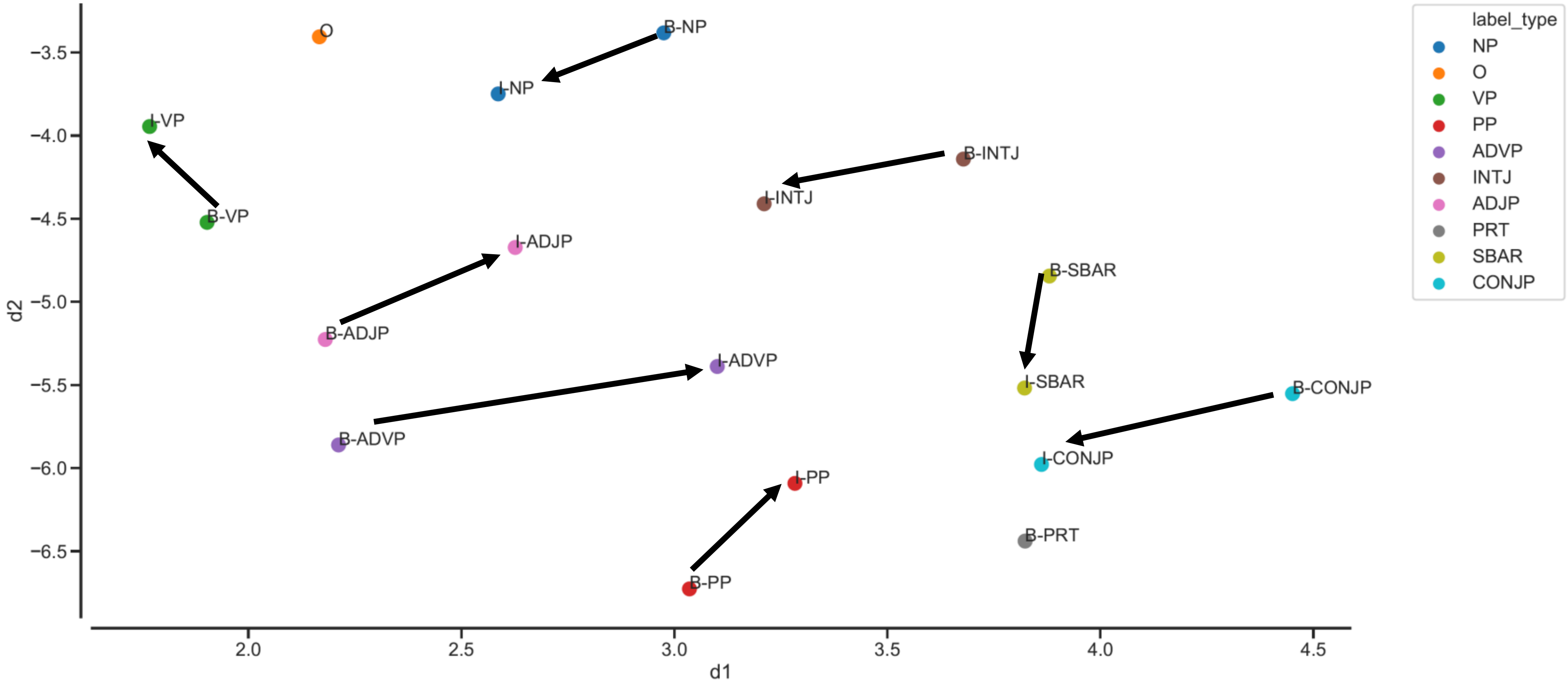
Label embeddings (NER)

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

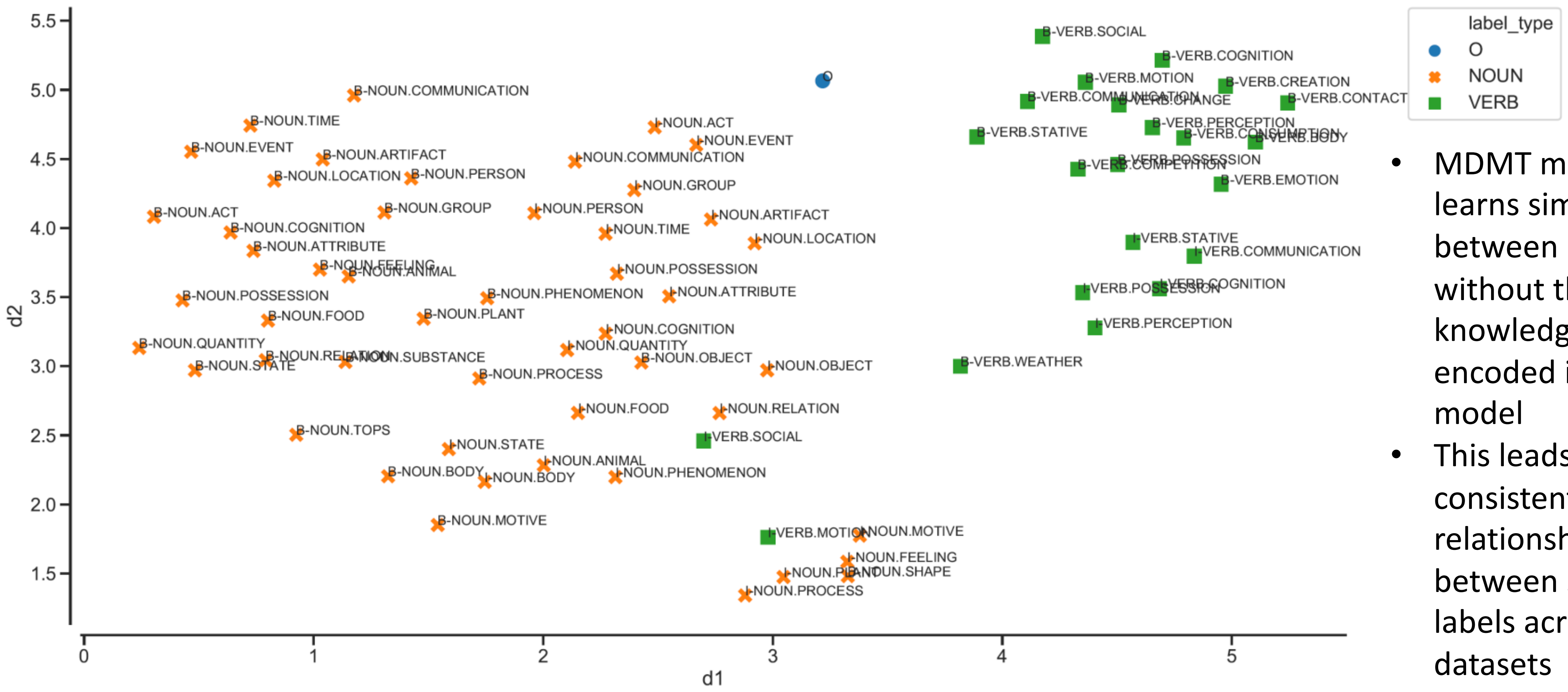


Label embeddings (chunking)

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

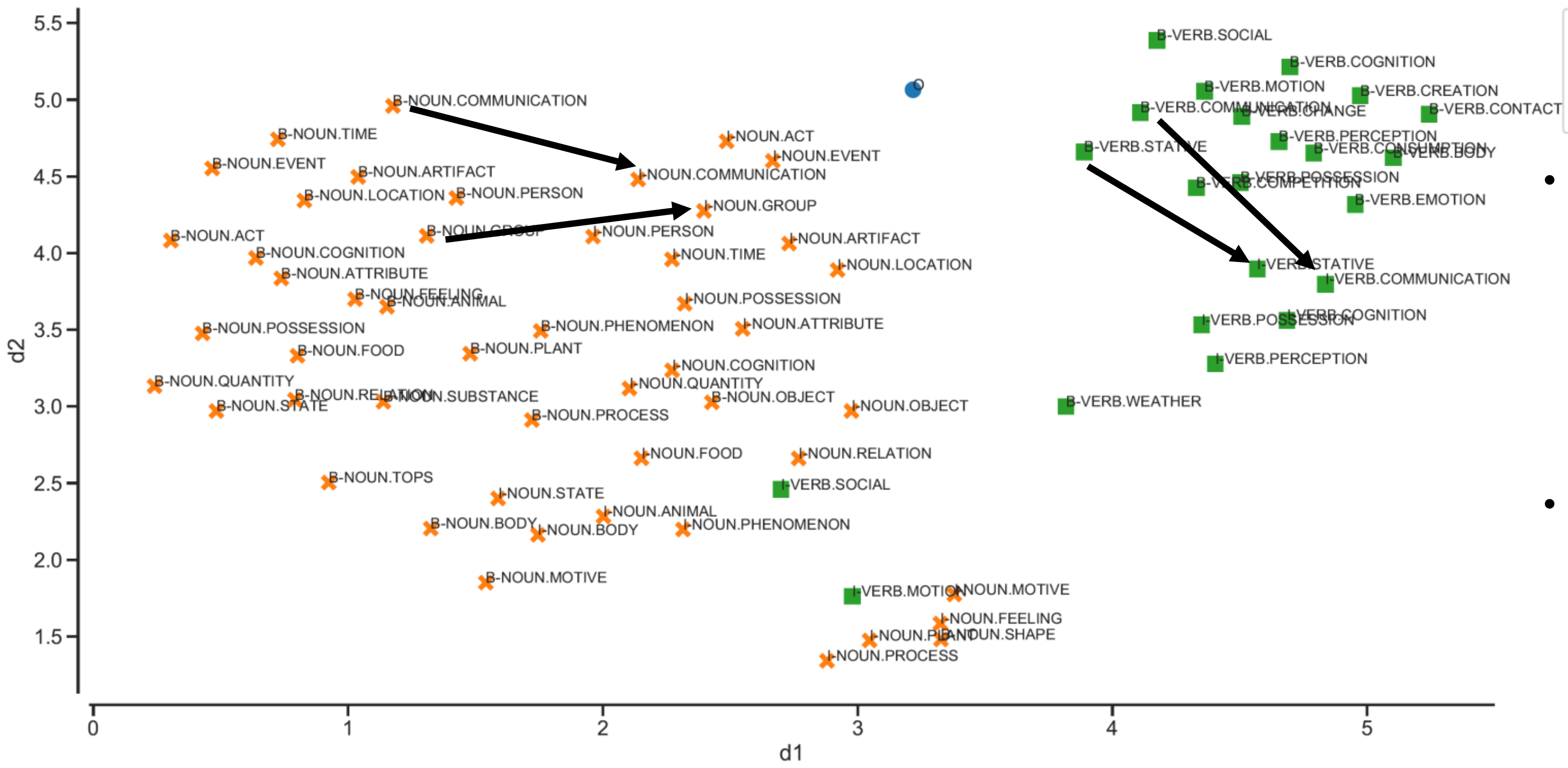


Label embeddings (super-sense tagging)



- 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

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	

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

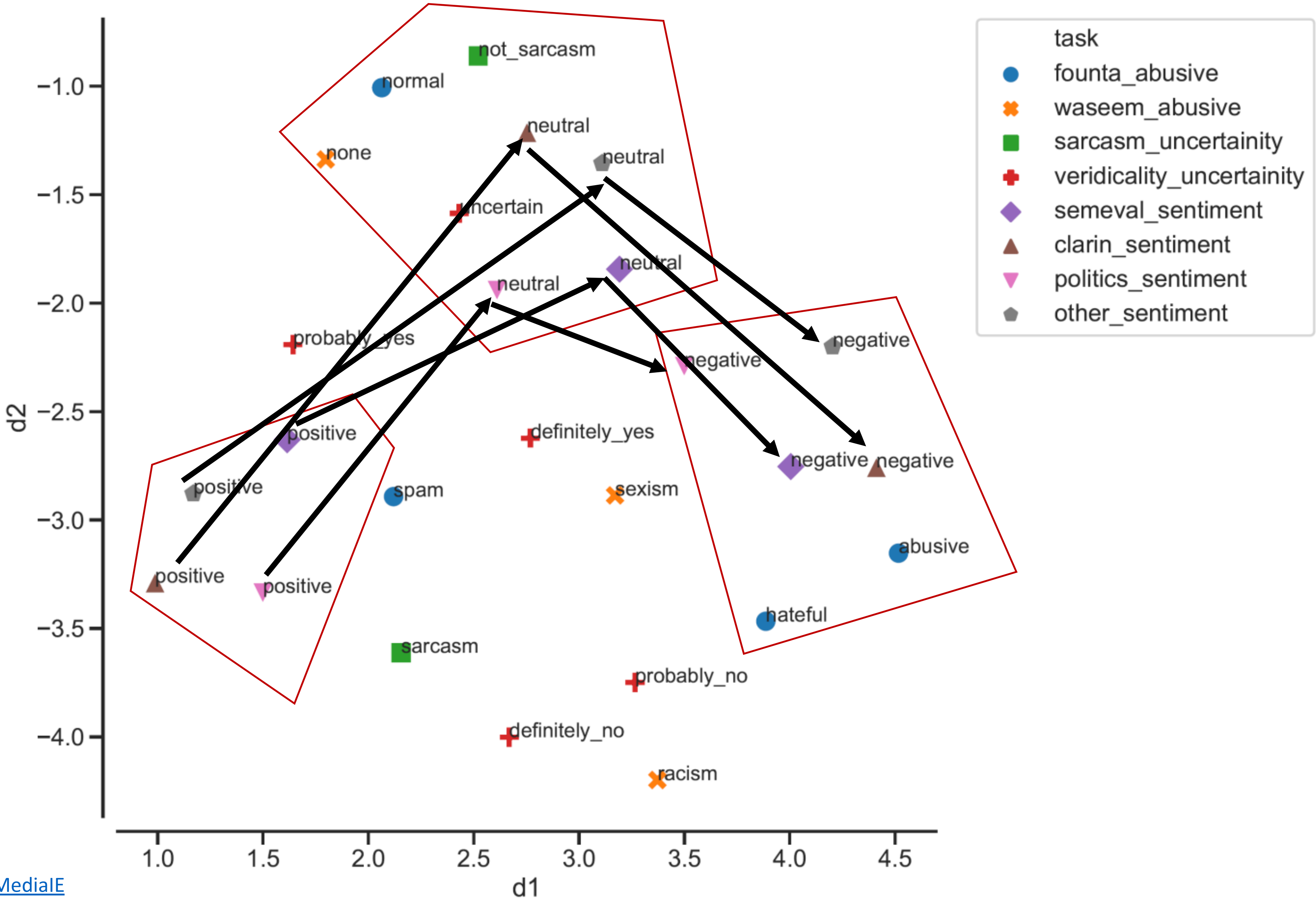
file	Founta		WaseemSRW	
model	r	v	r	v
S bilstm	8	79.33	8	81.72
MD bilstm	9	79.03	9	81.31
MTS bilstm	11	61.48	11	68.57
MTL bilstm	10	69.26	10	70.13
S bilstm *	1	**80.6**	3	**82.95**
MD bilstm *	2	**80.35**	2	**83.22**
MTS bilstm *	6	80.11	7	81.99
MTL bilstm *	4	80.23	5	82.78
S cnn *	3	**80.25**	4	82.89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

Use Wikipedia lookup + Dense retrieval for Entity Linking

Knowledge Base:
July 2022 Wikipedia - 6.5M Entities. Filtered to remove miscellaneous pages using Wikidata

TweetNERD Dataset[2]:
Dataset of over 340k+ Labeled Tweets. Evaluated on Academic and OOD Split

Lookup Retrieval:
Alias Table using Wikidata Aliases and Labels, ranked using probability of entity given surface form

Dense Retrieval:
Pre-trained BLINK[1] Encoders, embeddings indexed using FAISS. First 4 sentences of Wikipedia and annotated spans

Hybrid Retrieval:
Combine candidates from both Dense and Lookup

[1] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.
[2] Mishra, Shubhanshu, Saini, Aman, Makki, Raheleh, Mehta, Sneha, Haghighi, Aria, & Mollahosseini, Ali. (2022). TweetNERD - End to End Entity Linking Benchmark for Tweets (0.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6617192

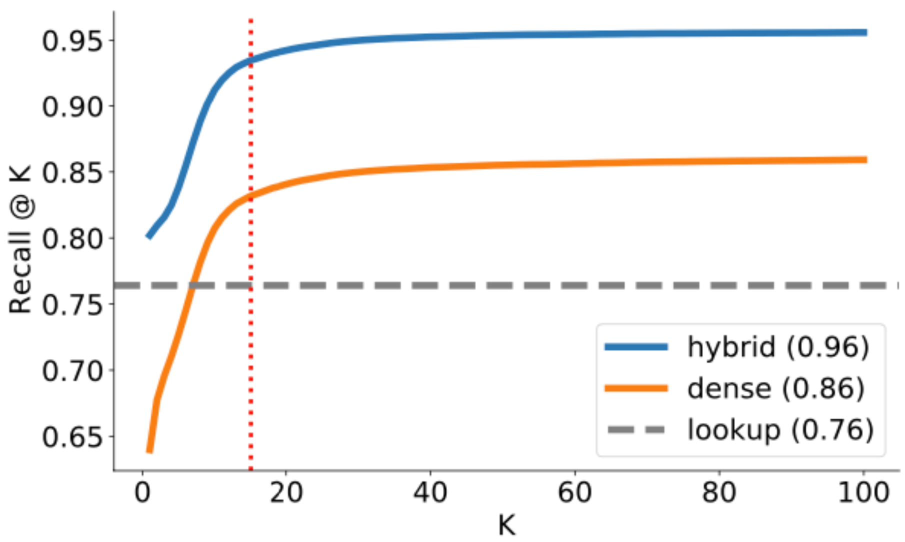
Recall@16 Using Gold Spans

Data Split	Dense	Lookup	BM25	Hybrid
Academic	<u>0.783</u>	0.741	0.221	**0.916**
OOD	0.772	<u>0.847</u>	0.556	**0.933**
Overall	<u>0.779</u>	0.717	0.362	**0.930**

Recall@16 Using NER Spans

Data Split	Dense	Lookup	BM25	Hybrid
Academic	<u>0.761</u>	0.613	0.164	**0.880**
OOD	0.754	<u>0.757</u>	0.440	**0.903**
Overall	<u>0.759</u>	0.715	0.245	**0.887**

Recall@K Using Gold Spans



Liam Hebert, Raheleh Makki, Shubhanshu Mishra, Hamidreza Saghir, Anusha Kamath, and Yuval Merhav. 2022. Robust Candidate Generation for Entity Linking on Short Social Media Texts. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 83–89, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

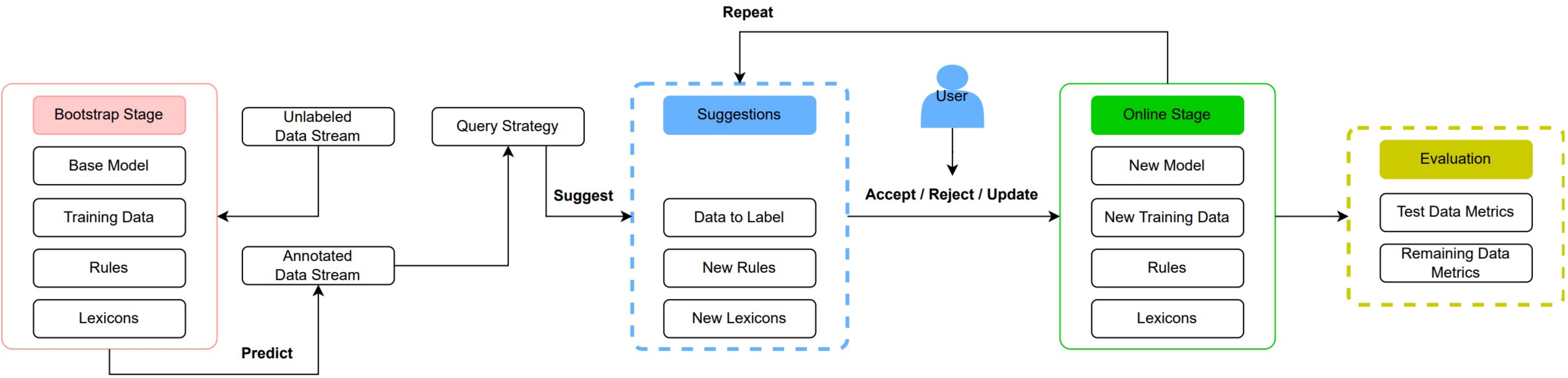
- Given a large unlabeled corpus, can we label it efficiently using fewer human annotations?
- Can existing models be updated efficiently to work with new data?
- Proposal:
  - Use active learning for data labeling
  - Use incremental learning algorithms for model updates
- Highly application to social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

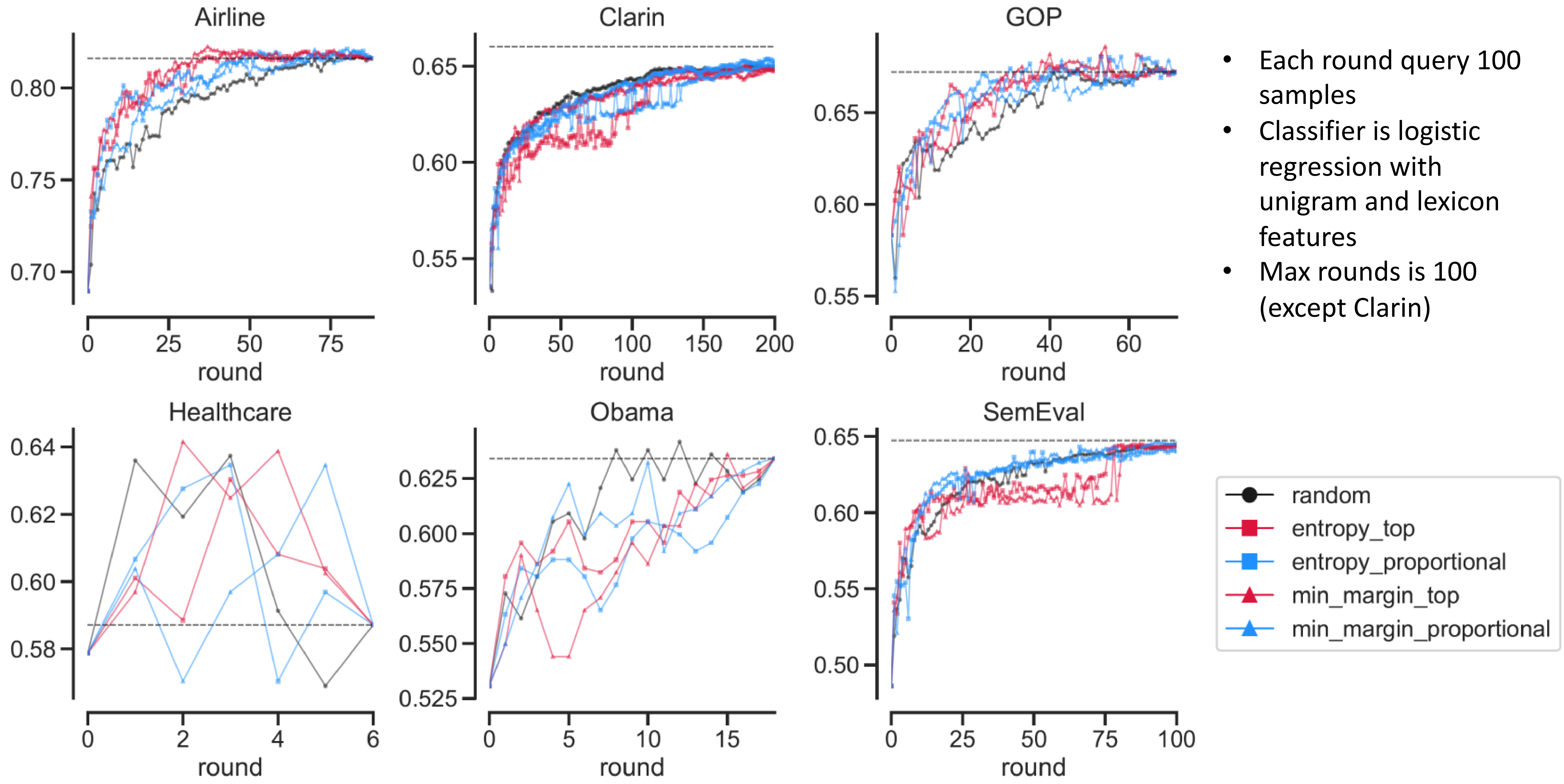
https://shubhanshu.com/phd_thesis/

Table 2: Performance of query strategies across datasets using around 10% training dataset.

task	dataset	round	N	N_{left}	$\%_{used}$	Full	Rand	E_{top}	E_{prop}	M_{top}	M_{prop}
Test Dataset											
ABUSIVE	Founta	42	41,861	37,661	0.10	0.79	0.77	0.78	0.78	0.79	0.77
	WaseemSRW	14	13,072	11,672	0.11	0.82	0.79	0.78	0.77	0.78	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	0.82	0.76	0.78	0.79	0.77	0.77
	Clarin	45	44,299	39,799	0.10	0.66	0.63	0.61	0.62	0.63	0.63
	GOP	8	7,121	6,321	0.11	0.67	0.63	0.64	0.63	0.62	0.64
	Healthcare	1	590	490	0.17	0.59	0.64	0.60	0.61	0.60	0.60
	Obama	2	1,777	1,577	0.11	0.63	0.56	0.60	0.58	0.59	0.57
	SemEval	13	12,145	10,845	0.11	0.65	0.59	0.60	0.61	0.58	0.61
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	0.78	0.77	0.76	0.77	0.76	0.79
	Swamy	1	555	455	0.18	0.39	0.39	0.40	0.39	0.34	0.31
Remaining Dataset											
ABUSIVE	Founta	42	41,861	37,661	0.10	NaN	0.77	0.80	0.78	0.81	0.78
	WaseemSRW	14	13,072	11,672	0.11	NaN	0.78	0.79	0.77	0.80	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	NaN	0.75	0.79	0.79	0.80	0.78
	Clarin	45	44,299	39,799	0.10	NaN	0.62	0.62	0.62	0.64	0.63
	GOP	8	7,121	6,321	0.11	NaN	0.62	0.64	0.62	0.63	0.63
	Healthcare	1	590	490	0.17	NaN	0.53	0.56	0.53	0.47	0.50
	Obama	2	1,777	1,577	0.11	NaN	0.54	0.56	0.57	0.56	0.56
	SemEval	13	12,145	10,845	0.11	NaN	0.61	0.62	0.62	0.63	0.62
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	NaN	0.80	0.82	0.84	0.82	0.81
	Swamy	1	555	455	0.18	NaN	0.37	0.40	0.40	0.33	0.36

https://shubhanshu.com/phd_thesis/

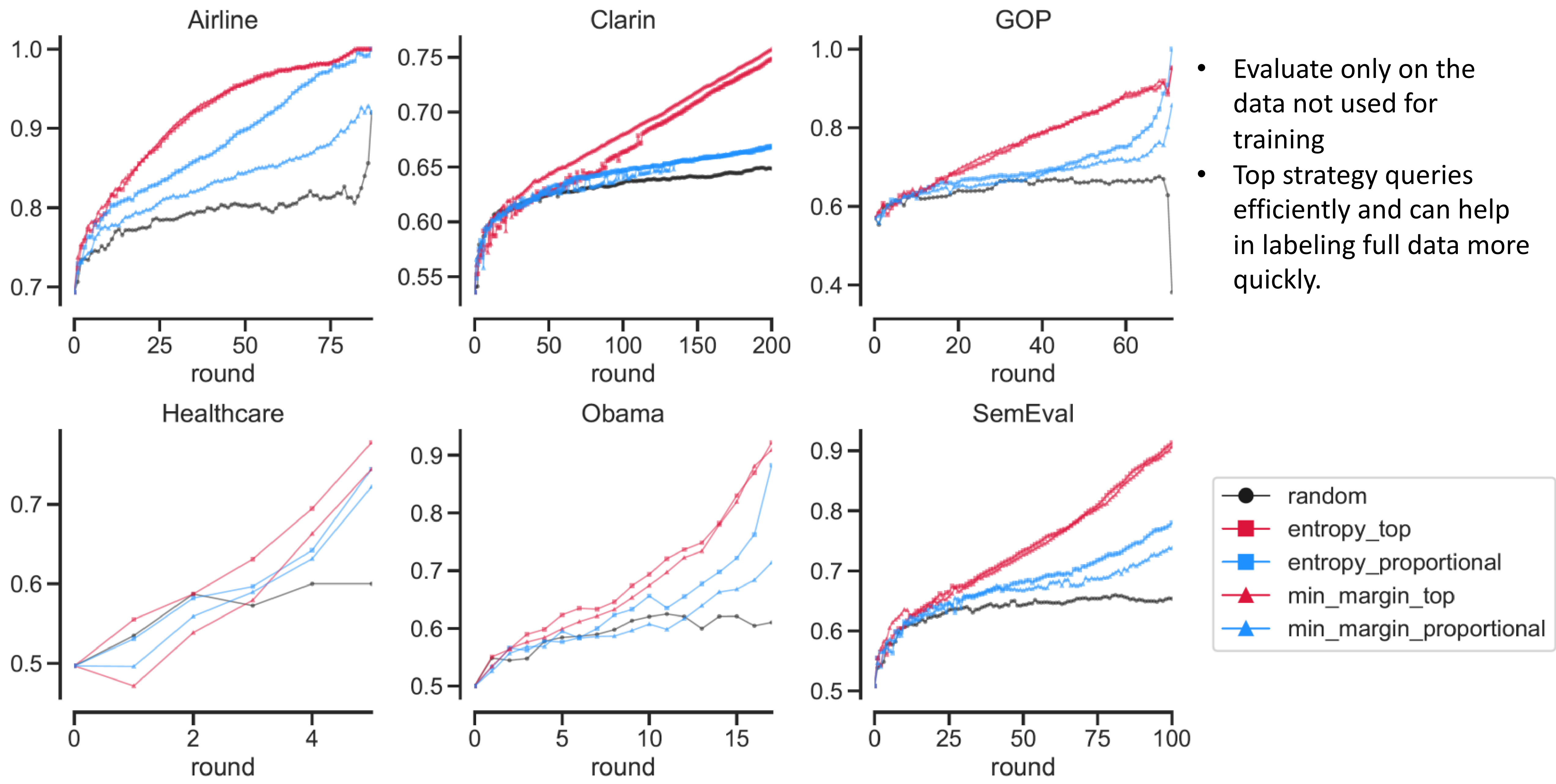
Mishra et al. (2015)



- 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

Less languages to learn: Multilingual learning to improve coverage

Stripe org acquires Nigeria loc 's Paystack org for $200M+ to expand into the African continent loc https://tcrn.ch/3j2mnS3 by @ingridlunden

Stripe org rachète la startup nigériane loc Paystack org pour 200 millions de dollars afin de s’implanter sur le continent Africain loc https://tcrn.ch/3j2mnS3 @ingridlunden

स्ट्राईप org ने $200M+ में नाइजीरिया loc के पेस्टैक org को अफ्रीकी महाद्वीप loc में विस्तारित करने के लिए अधिग्रहित किया https://tcrn.ch/3j2mnS3 @ingridlunden

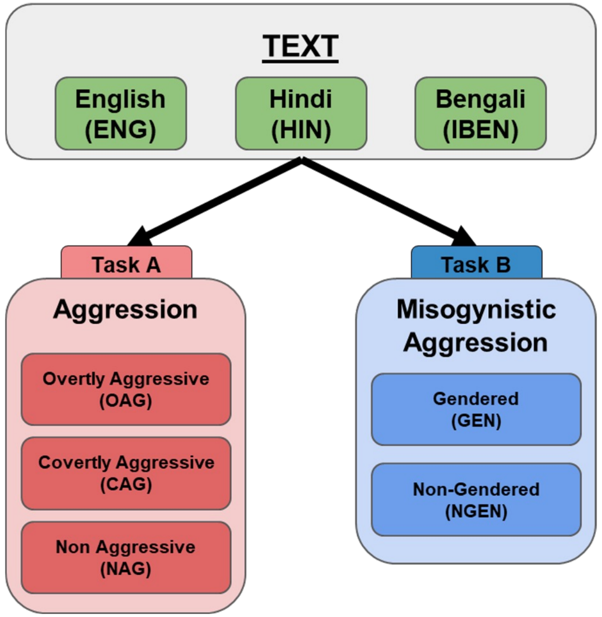
NER trained on tweets using Multilingual Word Embeddings and BiLSTM

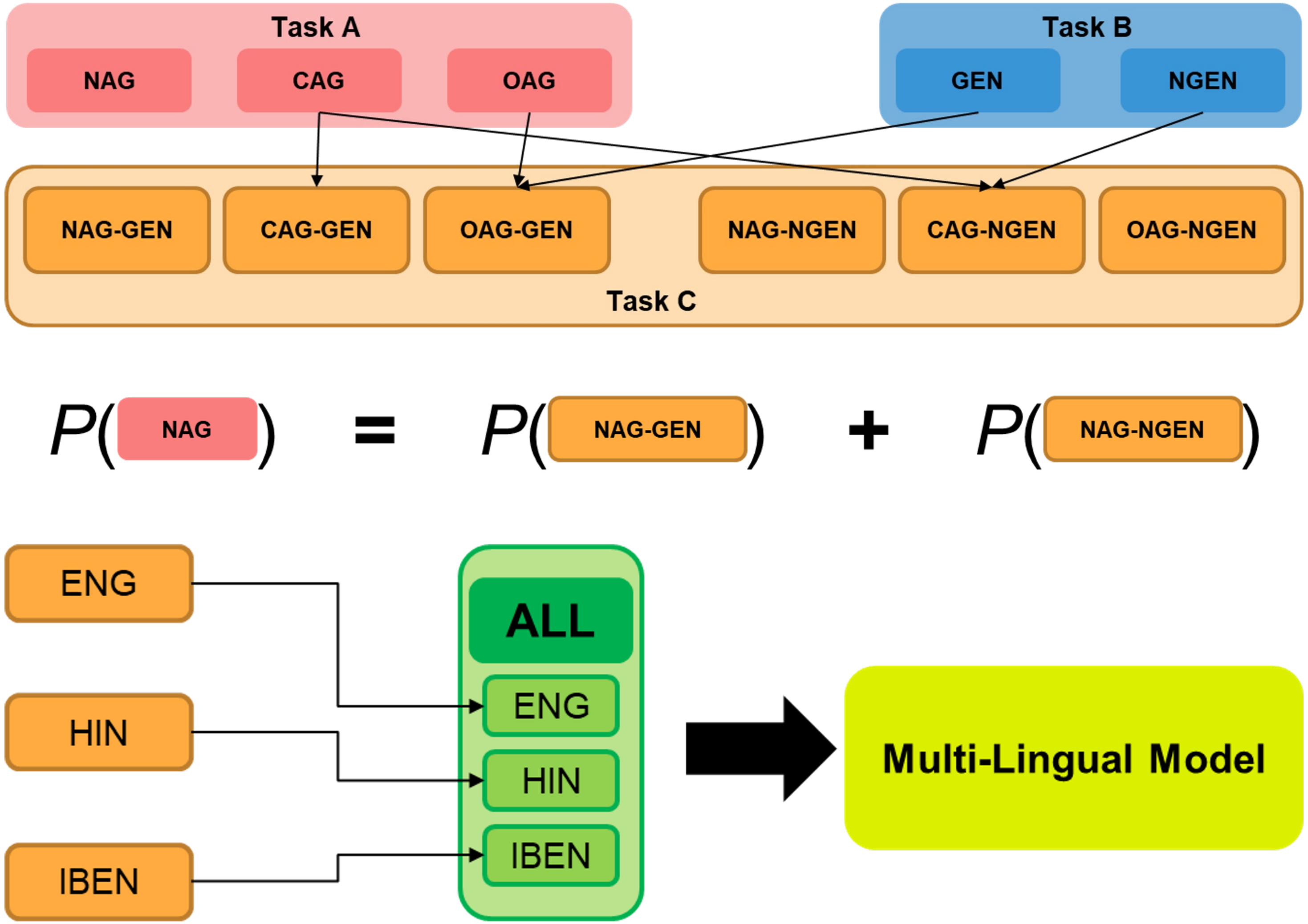
Language Testing Dataset	English CoNLL-03	German CoNLL-03	Dutch CoNLL-02	Spanish CoNLL-02	French xLIME	Italian xLIME	Turkish JRC	Hindi SEAS	Arabic CS-18
Lookup	36.6	22.8	36.8	29.7	15.6	23.3	22.9	**20.4**	16.7
Mono Training	40.2	35.5	39.4	27.4	27.7	**29.3**	24.8	11.8	**22.8**
Mul Training	38.3	36.6	43.2	29.1	26.4	28.9	28.0	9.8	14.0
Mono Training + WikiANN	**47.2**	**41.2**	**55.4**	37.6	30.3	28.4	27.8	14.0	21.9
Mul Training + WikiANN	43.2	39.6	52.8	**44.0**	**32.6**	25.4	**28.6**	8.3	11.3

Table 1: Entity-Level Micro-Average F1-scores for the PERSON, LOCATION and ORGANIZATION types

Table Source: Ramy Eskander, Peter Martigny, Shubhanshu Mishra. Multilingual Named Entity Recognition in Tweets using Wikidata in WeCNLP 2020

Multilingual transformer models for hate and abusive speech





https://github.com/socialmediaie/TRAC2020

Multilingual learning for hate speech detection

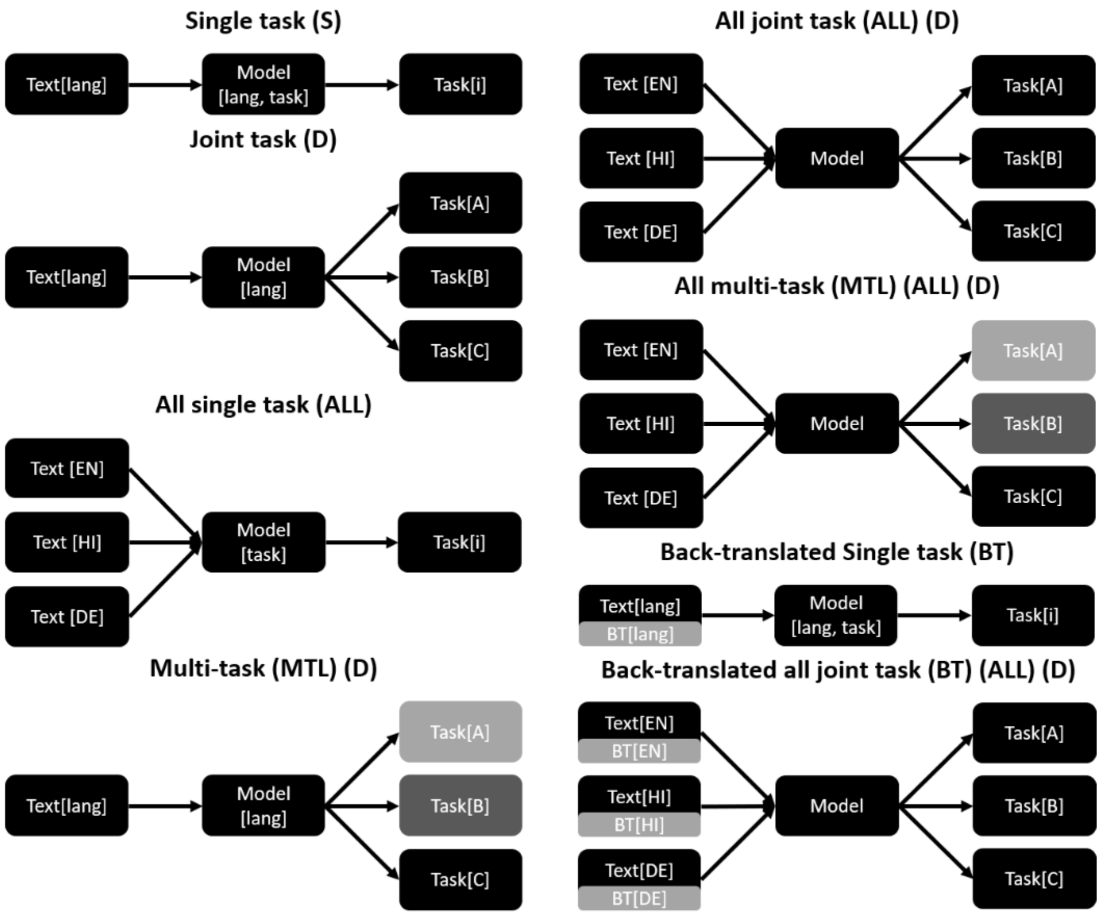


Fig. 2: An overview of various model architectures we used. Shaded task boxes represent that we first compute a marginal representation of labels only belonging to that task before computing the loss.

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

Multilingual Language Model Pretraining

	Hindi		Japanese		Arabic	
NER	F_1	Δ%	F_1	Δ%	F_1	Δ%
mBERT	21.1	0.0	16.5	0.0	32.1	0.0
+TPP (ONE)	**24.3**	15.2	**29.9**	81.4	**39.4**	22.8
+TPP (ALL)	23.2	10.3	27.4	66.4	38.5	19.9
Sentiment	F_1	Δ%	F_1	Δ%	F_1	Δ%
mBERT	31.7	0.0	55.0	0.0	51.5	0.0
+TPP (ONE)	**32.7**	3.0	66.4	20.6	58.3	13.2
+TPP (ALL)	32.4	2.3	**67.7**	23.1	**58.5**	13.7
UD POS	acc.	Δ%	acc.	Δ%	acc.	Δ%
mBERT	67.4	0.0	52.7	0.0	64.0	0.0
+TPP (ONE)	**71.5**	6.0	**57.6**	9.2	**67.1**	4.8
+TPP (ALL)	66.4	-1.5	52.7	0.1	65.0	1.5

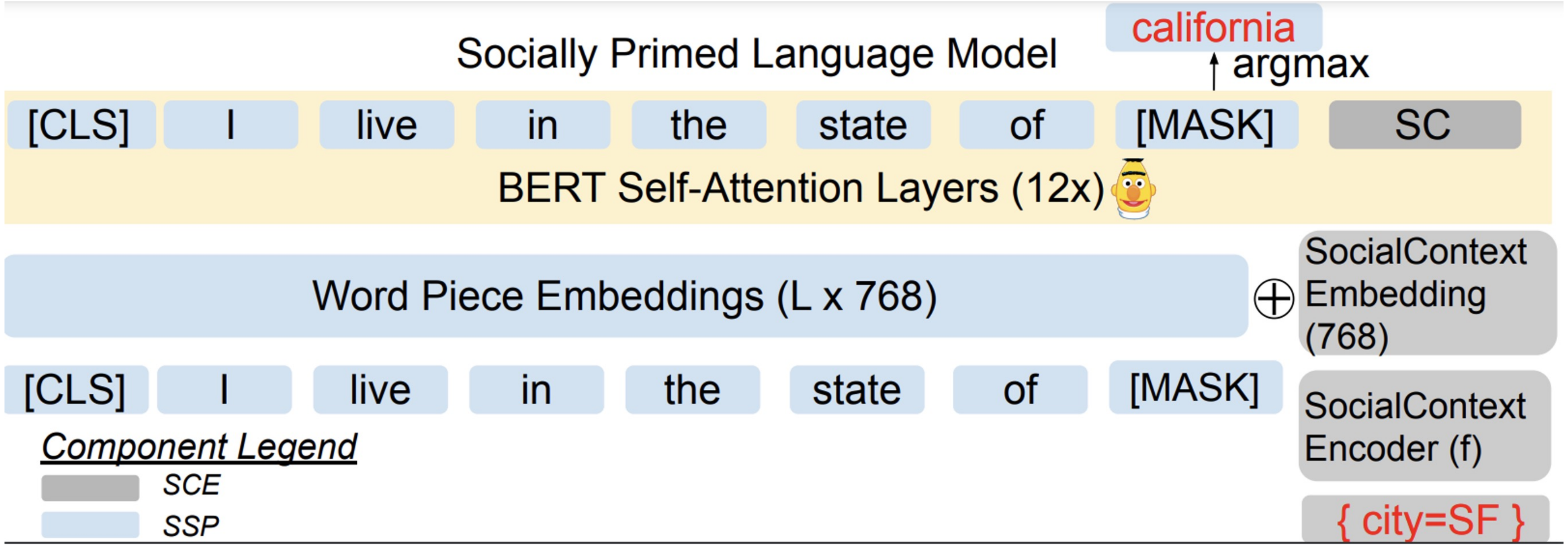
- **NER:** 37% relative improvement in F1.
- **Sentiment:** 12% relative improvement in F1.
- **UD POS:** 6.7% relative improvement in accuracy.

Shubhanshu Mishra and Aria Haghighi. 2021. Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction. In Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), pages 381–388, Online. Association for Computational Linguistics.

Less context to learn

Include Tweet context: Time, Geolocation, Meta-data, and Social Engagements

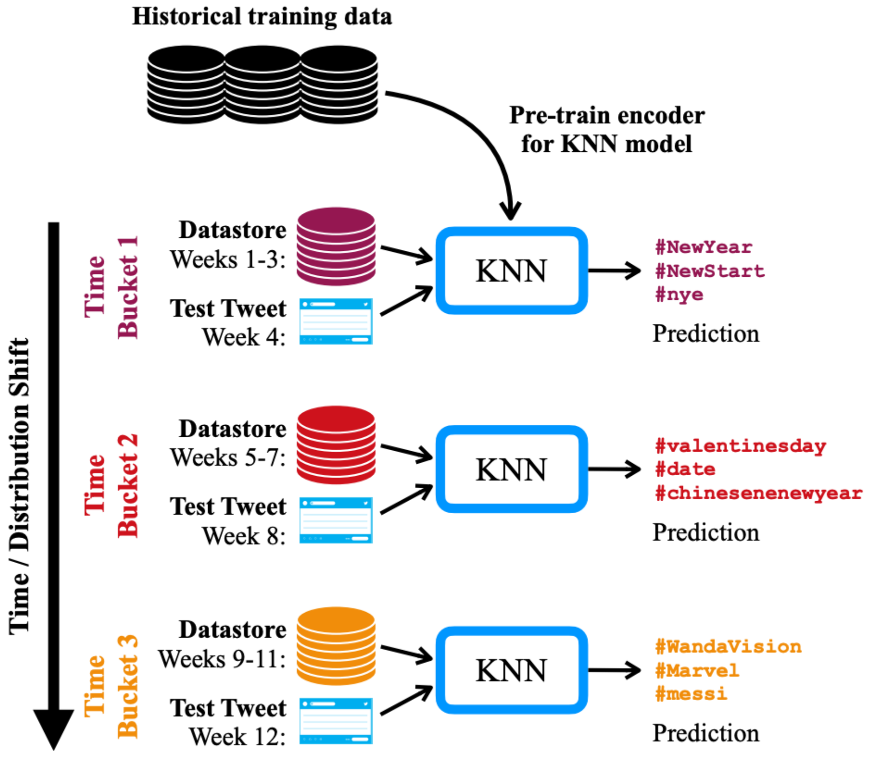
Use time and location

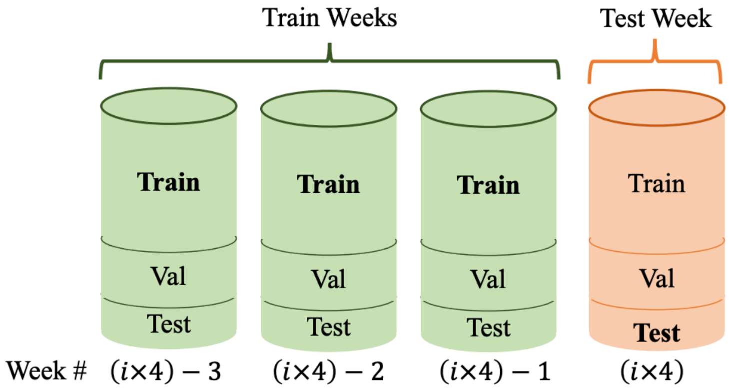


Input Sentence	Social Context	Top 10 predicted tokens
I reside in the state of [MASK]	*San Diego*	*california, ca, texas, mexico*
I reside in the state of [MASK]	*Dallas*	*texas, houston, mexico, california, tx*
I reside in the state of [MASK]	*Tampa*	*florida, georgia, fl, texas, jacksonville*
The most popular nfl team in our state is [MASK]	*San Diego*	*. the 49ers seattle patriots*

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics

Retrieve temporally similar content





	Non-temporal		Temporal W/o Adaptation		Temporal W/ Adaptation	
	@5	@1	@5	@1	@5	@1
Frequency baseline	1.68	0.47	1.11	0.26	1.69	0.46
Classifier	**39.87**	**14.80**	14.79	5.41	23.25	8.34
Seq2Seq	34.36	12.96	15.97	6.10	23.49	8.30
KNN-Clsf	39.54	13.15	**18.45**	**6.65**	**26.21**	**9.60**

Fatemehsadat Mireshghallah, Nikolai Vogler, Junxian He, Omar Florez, Ahmed El-Kishky, Taylor Berg-Kirkpatrick: "Non-Parametric Temporal Adaptation for Social Media Topic Classification", 2022; arXiv:2209.05706

Continually train across time

Models	2020-Q1	2020-Q2	2020-Q3	2020-Q4	2021-Q1	2021-Q2	2021-Q3	2021-Q4	Change
Barbieri et al., 2020	9.420	9.602	9.631	9.651	9.832	9.924	10.073	10.247	N/A
2019-90M	4.823	4.936	4.936	4.928	5.093	5.179	5.273	5.362	N/A
2020-Q1	4.521	4.625	4.699	4.692	4.862	4.952	5.043	5.140	-
2020-Q2	4.441	4.439	4.548	4.554	4.716	4.801	4.902	5.005	-4.01%
2020-Q3	4.534	4.525	4.450	4.487	4.652	4.738	4.831	4.945	-2.15%
2020-Q4	4.533	4.524	4.429	4.361	4.571	4.672	4.763	4.859	-2.81%
2021-Q1	4.509	4.499	4.399	4.334	4.439	4.574	4.668	4.767	-2.89%
2021-Q2	4.499	4.481	4.376	4.319	4.411	4.445	4.570	4.675	-2.83%
2021-Q3	4.471	4.455	4.335	4.280	4.366	4.394	4.422	4.565	-3.26%
2021-Q4	4.467	4.455	4.330	4.263	4.351	4.381	**4.402**	**4.463**	-2.24%
2021-124M	**4.319**	**4.297**	**4.279**	**4.219**	**4.322**	**4.361**	4.404	4.489	N/A

	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
RoBERTa-Base	30.8	76.6	44.9	55.2	78.7	72.0	70.9	61.3
TweetEval	31.6	79.8	55.5	62.5	81.6	72.9	72.6	65.2
BERTweet	33.4	79.3	56.4	**82.1**	79.5	73.4	71.2	67.9
TimeLM-19	33.4	**81.0**	**58.1**	48.0	**82.4**	73.2	70.7	63.8
TimeLM-21	**34.0**	80.2	55.1	64.5	82.2	**73.7**	**72.9**	66.2
Metric	M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG (F1)	TE

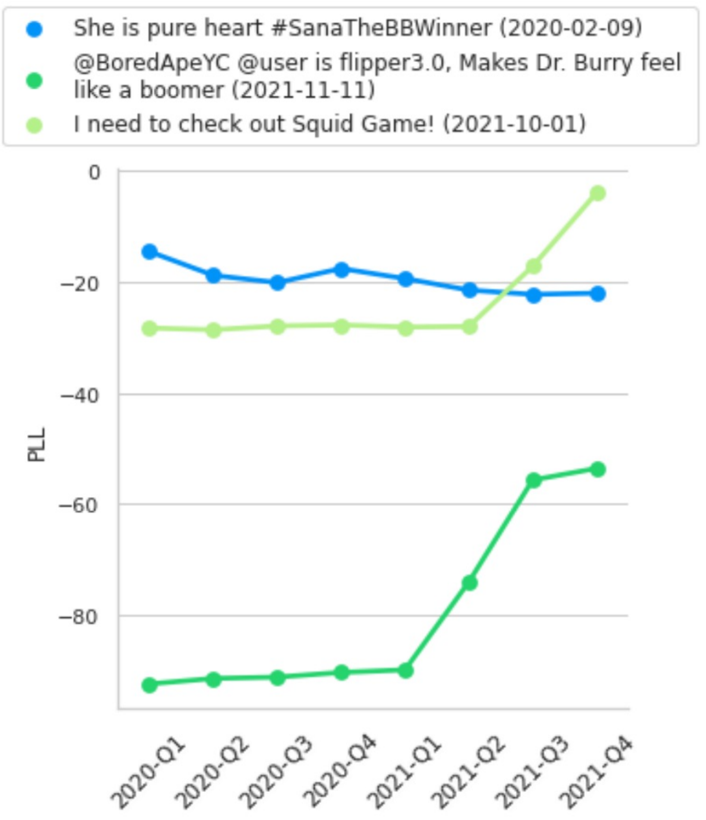
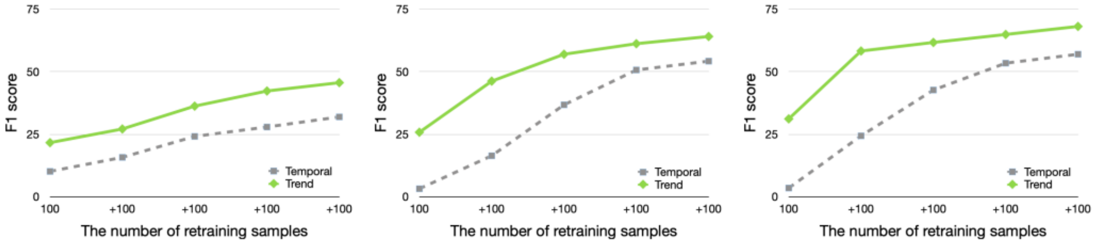


Figure 2: PLL scores of TimeLMs language models trained over different periods for three selected tweets.

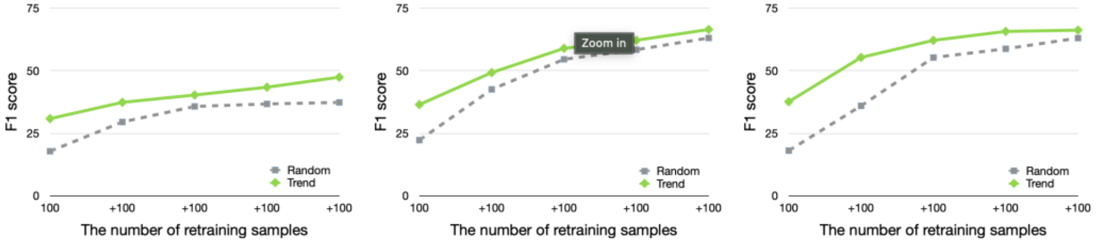
Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, Jose Camacho-Collados: "TimeLMs: Diachronic Language Models from Twitter", 2022; arXiv:2202.03829

Train using Trending Content



(a) BiLSTM + CRF (b) BERT + CRF (c) BERTweet + CRF

Figure 2: **Data can only be accessed year by year -** Each step represents a year from 2014 to 2018. At each step, we add instances from its respective year to the training set. For Temporal, we randomly select instances from that given year. For Trend, we rank instances based on their trending score. We experiment with 50 (Appendix B), 100 and 200 (Appendix B) instances per step to show the impact of training size.



(a) BiLSTM + CRF (b) BERT + CRF (c) BERTweet + CRF

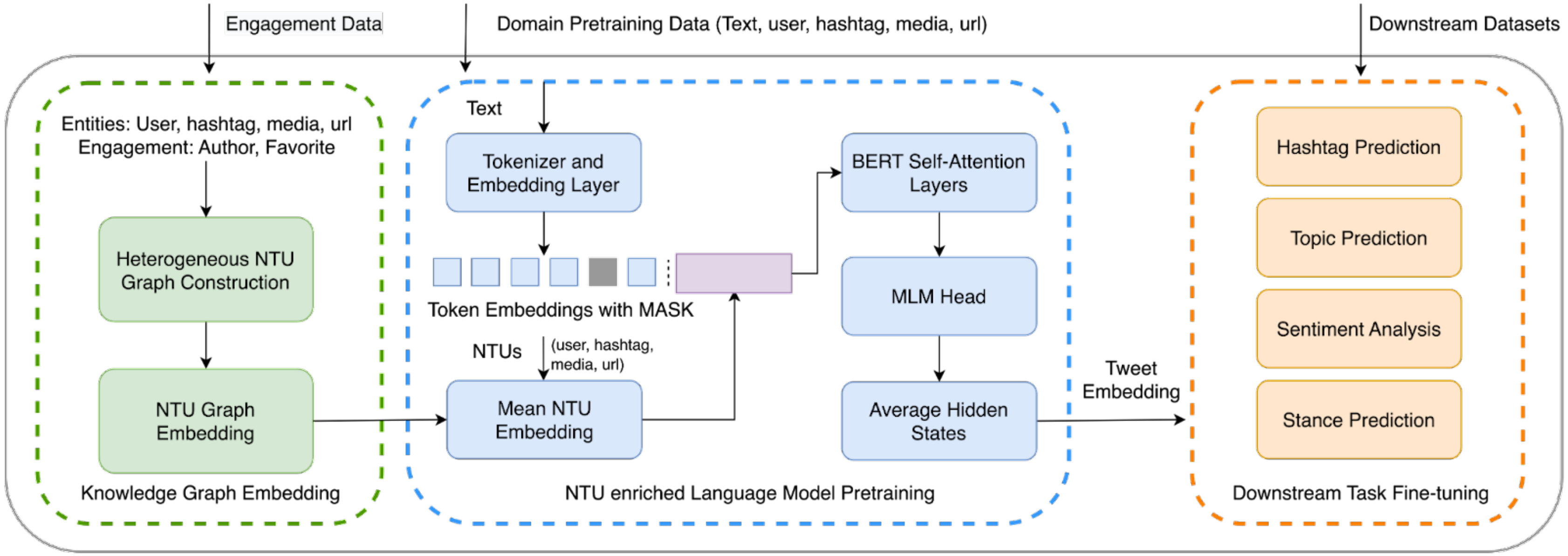
Figure 3: **Data from all years is available -** At each step, we add new instances to our training set. For Random, we randomly select instances from the available data. For Trend, we rank all available instances from most trending to less trending based on their trending scores. We then use this ranking to select the instances. At each step, we choose the instances with the highest trending scores that have not yet been added to the training set. We experiment with 50 (Appendix B), 100 and 200 (Appendix B) instances per step to show the impact of training size.

Model	**Random data**			**Trending data**		
	P	**R**	**F1**	**P**	**R**	**F1**
BiLSTM + CRF	59.38	48.02	53.10	63.42	54.83	58.81
BERT + CRF	62.26	73.23	67.30	70.07	69.93	70.00
BERTweet + CRF	60.64	64.84	62.67	65.45	70.46	67.86

Table 1: Performance comparison on random data and trending data, including persc.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2021. Mitigating Temporal-Drift: A Simple Approach to Keep NER Models Crisp. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 163–169, Online. Association for Computational Linguistics.

Use non-textual units in social media posts



Author: *user1*
Tweet: Our paper was accepted at *@WNUT* with *@user2* *@user3* *#nlproc* *#socialmedia*
Favorited by: *user4*, *user5*

Table 1: Example tweet with engagement data of author, mentions, Hashtags, and favorites

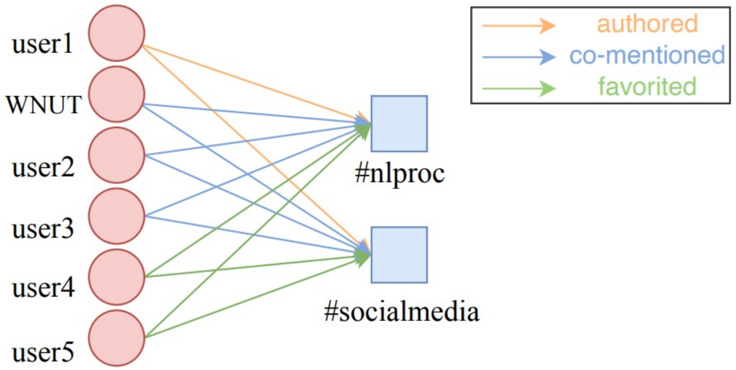


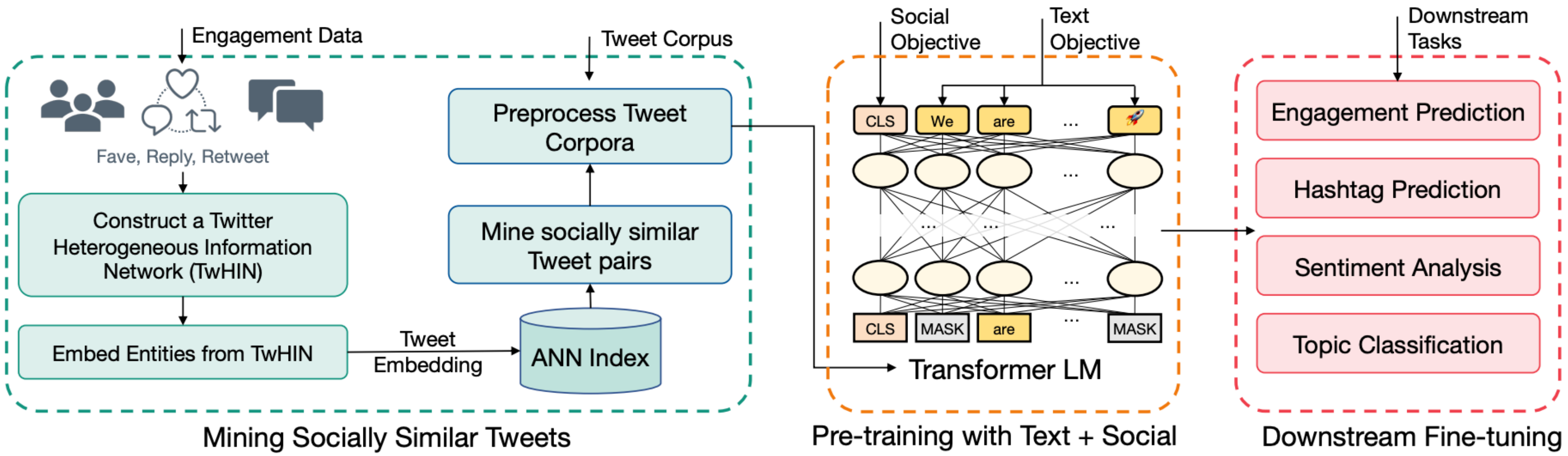
Figure 2: Graph construction with the example data in Table 1 for training NTULM user-Hashtag embeddings.

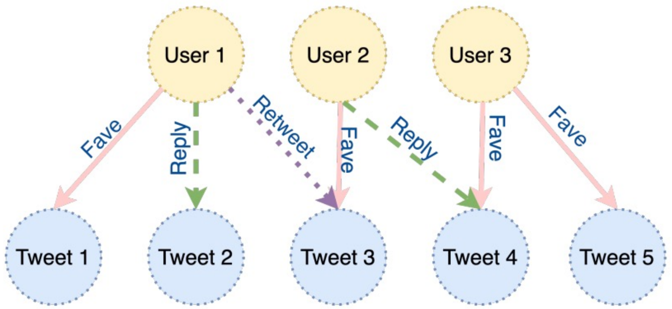
Model	NTUs	Perplexity bits	Topic MAP	TweetEval mean F1	SemEval 1 mean F1	SemEval 2 mean F1	Hashtag Recall@10	SMIE mean F1
BERT	-	4.425	0.327	0.577	0.527	0.515	0.689	0.548
NTULM	**author**	4.412	0.325	0.579	0.527	**0.548**	0.693	0.548
NTULM	**Hashtag**	4.391	0.339	0.586	0.534	0.545	0.711	0.539
NTULM	**author+Hashtag**	**4.344**	**0.343**	**0.590**	**0.534**	0.545	**0.720**	**0.549**

Table 2: NTULM compared with BERT (MLM fine-tuned, section 4.2). We report the perplexity, mean average precision (MAP) in Topic, Recall@10 in Hashtag Prediction, and mean F1 score in the rest.

Jinning Li, Shubhanshu Mishra, Ahmed El-Kishky, Sneha Mehta, and Vivek Kulkarni. 2022. NTULM: Enriching Social Media Text Representations with Non-Textual Units. In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), pages 69–82, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Use user co-engagement





Method	SE2017 en	SE2018 en	SE2018 es	ASAD ar	COVID-JA ja	SE2020 hi+en	SE2020 es+en	Avg.
mBERT	66.44	27.73	32.14	68.98	41.87	66.44	45.55	49.88
BERTweet	**72.95**	**33.44**	-	-	-	-	-	-
XLM-T	71.50	31.99	35.88	80.87	42.09	70.96	51.03	54.90
TwHIN-BERT-MLM	71.98	31.73	**36.06**	80.36	43.09	71.09	50.83	55.02
TwHIN-BERT	72.40	31.38	36.00	**81.17**	**43.79**	**72.01**	**51.67**	**55.49**

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, Ahmed El-Kishky: "TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations", 2022; arXiv:2209.07562.

Improving sentiment classification using user and tweet metadata

Sentiment is usually identified as **positive**, **negative**, and **neutral**.

Twitter Sentiment Corpora

Tweet:
- Text
  - *Sentiment*
- # of URLs, Hashtags, Mentions
- Created at
- Retweets
- Replies
- Is reply or quote?
- User:
  - Created at
  - # Followers, Friends, Statuses
  - Is verified or has profile URL?

What we use

What we discard

- Are our corpora biased to certain meta-data attributes?
- Can those biases propagate into systems trained on these corpora?
- How correlated are these meta-data features with the annotated sentiment?
- Do these correlations hold outside of the annotated data for the same users?
- Can sentiment classifiers exploit this bias to do well on these datasets?

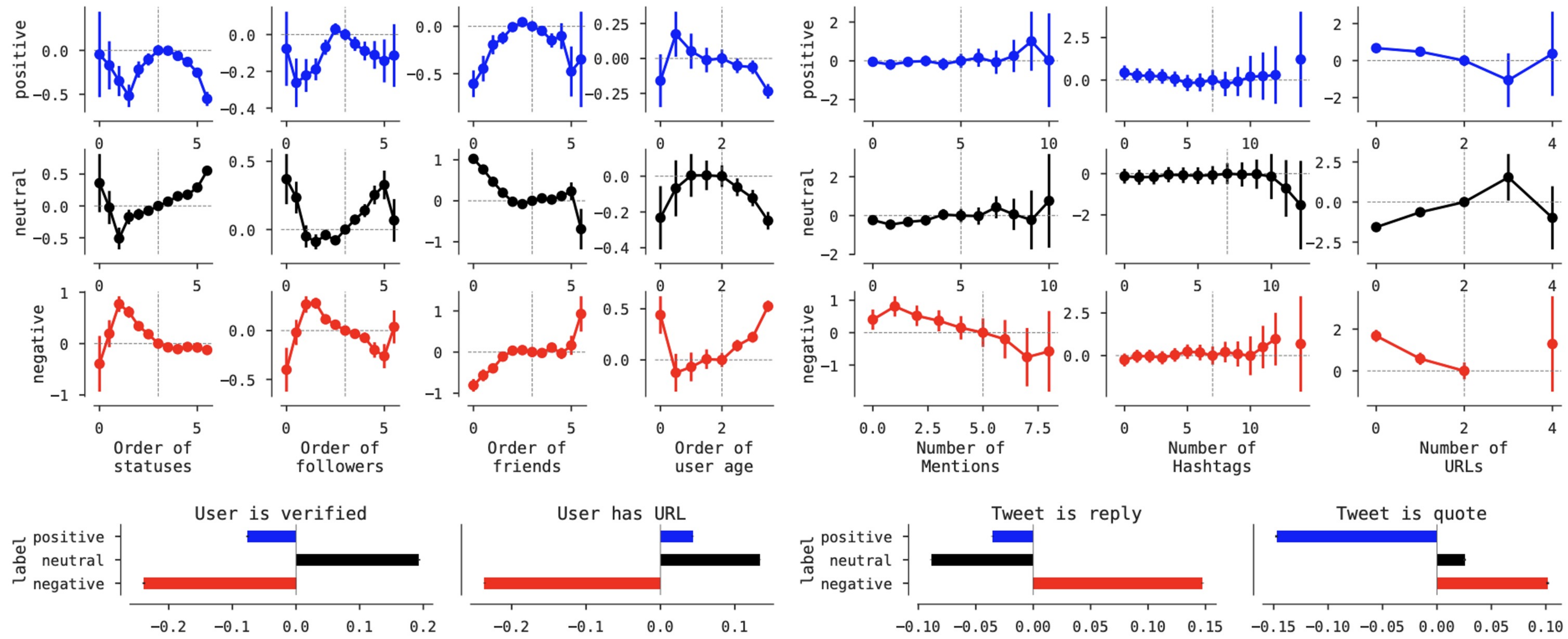
Mishra, S., & Diesner, J. (2018, July 3). Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. Proceedings of the 29th on Hypertext and Social Media. HT '18: 29th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3209542.3209562

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/CIKM2022/

Initial setup

- Open google Colab notebook specified at: https://socialmediaie.github.io/tutorials/CIKM2022/#software-setup
- On Colab click **Connect**
- Follow along during the session.
- 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

List of social media IE tools

Ours:
- 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/)

Others:
- TweetNLP - https://github.com/cardiffnlp/tweetnlp - Uses transformer models
- TweeBankNLP - https://github.com/mit-ccc/TweebankNLP - User transformer models + Stanza and supports token level tasks like NER, POS, Dependency Parsing
- TwitterNER - https://github.com/socialmediaie/TwitterNER (more lightweight NER focused on English tweets)
- ConText - https://github.com/uiuc-ischool-scanr/ConText (generate networks from text data)
- Bertweet – large scale pre-trained Roberta model - https://huggingface.co/vinai/bertweet-base
- BERTweet NER - https://huggingface.co/socialmediaie/bertweet-base_wnut17_ner
- Twitter Roberta - https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Using SocialMediaIE for IE from text

- Notebook link: https://github.com/socialmediaie/tutorials/blob/master/docs/CIKM2022/CIKM_2022_Tutorial_SocialMediaIE.ipynb (Click on Open in Colab)
- 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://socialmediaie.github.io/PredictionVisualizer/ 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

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/

Using Twitter API and Tweet Downloader

Key benefits	Access Twitter's real-time and historical public data with additional features and functionality that support collecting more precise, complete, and unbiased datasets. **More details on included endpoints**
Tweet cap	10 million Tweets / month
Query rules	1024 characters, 1000 streaming rules
Streaming rates	50 requests / 15 minutes, per app
Technical support	Developer documentation, tutorials, support content, and community forums
Cost	Free

New Tweet downloader

Along with the SDKs, we have a new addition to the Twitter API Tools 26 called the Tweet Downloader. The downloader provides Academic Researchers a quick and easy way to access historical Twitter data from the full-archive search endpoint 11 via a no-code web interface. Like the Query Builder 7, the UI offers the same, easy-to-use form to build and group search queries where you can then save the matching Tweets in either JSON or CSV format to your machine.

To get started with the downloader, you must provide a Bearer Token with Academic Research access 14. This tool is only available to developers with access to the full-archive search endpoints (available via Academic Research access).

Tweets Downloader features:

- Build search query via a user-friendly web interface
- No coding needed
- Start/End date picker
- Ability to download data in either CSV or JSON format
- Run multiple queries at the same time
- Continue downloading if there are API timeouts or page reloads

- How to build Twitter search queries: https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query
- Academic Research Access: https://developer.twitter.com/en/products/twitter-api/academic-research
- Twitter API v2 Docs: https://developer.twitter.com/en/docs/twitter-api
- A guide to teaching with the Twitter API v2

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
```

Responsible handling of social media data

Personally Identifiable Information (PII) and Ownership

- **Status quo**:
  - Tech innovation often precedes policy
  - Collection, storage, fusion, mining of large-scale user data/ personally identifiable data fast, cheap, easy
  - Technically feasible versus legal versus ethical
- **Common misassumptions**:
  - Publicly available data can be accessed, downloaded, stored, analyzed
  - People who post information online don’t expect privacy and do consent to the data being used for research
  - Creator of data (author) is the owner of the data
  - Anonymity equals privacy

Why is responsible data handling hard? Multitude of regulations!

1. **Governmental, institutional, and communal norms and regulations**
   - Health Insurance Portability and Accountability Act (HIPAA), Fair Information Practice Principles (FIPPs), Menlo Report (Ethical Principles Guiding Information and Communication Technology Research, 2012), and many more
   1. **Intellectual property**
   2. **Privacy and security law and regulations**
2. **Terms of use/ service**
3. **Technical constraints** (robots.txt, APIs)
4. **Personal values**
   - People apply them consciously or unconsciously
   - Depend on gender (Gilligan 1987), culture (Graham at el. 2011)
   - 16+: Conventional morality (comply with (group) norms) versus 10-15% post-conv. morality (own principles) (Kohlberg 1984)

Learn more at

•Diesner, J., & Chin, C. (2015). Usable Ethics: Practical considerations for responsibly conducting research with social trace data. Workshop: Beyond IRBs: Ethical Review Processes for Big Data Research, Future of Privacy Forum, Washington DC.
•Diesner, J., & Chin, C. (2016). Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. ETHI-CA² Workshop (ETHics in Corpus Collection, Annotation & Application), 10th Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia.
•Diesner, J., & Chin, C. (2016). Seeing the forest for the trees: Understanding and implementing regulations for the collection and analysis of human centered data. Human-Centered Data Science (HCDS) Workshop, 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016).

Data are Property, Property is Protected

- United States Constitution, Article I, Section 8:
  "The Congress shall have Power […] To promote the **Progress** of Science and useful Arts, by securing for limited Times to Authors and Inventors the **exclusive Right** to their respective Writings and Discoveries."
  - Copyright, and fair use (depends on country)
  - Patents
  - Trademarks
  - Trade Secrets

Data are Property, Property is Protected

©

1. **Copyright**

- Protectable: original, fixed in a tangible medium, non-functional expressions
- Protection starts at moment of creation
- Not protectable: facts and ideas
- © means: Nobody can copy, distribute, create derivatives, perform, display expressions that are copyright protected without consent from the copyright owner.
- Neither registration with US copyright office nor copyright symbol required for ensuring copyright protection
  - Registration required for suing for copyright infringement
- Exception: fair use: allows the use of very small portions of copyrighted material without permission from owner, e.g., for education, research, news reporting (depends on country)
- Baseline: creator of data is owner, exceptions:
  - Contracts (work for hire)
  - Terms of service

Building on top of classic copyright

- Creative Commons
  - Attribution: You let others copy, distribute, display, and perform your copyrighted work — and derivative works based upon it — but only if they give credit the way you request.
  - Share Alike: You allow others to distribute derivative works only under a license identical to the license that governs your work.
    - Also known as Copyleft
  - Noncommercial: You let others copy, distribute, display, and perform your work — and derivative works based upon it — but for non-commercial purposes only.
  - No Derivative Works: You let others copy, distribute, display, and perform only verbatim copies of your work, not derivative works based upon it.

http://creativecommons.org/about/licenses

Using Creative Commons Licenses to ensure Open Knowledge

- Wikipedia
  - Text available under Creative Commons Attribution ShareAlike License
  - Constraints:
    - Attribution—You must attribute the work in the manner specified by the author or licensor
    - Share Alike—If you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar or a compatible license.

In what sense are online data public?

- Open Data, Open Science, Open you name it…
  - Gratis (free as in free beer) versus libre (free as in free speech) (Floss, Stallman, GNU)
  - User-generated data from 3rd party platforms often “free to see”
- GNU General Public License (GPL), originally authored by Richard Stallman
  - “"free" in the sense of freedom: […] freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially.”
  - Author and publisher get credit for their work without being responsible for modifications made by others
  - State of the art for Open Source (software) development projects
    - For example Github, Sourceforge

Scraping public data – technical aspects

- APIs -> use that if available before thinking about scraping
- robots.txt -> consider those if you collect data directly
- Scraping: manual (time consuming) or automated (crawler) (noisy)

Scraping - legal aspects

- End user license agreements (EULA):
  - Shrink wrap contracts: unsigned permit understand-ings (assumption: user agrees by opening the product)
  - For websites aka browse-wrap, click-wrap, web-wrap
    - Can change often
    - Vague
    - Inconstant across sites
    - Lack context
  - For more, see Fiesler, Bearn and Keegan (2020)

Scraping public data - legal aspects

- ToS are contract law
- Problem: ToS can violate the Computer Fraud and Abuse Act CFAA (1984), which is a federal law against hacking (unauthorized access to a computer)
  - Does the CFAA make ToS violates a federal crime?
    - United States of America v. Aaron Swartz (2011)
    - Browsewrap agreements not enforceable:
      - “Terms of Use” hyperlinks “not sufficiently conspicuous” (obvious) for “reasonably prudent internet consumer” (plaintiff did not manifest unambiguous assent to be bound by Terms of Use“) (Long v. Provide Commerce, Inc., 2016 WL 1056555, Cal Ct. App., 03/17/2016)

Scraping public data - legal aspects

- Creating sock puppet accounts and collect data to research algorithmic discrimination online does not violate CFAA
- Sandvig v. Barr, filed by American Civil Liberties Union (ACLU) on behalf of academics, computer scientists, journalists, ruling came out in March 2020
- "Researchers who test online platforms for discriminatory and rights-violating data practices perform a public service. They should not fear federal prosecution for conducting the 21st-century equivalent of anti-discrimination audit testing." (https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti)

Scraping public data - legal aspects

- HiQ versus LinkedIn, legal decision (2017, see Fiesler for details):
  - HiQ, a talent management service, violated ToS -> received cease and desist letter from LinkedIn to stop scraping
  - By virtue of being published publicly, a website authorizes the public to access it
  - Revoking access on case-by-case basis problematic (can be discriminatory)
  - CFAA does not apply to scrape non-password protected data
- Line in the sand: access control such as passwords

Scraping public data - ethical aspects

- Scraping can be illegal but ethical
- Scraping can be legal but unethical
- Expectations of users: data available beyond website (download, redistributed) and for research?
- Contextual privacy (Nissenbaum 2020)
- More on data: working with online data kind of archival research (Kosinski et al. 2015)
  - No consent needed if 1) users consciously made their data public, 2) collected data anonymized, 3) researchers do not interact with participants, 4) no identifiable user information published

How to decide if and how to collect online data?

- Scraping, crawling: Consider access control
- Do a holistic case by case assessment (Fiesler, 2019)
  - Amplification?
  - Inference?
  - Seek guidance from members of community
- **Include information about your ethical consideration and reasoning in your papers**

Social data from online sources: collection/ acquisition methods

- (omitting interaction methods (elicitation, user studies) and crowdsourcing)
- Reuse existing data
  - Benchmarks
  - Archival data, repositories: https://datasetsearch.research.google.com https://www.kaggle.com/datasets https://datacatalogue.cessda.eu/ (social science data)
- APIs, scraping, crawling
  - Code -> maintenance, chasing a moving target, dependencies
- 3rd party services (e.g., BrandWatch, Crimson Hexagon, Pushshift)
- Buy data
- **Shared issues:**
  - **provenance, quality and biases, sampling, context of data production and collection impact data, ethics**

Data documentation – Why?

- Make important aspects explicit
- Avoid pitfalls with important aspects, e.g., discriminatory outcomes
  - Digital social data increasingly used to develop policy, decision making, design products and services
  - Not just an observational tool
- Standardization to ease collaboration/ communication, esp. in interdisciplinary teams
- Improve transparency, accountability, reproducibility, responsibility
  - Starting with collection, preprocessing, representation/ indexing/ storage, provenance
- Select appropriate datasets

Pitfalls of working with digital (social) data

- Concerns/ pitfalls
  - Opportunistic use
  - Biases introduced in data itself or measurement (data collection, methods, can depend on research context -> case by case assessment necessary, again)
    - Social, technical, and methodological roots of biases
  - Lack of consensus on vocabulary and taxonomy for biases and measurement issues
  - Data quality
- For more: boyd and Crawford 2012, Olteanu 2019

Datasheets for datasets

- Gebru et al. 2020, industry (Google, Microsoft)/ academia project, originally for ML datasets
- Audiences:
  - Dataset creators: **reflect** on:
    - Workflow: process of creation, distribution, maintenance of dataset;
    - Assumptions, risks or harms, implications of use
  - Consumers: informed choice about use
- Producing a datasheet not an automated process, dependent on domain and specific case
- Design:
  - iterative
  - Yes/ no questions discouraged

Datasheets for datasets: Questions for developers

- Workflow questions
  - Motivation
  - Composition
  - Collection
  - Preprocessing, cleaning, labeling

Read before the step, answer once task is completed

- Uses
- Distribution
- Maintenance

Provide answers before executing the step

Model cards for model reporting

- By Mitchell et al, 2019 (Google)
- Machine learning models involved in high-stakes tasks, incl. hiring, law enforcement, health care, education
- Goals:
  - Standardize ethical practice and reporting
  - Allow others to assess and compare models for deployment in terms of performance AND ethical, inclusive and fair considerations
  - Identify systematic errors of model performance before deployment
  - Inform users about what ML systems can and cannot do
  - Types of errors a ML system will make
  - Create more fair and inclusive outcomes with using ML systems

Model cards for model reporting

- Model cards: Transparent model reporting in terms of:
  - Performance characteristics (metrics, what feature impact performance)
  - Intended use contexts
  - Benchmarking (evaluating) human-centric ML systems under predefined conditions, here via **disaggregated evaluation** by unitary and intersectional groups (cultural, demographic, phenotype; incl. race and gender)
- Alternative solutions:
  - Qual and quant algorithmic auditing by 3rd parties
  - Adversarial testing by technical and non-technical analysis
  - Inclusive user feedback

Model cards for model reporting: sections

- Model details
  - Not a requirement to compromise private information or reveal proprietary training methods
- Intended use, incl. out of scope use
- Factors, incl. groups, instrumentation, environment
- Metrics
- Evaluation data
- Training data
- Quantitative analysis, disaggregated (broken down by factors) -> aim for parity (as a dimension of fairness)
- Ethical considerations
- Caveats and recommendations

Other data documentation efforts

- DDI: https://ddialliance.org/: “The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences.”
- Industry-wide documentation of best documentation practices in ML and AI: https://partnershiponai.org/workstream/about-ml/
- Dataset Nutrition Labels: Holland, S., et al. (2020). "The dataset nutrition label." Data Protection and Privacy: Data Protection and Democracy 1.
- Factsheets: Arnold, M., et al. (2019). "FactSheets: Increasing trust in AI services through supplier's declarations of conformity." IBM Journal of Research and Development 63(4/5): 6: 1-6: 13.
    - Characteristics of AI services

Readings on responsible use of social media data

- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.
- Berreby, D (2017). Click to agree with what? No one reads terms of service, studies confirm. https://www.theguardian.com/technology/
- Covey, D.T. (2010). Open Access & Copyright. Carnegie Mellon University. http://works.bepress.com/denise_troll_covey/48
- Creative commons: Lessig, L. (2006). Code: Version 2.0. NY: Basic Books. URL: http://codev2.cc
- Diesner, J., & Chin, C. (2015). Usable Ethics: Practical considerations for responsibly conducting research with social trace data. Workshop: Beyond IRBs: Ethical Review Processes for Big Data Research, Future of Privacy Forum, Washington DC.
- Diesner, J., & Chin, C. (2016). Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. ETHI-CA² Workshop (ETHics in Corpus Collection, Annotation & Application), 10th Language Resources and Evaluation Conference (LREC), Portoroz, Slovenia.
- Diesner, J., & Chin, C. (2016). Seeing the forest for the trees: Understanding and implementing regulations for the collection and analysis of human centered data. Human-Centered Data Science (HCDS) Workshop, 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016), San Francisco, CA.
- Fiesler, C. "Law & Ethics of Scraping: What HiQ v LinkedIn Could Mean for Researchers Violating TOS", 2017 (https://cfiesler.medium.com/law-ethics-of-scraping-what-hiq-v-linkedin-could-mean-for-researchers-violating-tos-787bd3322540)
- Fiesler, C., Beard, N., & Keegan, B.C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. Proceedings of the international AAAI Conference on Web and Social Media (ICWSM).
- Fiesler, C., et al. (2016). Reality and perception of copyright terms of service for online content creation. Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing.
- Fiesler, Casey. (2019). Scientists Like Me Are Studying Your Tweets Are You OK With That. https://howwegettonext.com/scientists-like-me-are-studying-your-tweets-are-you-ok-with-that-c2cfdfebf135
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist, 70(6), 543-556.
- Nissenbaum, H. (2011). A contextual approach to privacy online. Daedalus 140(4): 32-48.
- Nissenbaum, H. (2020). Privacy in context, Stanford University Press.
- Todd, G.V. & Crews, K. (2004). An Intellectual Property Primer: Protecting Your Investment with Copyright, Patent, Trade Secret, and Trademark. Presented at the Indiana Health Industry Forum, Intellectual Property Workshop, June 16, 2004. http://www.copyright.iupui.edu/IPPrimer.pdf
- Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. Ethics and information technology 12(4): 313-325.
- Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. Social Media+ Society 4(2).

2017/mar/03/terms-of-service-online-contracts-fine-print

Vaccaro, K., et al. (2015). Agree or cancel? research and terms of service compliance. ACM CSCW Ethics Workshop: Ethics for Studying Sociotechnical Systems in a Big Data World.

Readings on data and model documentation

- Bender, E. M. and B. Friedman (2018). "Data statements for natural language processing: Toward mitigating system bias and enabling better science." Transactions of the Association for Computational Linguistics 6: 587-604.
- Boyd, D. and K. Crawford (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, communication & society 15(5): 662-679.
- Connaway, L.S. & M.L. Radford (2016). Research Methods in Library and Information Science, Libraries Unlimited, 6th Edition (available as e-book from library)
- Gebru, T., et al. (2018). "Datasheets for datasets." arXiv preprint arXiv:1803.09010.
- Hind, M., et al. (2018). "Increasing trust in AI services through supplier's declarations of conformity." arXiv preprint arXiv:1808.07261.
- Holland, S., et al. (2020). "The dataset nutrition label." Data Protection and Privacy: Data Protection and Democracy (2020)
- Mitchell, M., et al. (2019). Model cards for model reporting. Proceedings of the conference on fairness, accountability, and transparency.
- Olteanu, A., et al. (2019). "Social data: Biases, methodological pitfalls, and ethical boundaries." Frontiers in Big Data 2: 13.
- Sen, I., et al. (2021). "A total error framework for digital traces of human behavior on online platforms." Public Opinion Quarterly.

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/CIKM2022/
- 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]