Social Media
Information Extraction
multi-task, multi-lingual, & multi-contextual

Shubhanshu Mishra
Sr. Machine Learning Researcher
Content Understanding Research, Twitter
https://shubhanshu.com
https://socialmediaie.github.io/

Slides at: https://shubhanshu.com/talks

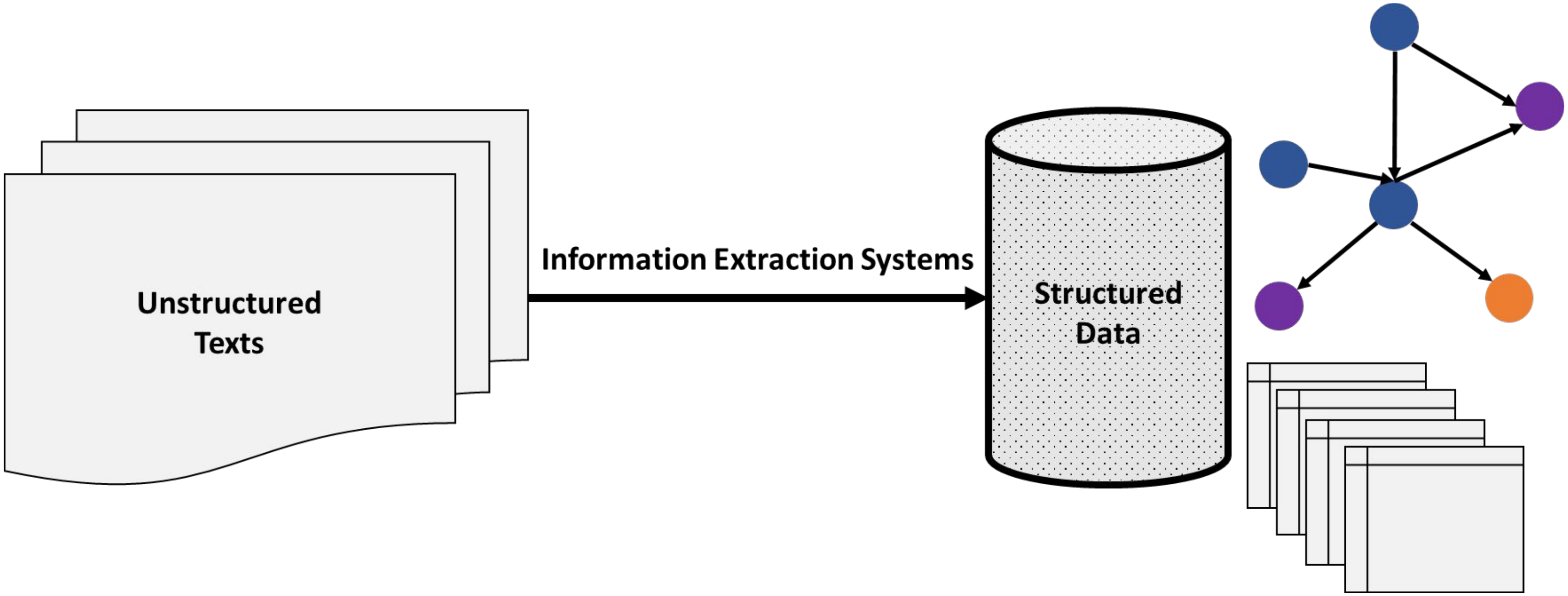
*** Most work presented here was done during my PhD at UIUC with multiple collaborators.**
Work done at twitter will be marked with Twitter logo.
Content and views expressed in this talk are solely the responsibility of the presenter.

Outline

- Definitions:
  - Information Extraction (IE)
  - Social Media
  - Digital Social Trace Data - DSTD
- Challenge of Social Media IE
- Tasks
  - Text Classification: Topics, Sentiment, Spam
  - Token Level Classification: NER + Linking, Phrases, Command Word Extractions
  - Document Similarity and Ranking: Search, Recommendations
- Applications
- Datasets
- Challenges
  - Less data to learn: Solution - Multi-task learning to improving efficiency
  - Less languages to learn: Solution - Multilingual learning to improve coverage
  - Less context to learn: Solution - LMSOC, NTULM
- Notes on bias of ML systems
  - NER Bias
- Conclusion

Definitions

Information extraction (IE)



"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

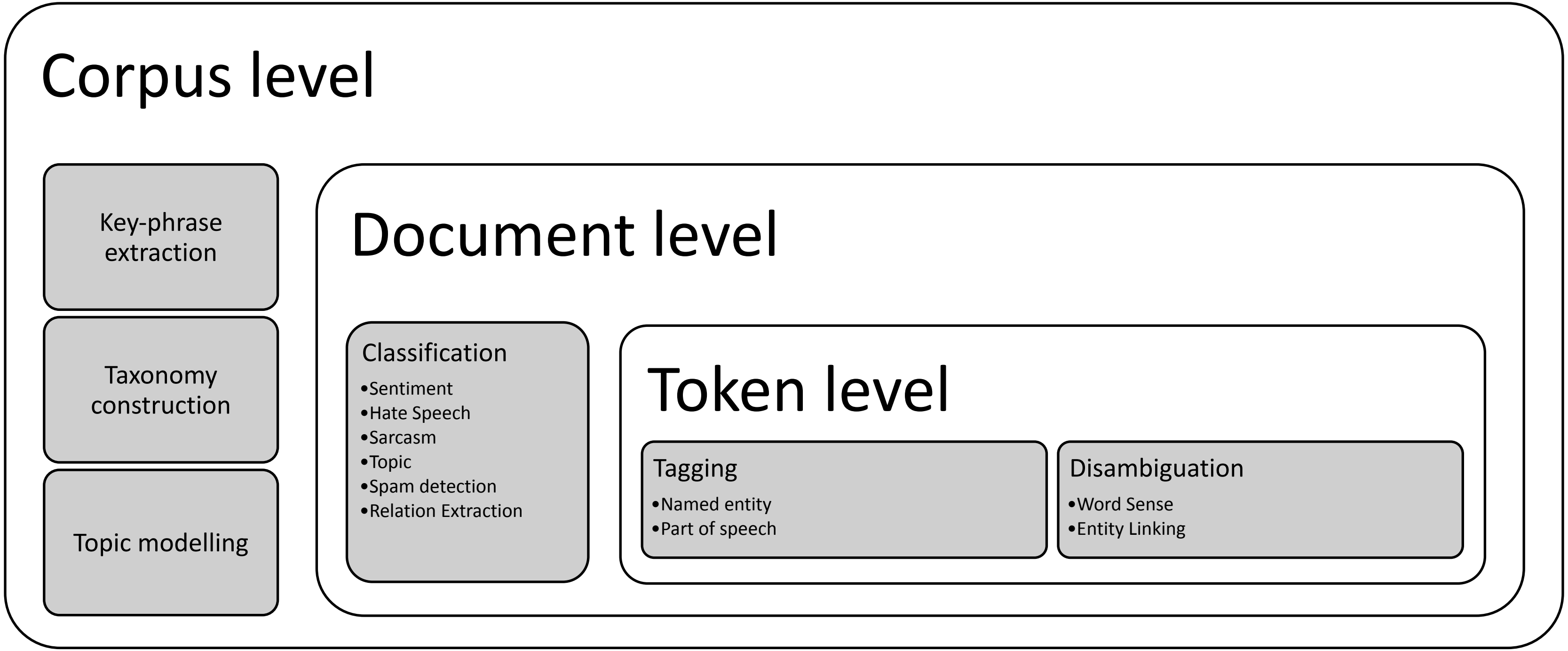
- **Messages sorted by:** [date][thread][subject][author]
- **Next message:** cyn@prism.nmt.edu: "Need help!"
- **Previous message:** jremick@u.washington.edu: "Where I am in here"

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]

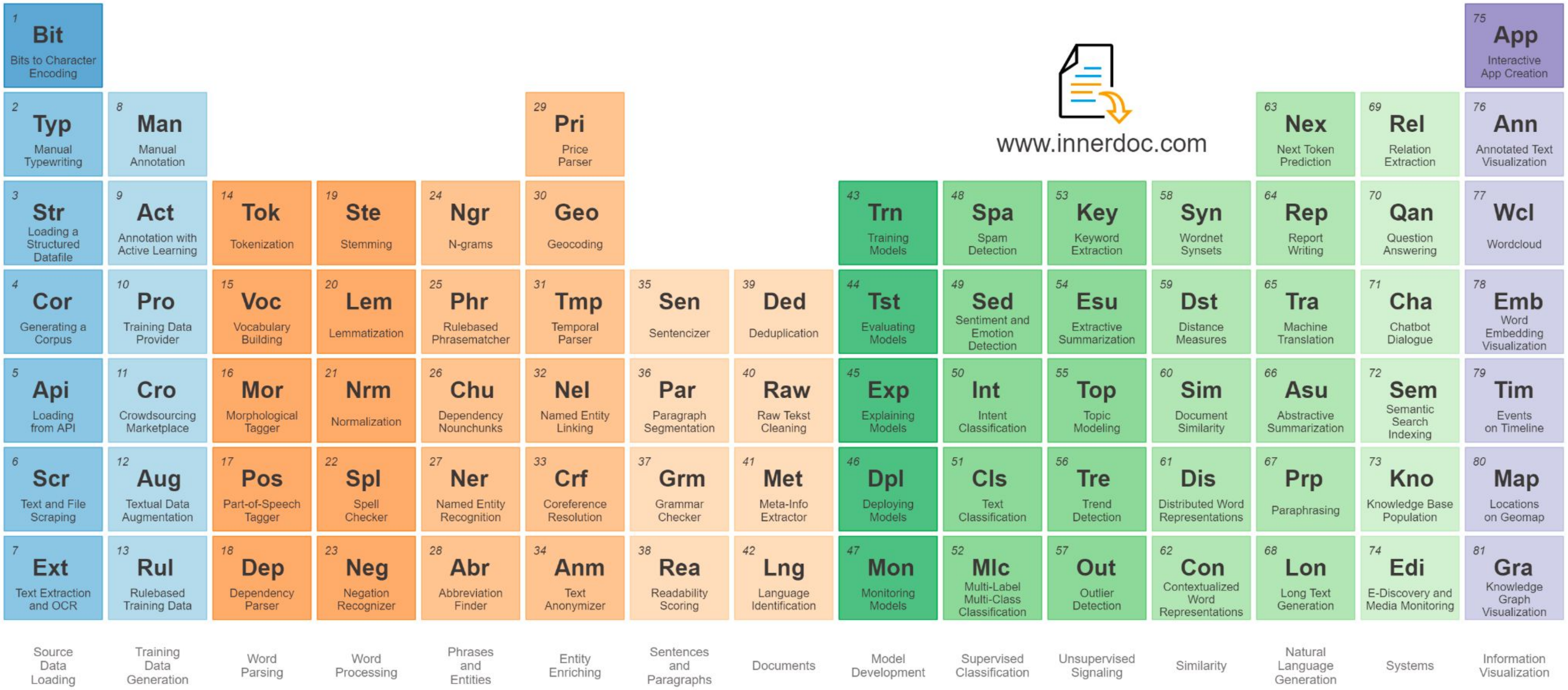
- **Next message:** [redacted] "Need help!"
- **Previous message:** [redacted] "Where I am in here"

1995 - Usenet

Information extraction tasks



Periodic Table of Natural Language Processing Tasks



Source: https://www.innerdoc.com/periodic-table-of-nlp-tasks/

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

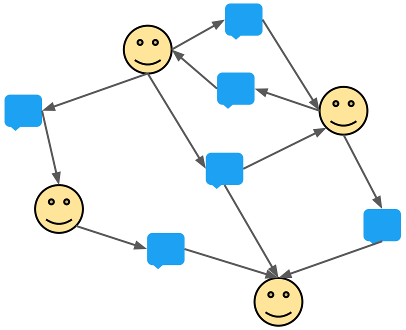
john oliver coined the term donal drumph as a joke on his show #LastWeekTonight

Predict

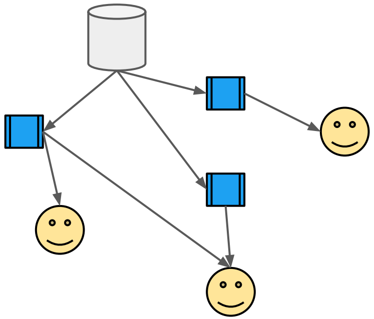
Output

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	

Social 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)**."

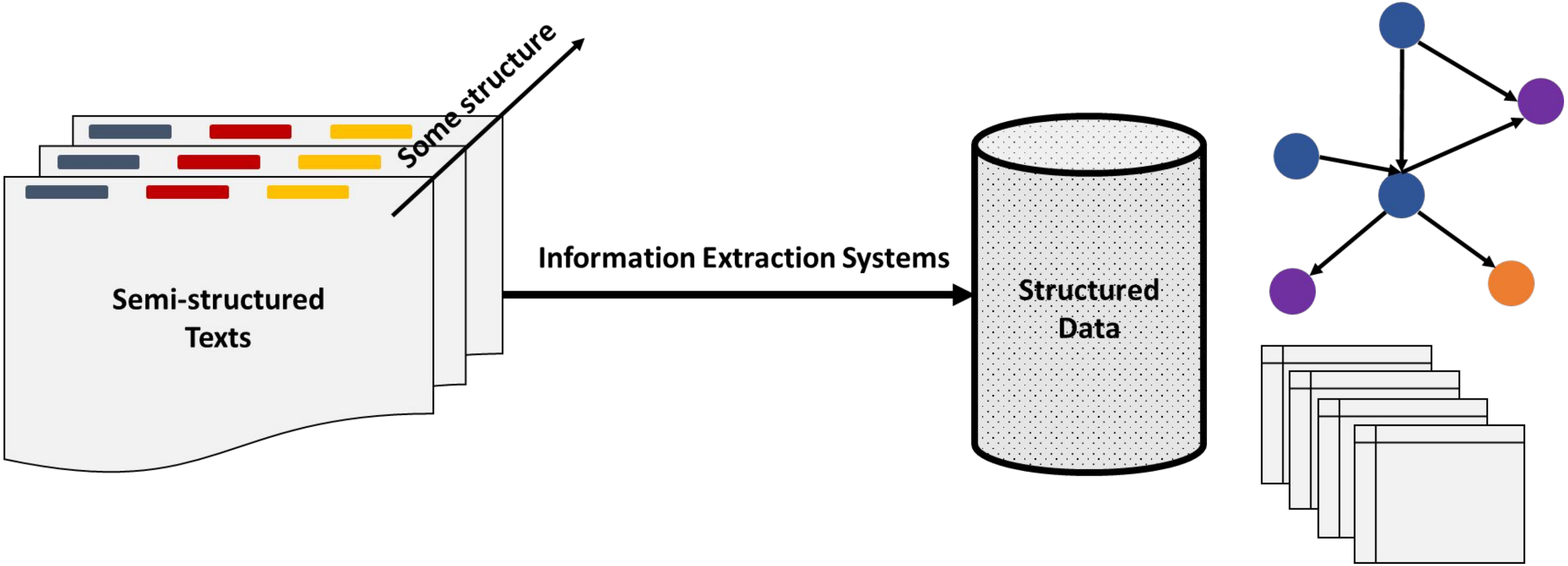
"For instance, a newspaper is delivered to many subscribers and a radio station broadcasts the same programs to an entire city."

"**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]

Information extraction from semi-structured data



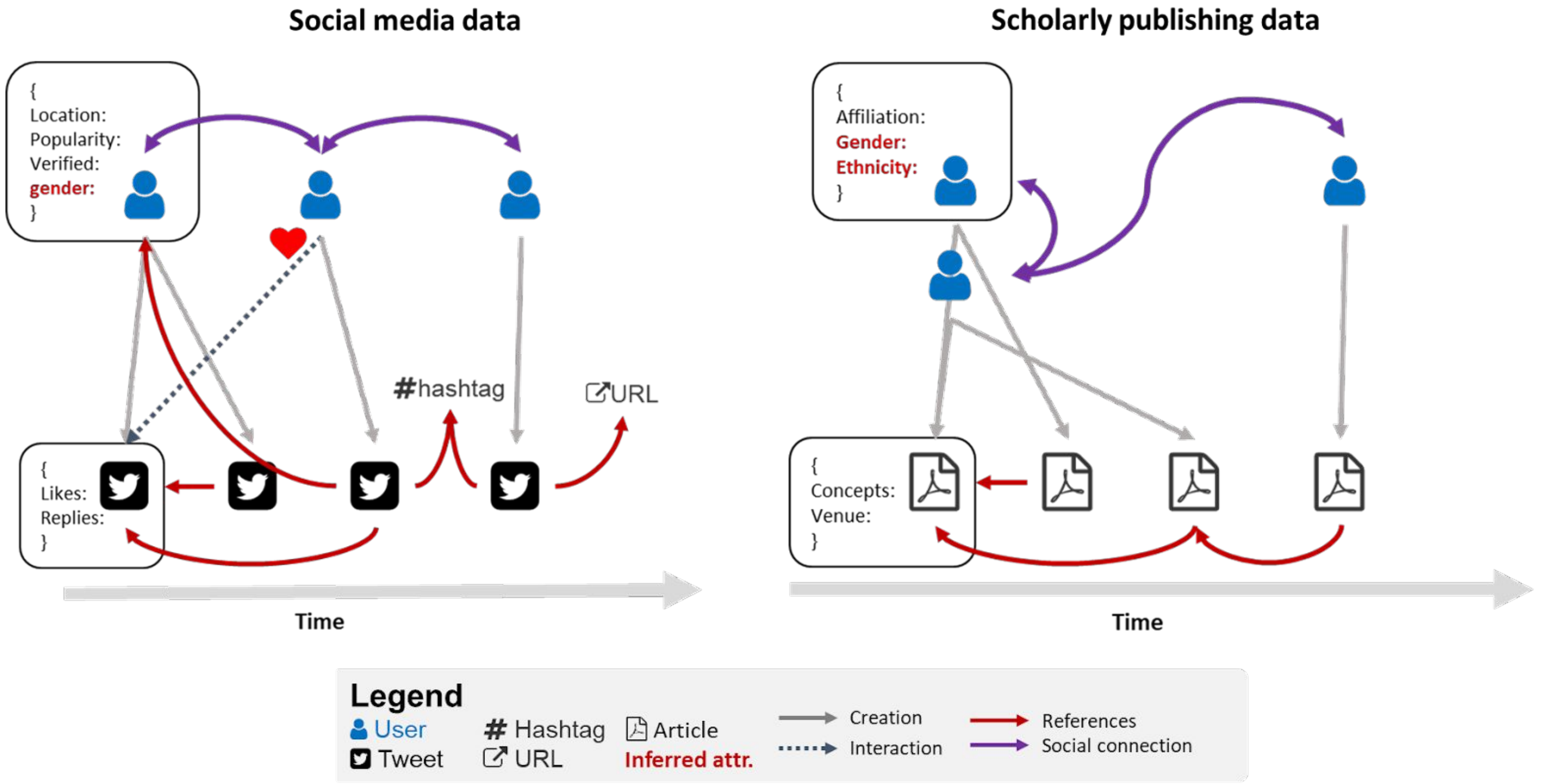
However, not all data is unstructured. Many datasets of interest have some inherent structure imposed because of the data generating process.

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)



DSTD properties and examples

Property	Social Media	Scholarly data
Temporal information associated with each item of the data	Tweets ordered by time	Scholarly papers ordered by time
Presence of connection between various data items	User authors tweets, tweet are quoted in other tweets	Authors connected to papers, papers cite other papers
Optionally associated meta-data for data items	Likes, retweets, followers, location	Venue, topics, key words

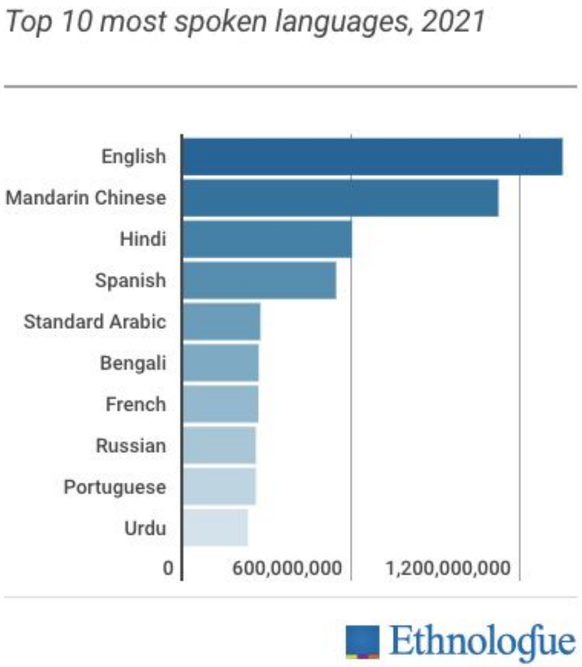
Challenge of Social Media IE

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



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。

Example of Named Entity Recognition on tweets

Here we go - Arsenal v Tottenham at Meadow Park! 💪

7:00 AM · Aug 25, 2019 · Twitter for iPhone

Twitter Specific Model

Here we go - Arsenal Organization 0.966 v Tottenham Organization 0.954 at Meadow Park Place 0.929 !

SpaCy (Open-source)

Here we go - Arsenal vs Tottenham PERSON at Meadow Park!

Google Natural Language API

Here we go - ⟨Arsenal⟩$_2$ ⟨v⟩$_1$ ⟨Tottenham⟩$_3$ at ⟨Meadow Park⟩$_4$!

OTHER 1. v Salience: 0.39	ORGANIZATION 2. Arsenal Wikipedia Article Salience: 0.23
LOCATION 3. Tottenham Wikipedia Article Salience: 0.22	LOCATION 4. Meadow Park Salience: 0.16

Source: Named Entity Recognition at Scale with Deep Learning Sijun He @SijunHe #TwitterCortex at #ODSCWest2019

NER performance difference

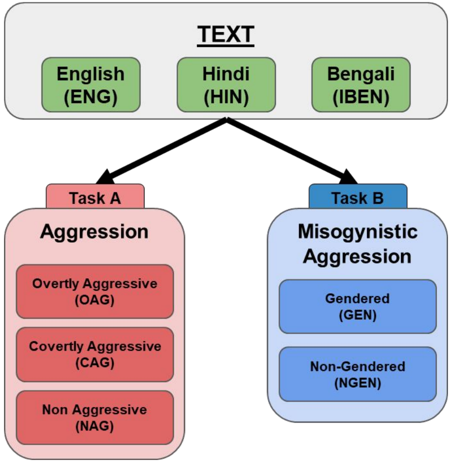
Named entity recognition performance over the evaluation partition of the Ritter dataset (best score in bold).

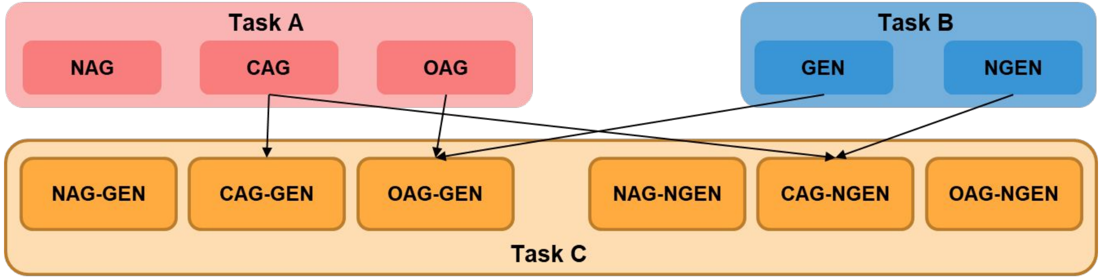
	Per-entity F1				Overall		
System	Location	Misc	Org	Person	P	R	F1
ANNIE	40.23	0.00	16.00	24.81	36.14	16.29	22.46
DBpedia Spotlight	46.06	6.99	19.44	48.55	34.70	28.35	31.20
Lupedia	41.07	13.91	18.92	25.00	38.85	18.62	25.17
NERD-ML	**61.94**	23.73	**32.73**	**71.28**	52.31	**50.69**	**51.49**
Stanford	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

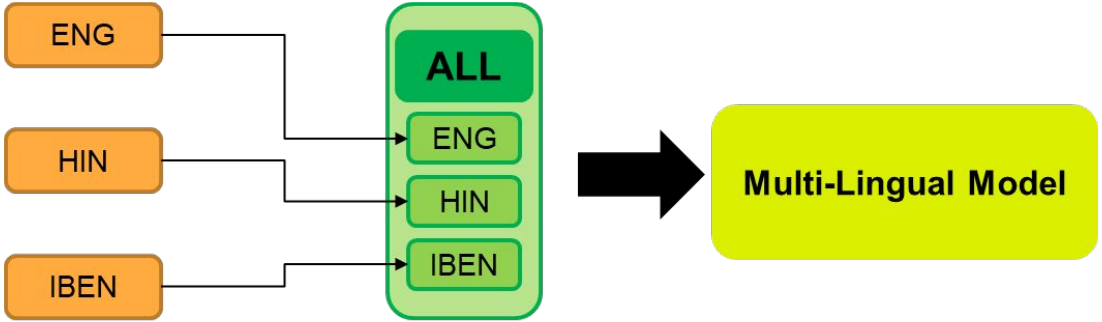
Tasks

Multilingual transformer models for hate and abusive speech





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



https://github.com/socialmediaie/TRAC2020

Named Entity Recognition (NER) on Tweets

Official ACM
@TheOfficialACM

Yoshua Bengio, Geoffrey Hinton and Yann LeCun, the fathers of #DeepLearning, receive the 2018 #ACMTuringAward for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing today. bit.ly/2HVJtdV

Real Madrid C.F. @realmadrid · Sep 5

Los jugadores del Real Madrid y del Castilla han guardado un minuto de silencio por el fallecimiento de Blanca Fernández Ochoa, medallista olímpica y leyenda del deporte español.

Yusaku Maezawa (MZ) 前澤友作
@yousuck2020

ZOZOTOWN新春セールが史上最速で取扱高100億円を先ほど突破！！日頃の感謝を込め、僕個人から100名様に100万円【総額1億円のお年玉】を現金でプレゼントします。応募方法は、僕をフォローいただいた上、このツイートをRTするだけ。受付は1/7まで。当選者には僕から直接DMします！ #月に行くならお年玉

Person
Location
Organization
Product
Other

Source: Named Entity Recognition at Scale with Deep Learning Sijun He @SijunHe #TwitterCortex at #ODSCWest2019

Applications

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

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

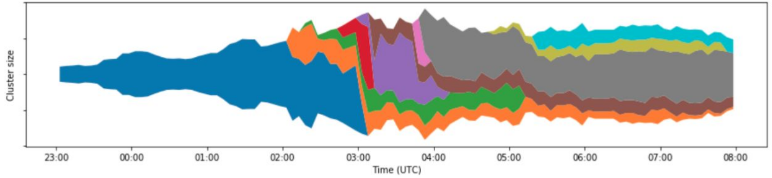
RPG Site
@RPGSite

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

Source: [Named Entity Recognition at Scale with Deep Learning Sijun He @SijunHe #TwitterCortex at #ODSCWest2019]

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] | Source: [Named Entity Recognition at Scale with Deep Learning Sijun He @SijunHe #TwitterCortex at #ODSCWest2019]

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

Inspired from Source: Named Entity Recognition at Scale with Deep Learning Sijun He @SijunHe #TwitterCortex at #ODSCWest2019

Datasets

Where is the data?

- **MetaCorpus**: A list of curated annotated datasets for various social media tasks and social media platforms. https://github.com/socialmediaie/MetaCorpus
- **MetaCorpus - benchmark**: A selected set of datasets which can be used for benchmarking multi-task learning or NLP for social media data

Text classification	
Sentiment	Datasets described in [32]
Abusive	Founta [19], WaseemSRW [44]
Uncertainity	**Sarcasm**: Riloff [36]; **Veridicality**: Swamy [42]
Sequence Tagging	
PoS tagging	**ark**: Owoputi [33, 34]; **ptb**: TwitIE [15] and Ritter [37]; **ud**: Tweetbankv2 [27], DiMSUM2016 [39], Foster [22], and lowlands [22, 23]
NER	Ritter [37],; WNUT 2016 [41], WNUT 2017 [14] Finin [18], Hege [20] , Broad [12], MultiModal dataset [46], YODIE [21], MSM2013 [7], and NEEL2016 [38]
Chunking	Ritter [37]
Supersense tagging	Ritter [37] and Johansen2014 [25]

Table 1: List of datasets used in our multi-dataset-multi-task learning models.

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

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

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

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

TweetNERD - End to End Entity Linking Benchmark for Tweets

[TweetNERD - End to End Entity Linking Benchmark for Tweets | Zenodo]

Largest dataset for Entity Linking for Tweets: 340K tweets annotated with Mentions and Entities Linked to Wikidata.

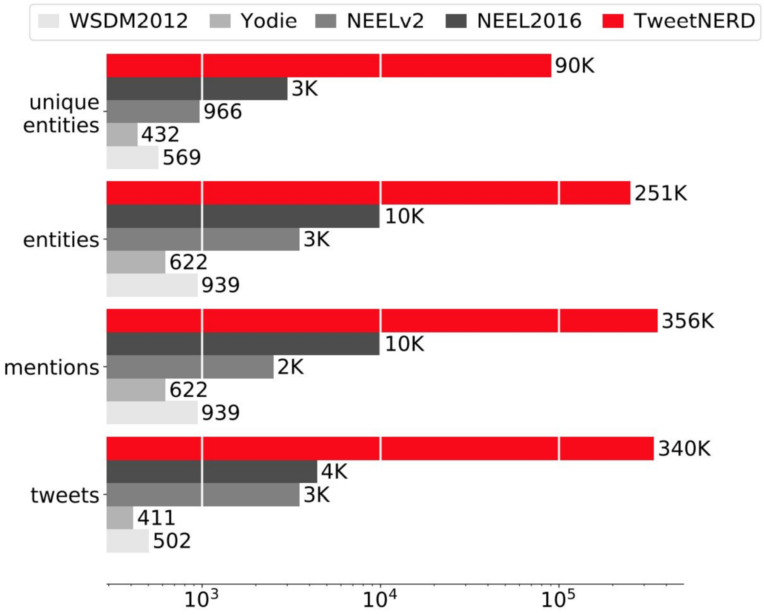


Figure 1: Comparison with existing Tweet entity linking datasets

TweetNERD - End to End Entity Linking Benchmark for Tweets

Table 3: Details of `TweetNERD-Academic` (same Tweet could occur in multiple datasets).

dataset	Tasks	Total Tweets	Found Tweets	Found %
Tgx [Dredze et al., 2016]	CDCR	15,313	9,790	63.9
Broad [Derczynski et al., 2016]	NER	8,633	6,913	80.1
Entity Profiling [Spina et al., 2012]	NER	9,235	6,352	68.8
NEEL 2016 [Rizzo et al., 2016]	NERD	9,289	2,336	25.1
NEEL v2 [Yang and Chang, 2015]	NERD	3,503	2,089	59.6
Fang and Chang [2014]	NERD	2,419	1,662	68.7
Twitter NEED [Locke, 2009]	NERD & IR	2,501	1,549	61.9
Ark POS [Gimpel et al., 2011]	POS	2,374	1,313	55.3
WikiD	NED	1,000	504	50.4
WSDM2012 [Meij et al., 2012]	Relevance	502	415	82.7
Yodie [Gorrell et al., 2015]	NERD	411	288	70.1

TweetNERD - End to End Entity Linking Benchmark for Tweets

Table 5: Evaluating `TweetNERD-OOD` and `TweetNERD-Academic` using existing systems.

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 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[3]	0.307	0.458	0.223	0.379

(c) End to end entity linking F1 scores.

Challenges

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: Multi-task learning to improving efficiency

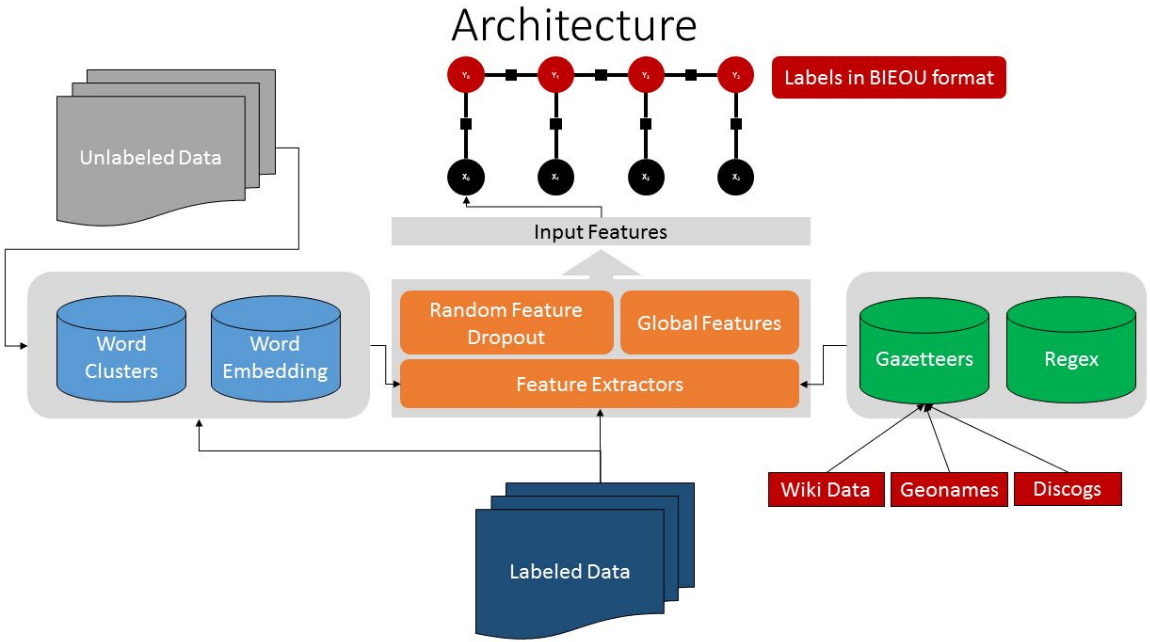
Multi-task learning

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/

System Name Precision Recall F1 Score

Evaluating Twitter NER (F1-score)

Mishra & Diesner (2016).

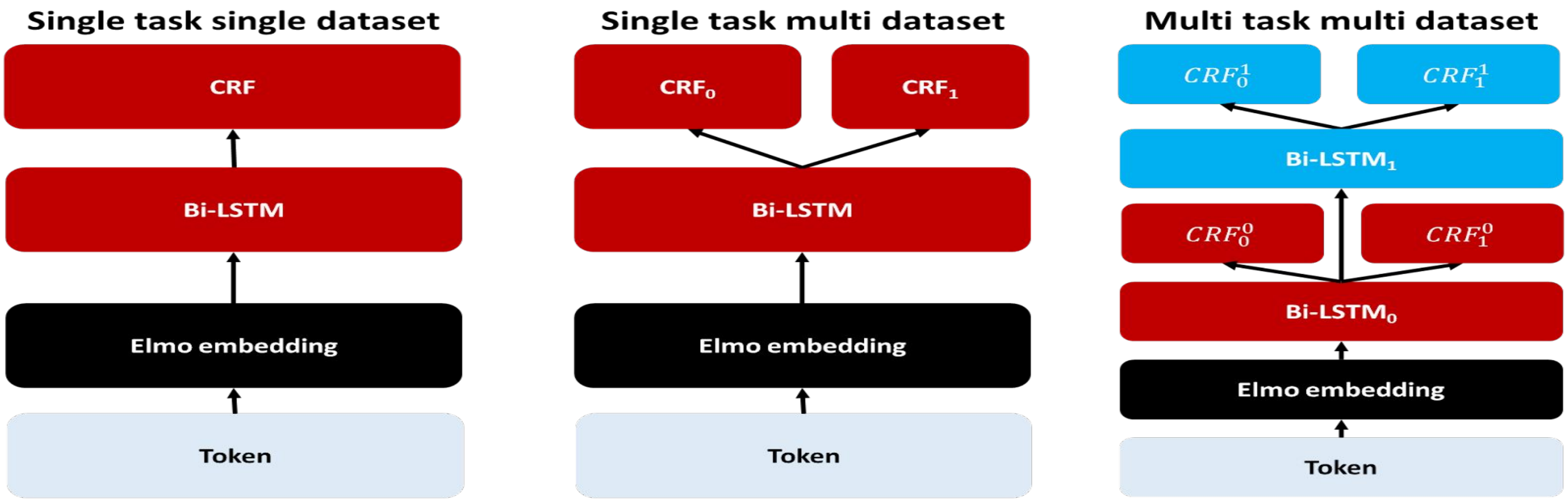
Rank	TD	TDT_E
10-types	**46.4**	**47.3**
No-types	**57.3**	**59.0**
company	42.1	46.2
facility	37.5	34.8
geo-loc	70.1	71.0
movie	0.0	0.0
music artist	7.6	5.8
other	31.7	32.4
person	51.3	52.2
product	10.0	9.3
sportsteam	31.3	32.0
tvshow	5.7	5.7

System Name	Precision	Recall	F1 Score
Stanford CoreNLP	0.526838069	0.453416149	0.487377425
Stanford CoreNLP (with Twitter POS tagger)	0.526838069	0.453416149	0.487377425
TwitterNER	**0.661496966**	0.380822981	0.483370288
OSU NLP	0.524096386	0.405279503	0.45709282
Stanford CoreNLP (with caseless models)	0.547077922	0.392468944	0.457052441
Stanford CoreNLP (with truecasing)	0.413084823	0.421583851	0.417291066
MITIE	0.340364057	0.457298137	0.390260063
spaCy	0.28426543	0.380822981	0.325535092
Polyglot	0.273080661	0.327251553	0.297722055
NLTK	0.149006623	0.331909938	0.205677171
TwitterNER (with Hege training data)	0.657213317	0.413819876	0.507860886
TwitterNER (with W-NUT 2017 training data)	0.675307842	0.404503106	0.505948046
TwitterNER (with Finin training data)	0.598086124	0.388198758	0.470809793
TwitterNER (with W-NUT 2017 and Hege training data)	0.652276759	0.42818323	0.51699086

Source: https://blog.maxar.com/earth-intelligence/2017/named-entity-recognition-for-twitter

Code: https://github.com/humangeo/twitter-ner-eval

Multi-task-multi-dataset learning Mishra 2019, HT' 19



(A) (B) (C)

S - Single

MD – Multi-dataset
MTS – Multi task Shared

MTL – Multi task Stacked (Layered)

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

Evaluating MTL models Mishra 2019, HT' 19

Part of speech tagging (overall accuracy)

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

Super sense tagging (micro f1)

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

Chunking (micro f1)

Data	Our best	SOTA	Diff %
Ritter	88.92	None	NA

Named entity recognition (micro f1)

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

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

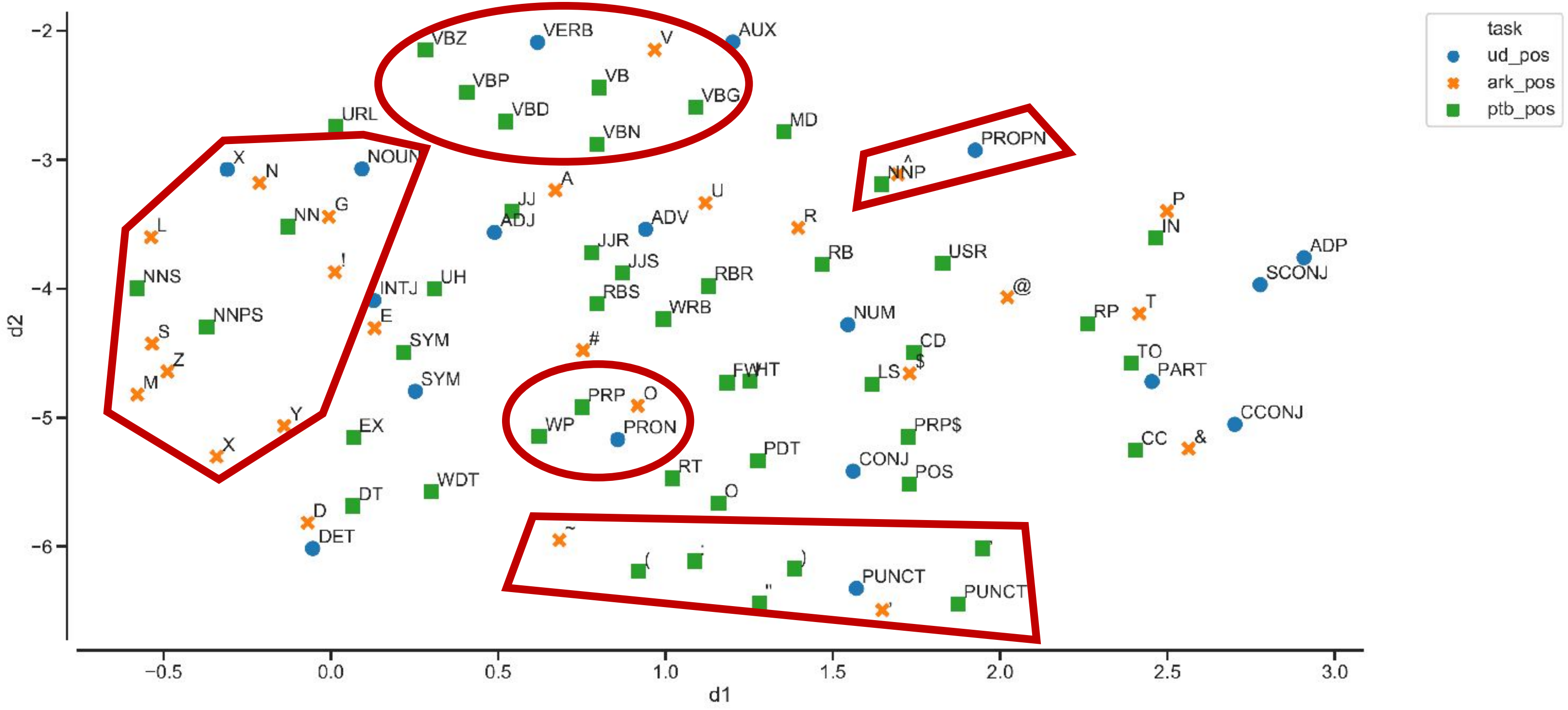
Training Mishra 2019, HT' 19

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

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

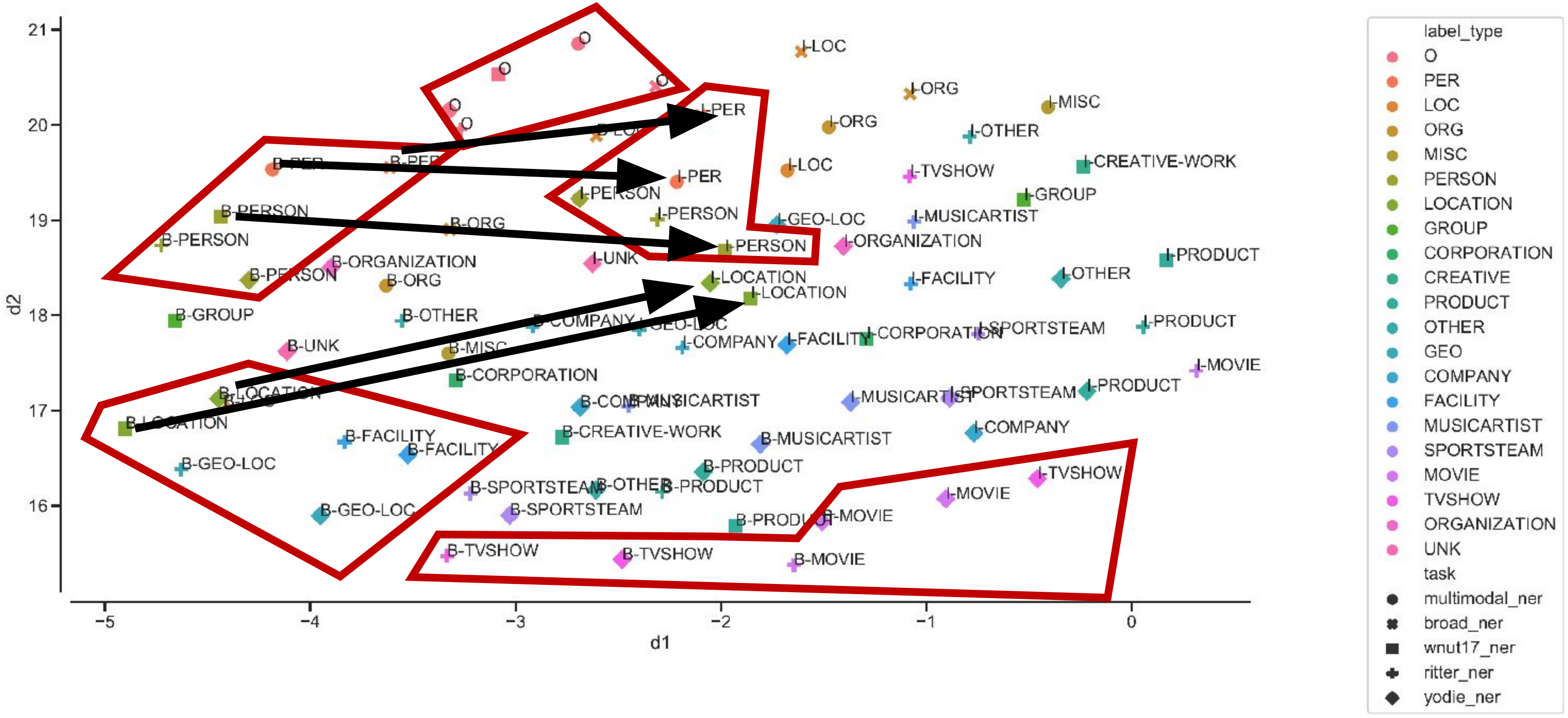
Label embeddings (POS)

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



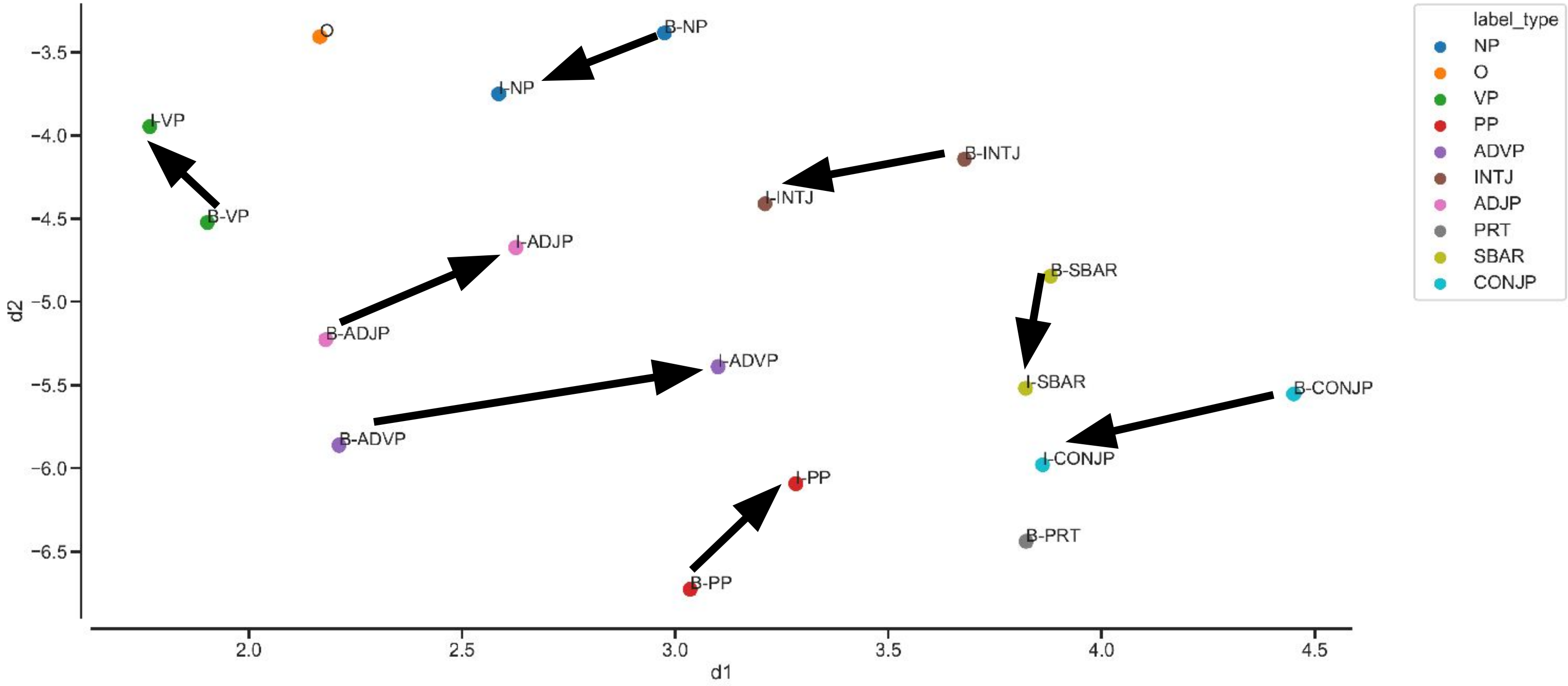
Label embeddings (NER)

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

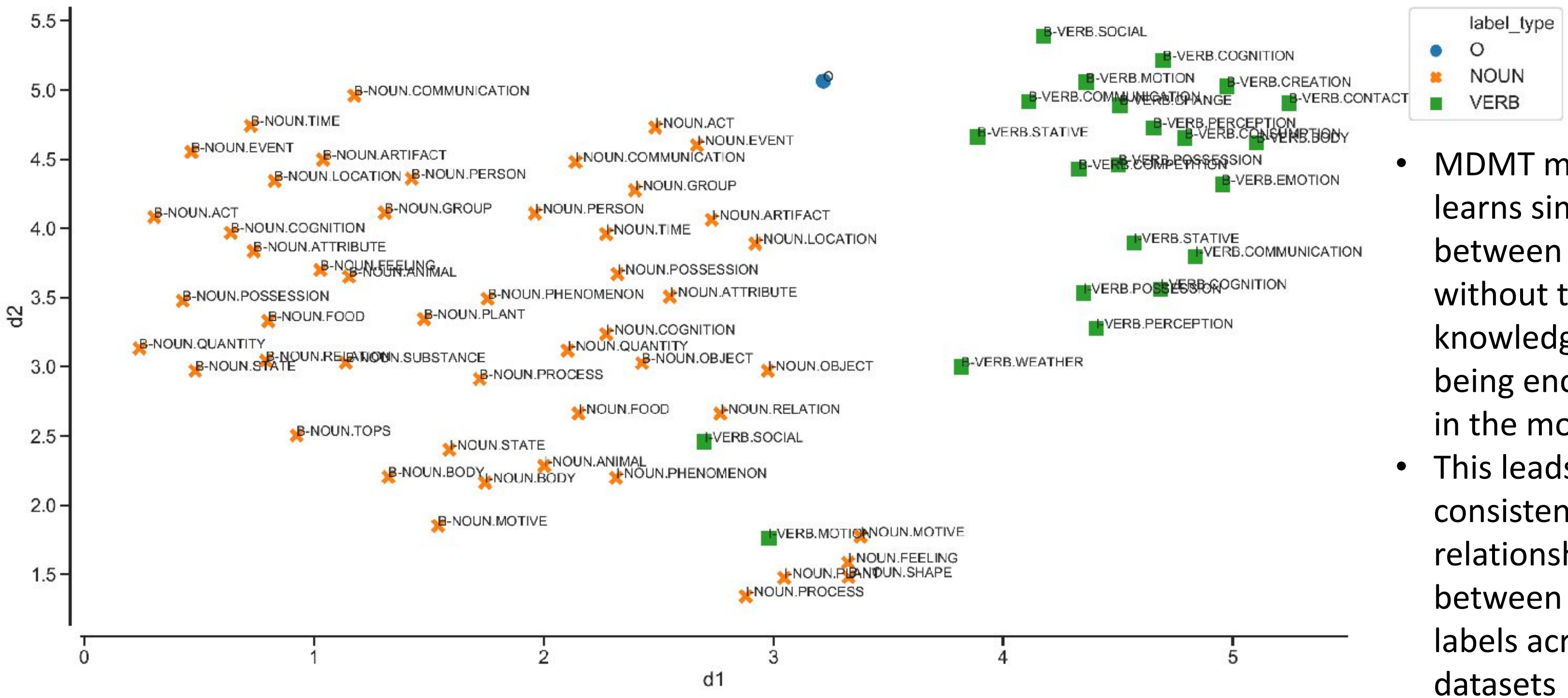


Label embeddings (chunking)

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

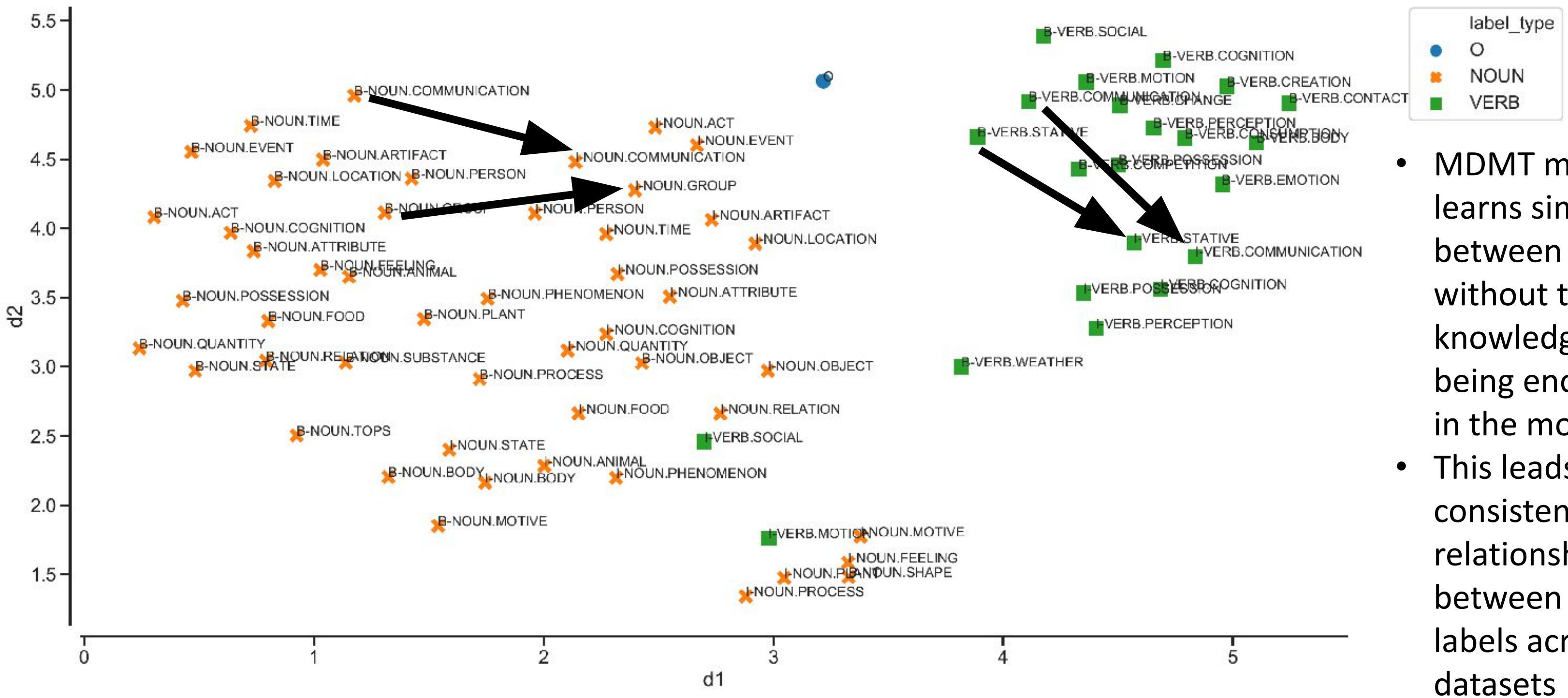


Label embeddings (super-sense tagging)



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

Label embeddings (super-sense tagging)



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

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

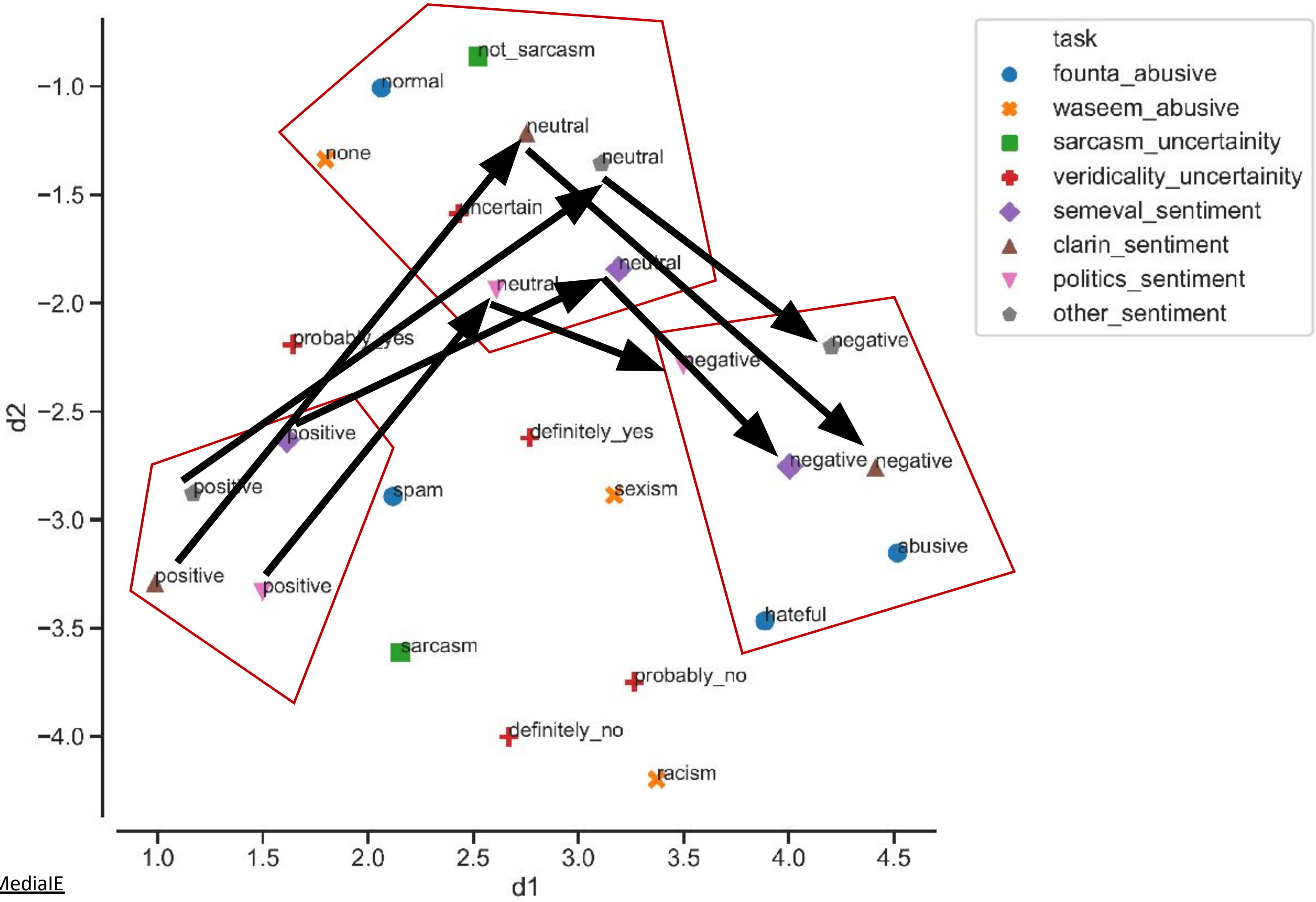
file	Founta		WaseemSRW	
model	r	v	r	v
S bilstm	8	79.33	8	81.72
MD bilstm	9	79.03	9	81.31
MTS bilstm	11	61.48	11	68.57
MTL bilstm	10	69.26	10	70.13
S bilstm *	1	**80.6**	3	**82.95**
MD bilstm *	2	**80.35**	2	**83.22**
MTS bilstm *	6	80.11	7	81.99
MTL bilstm *	4	80.23	5	82.78
S cnn *	3	**80.25**	4	82.89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

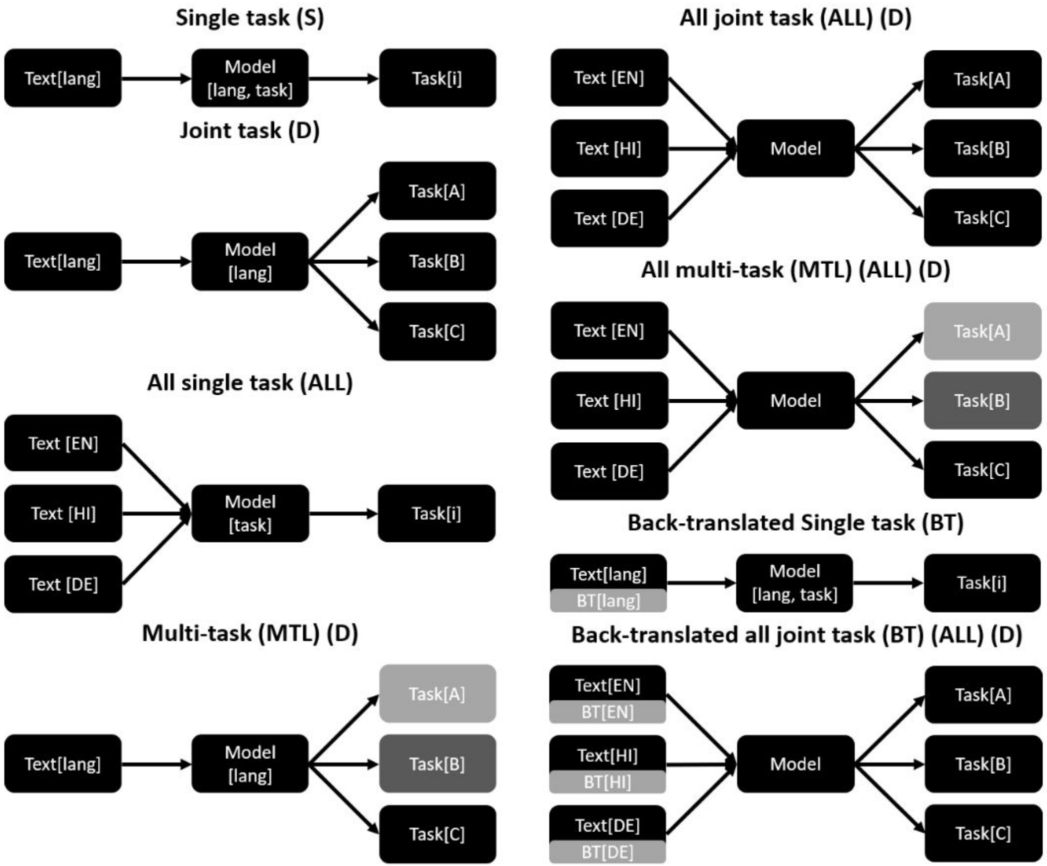


Fig. 2: An overview of various model architectures we used. Shaded task boxes represent that we first compute a marginal representation of labels only belonging to that task before computing the loss.

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

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

Less languages to learn: Multilingual learning with lang families

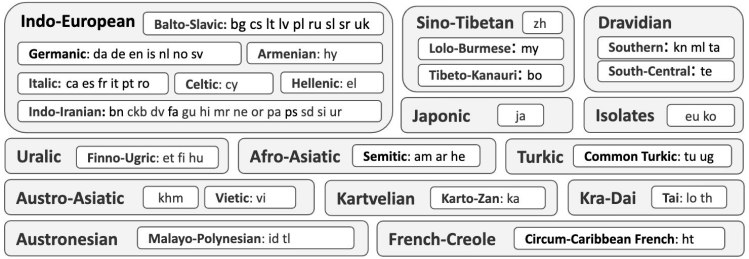


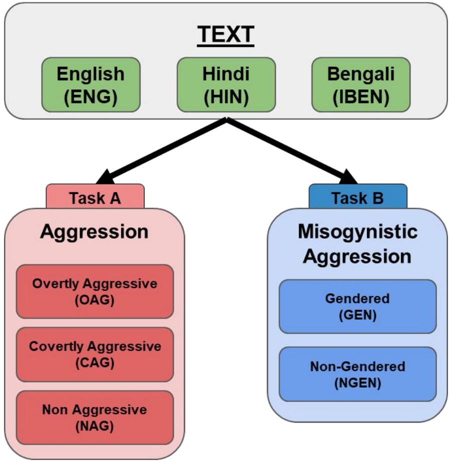
Figure 1: Our training languages, grouped into their families and sub-families

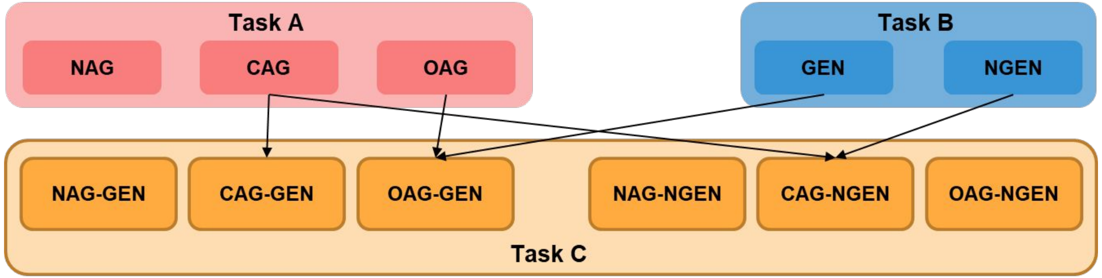
Lang.	Dataset	Monolingual			Mulilingual (Family-Based)			Multilingual (All-in-One)		
		mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE
en	CONLL'03	41.8	40.7	**43.1**	40.1	38.9	**42.9**	**37.9**	36.0	33.3
en	INH*	38.0	**43.2**	42.3	34.1	**42.5**	36.8	32.8	**38.6**	27.5
de	CONLL'03	44.9	42.0	**46.4**	42.3	40.9	**44.2**	38.1	**38.8**	29.0
nl	CONLL'02	44.5	43.3	**50.7**	**46.8**	43.6	42.2	**41.2**	35.8	25.2
es	CONLL'02	**31.2**	30.5	27.6	**31.5**	27.5	29.0	**29.0**	27.4	24.8
es	INH*	40.3	**41.8**	39.7	35.9	**39.0**	33.1	32.4	**37.2**	24.8
pt	INH*	33.0	**41.2**	38.1	29.1	**36.2**	26.3	27.6	**33.9**	18.5
fr	EuropeanaNP	**36.4**	35.4	34.4	**33.6**	31.3	29.7	**28.1**	26.8	22.0
it	xLiMe*	14.4	**17.7**	16.3	14.4	**18.9**	16.6	16.3	**19.3**	16.3
hi	SSEA	26.4	30.6	**33.7**	19.0	20.1	**29.4**	**19.1**	17.1	9.1
ur	SSEA	17.9	16.5	**20.5**	14.7	16.6	**19.6**	15.6	12.3	**15.8**
bn	SSEA	25.1	21.2	**45.3**	19.1	18.9	**36.8**	16.5	18.9	**19.3**
ar	Code-Switch'18*	26.8	**28.0**	27.6	23.4	25.5	**28.9**	21.9	**23.0**	23.0
ar	INH*	16.0	**20.4**	16.4	14.1	**20.7**	15.7	11.4	**16.2**	10.8
ja	INH	17.3	**23.9**	18.5	NA	NA	NA	17.2	**20.3**	15.1
tr	JRC*	31.5	**37.6**	31.2	NA	NA	NA	26.9	**32.1**	28.0
te	SSEA	13.0	10.8	**17.6**	NA	NA	NA	12.0	6.6	**18.0**
Average (Tweets)		27.2	**31.7**	28.7	25.2	**30.5**	26.2	23.3	**27.6**	20.5
Average (IEG)		42.3	42.3	**45.6**	40.8	41.5	**41.5**	**37.5**	37.3	28.8
Average (IEI)		31.1	**33.3**	31.2	28.9	**30.6**	26.9	26.7	**28.9**	21.3
Average (IEII)		23.1	22.8	**33.2**	17.6	18.5	**28.6**	**17.1**	16.1	14.7
Average (All)		29.3	30.9	**32.3**	28.4	30.0	**30.8**	24.9	**25.9**	21.2

Table 2: NER Results (entity-level micro-averaged F1) without the addition of the WikiAnn training sets. The best result per experimental pair ({test set, learning setting}) is in **bold**. The best result per test set is underlined. Tweet datasets are denoted by *. IEG = Indo-European, Germanic. IEI = Indo-European, Italic. IEII = Indo-European, Indo-Iranian.

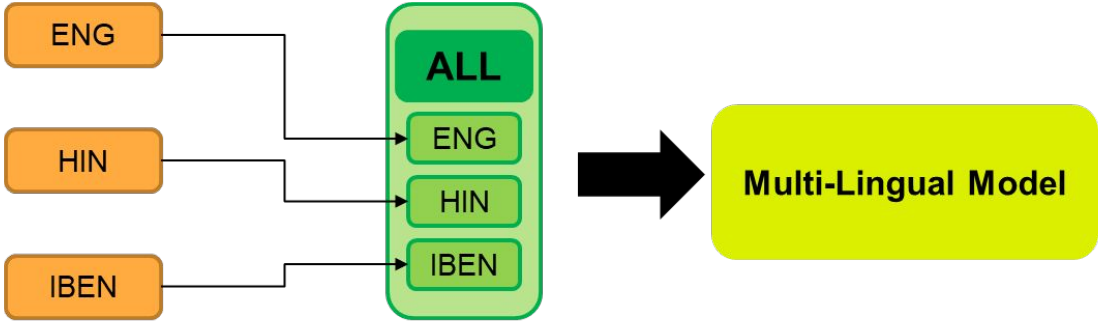
Table Source: Ramy Eskander et. al. Towards Improved Distantly Supervised Multilingual Named-Entity Recognition for Tweets (To appear at MRL EMNLP 2022)

Multilingual transformer models for hate and abusive speech





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



https://github.com/socialmediaie/TRAC2020

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.

Less context to learn: Include tweet context

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?

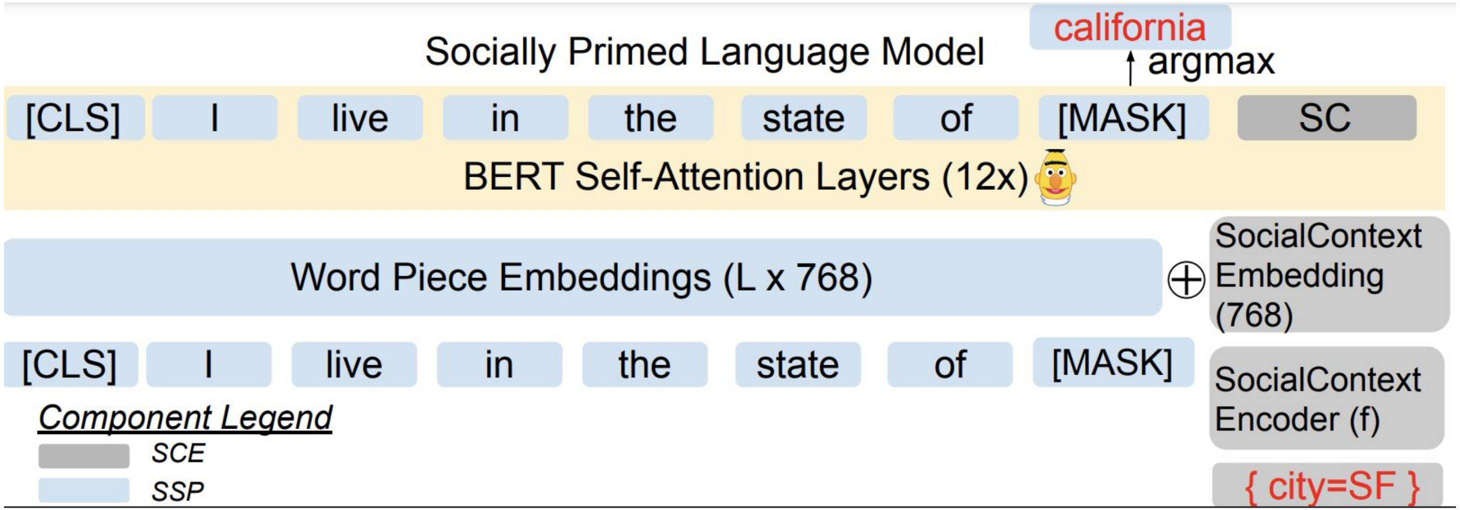
What we use

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?

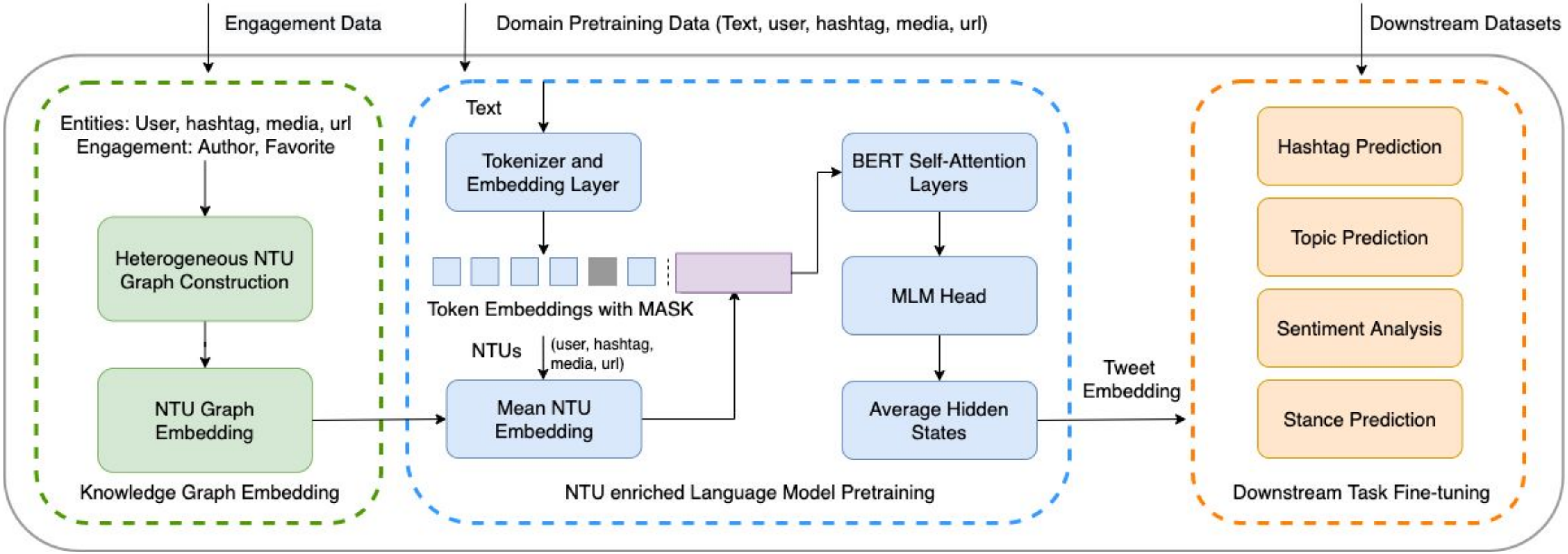
Mishra, S., & Diesner, J. (2018, July 3). Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. Proceedings of the 29th on Hypertext and Social Media. HT '18: 29th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3209542.3209562

Less context to learn: Include tweet context: LMSOC



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.

Less context to learn: Include tweet context: NTULM



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

Less context to learn: Include tweet context: NTULM

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, Vivek Kulkarni. 2022. NTULM: Enriching Social Media Text Representations with Non-Textual Units. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2022). Association for Computational Linguistics.

Bias of ML systems

Bias in Natural Language Processing

Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).	✓	✓	✓	✓

Source: *Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. "Mitigating gender bias in natural language processing: Literature review." DOI 10.18653/v1/P19-1159 (2019).*

NER Bias

	CNET	ELMo	GloVe	corenlp	spacy_lg	spacy_sm
			WINOGENDER			
Black Female	0.7039	0.8942	0.8931	0.7940	0.8908	0.3043
Black Male	0.8410	0.8986	0.9015	0.8862	0.7831	0.3517
Hispanic Female	0.8454	0.8308	0.8738	0.8626	0.8378	0.3726
Hispanic Male	0.8801	0.8603	0.7942	0.8629	0.8151	0.4628
Muslim Female	0.8537	0.8130	0.9074	0.8747	0.8287	0.4285
Muslim Male	0.7791	0.9265	0.9351	0.9477	0.8285	0.4976
White Female	0.9627	0.9116	0.9679	0.9723	0.9577	0.5574
White Male	0.9644	0.9068	0.9700	0.9688	0.9260	0.7732
OOV Name	0.4658	0.9318	0.7573	0.7724	0.2994	0.0824
			IN-SITU			
Black Female	0.8289	0.8802	0.9193	0.8134	0.6732	0.2104
Black Male	0.8964	0.8800	0.9206	0.8828	0.5922	0.2651
Hispanic Female	0.8934	0.8510	0.9091	0.8754	0.6736	0.3038
Hispanic Male	0.9151	0.8729	0.8404	0.8699	0.6692	0.3649
Muslim Female	0.9015	0.8348	0.9230	0.8817	0.5686	0.3409
Muslim Male	0.8574	0.9043	0.9407	0.9421	0.6890	0.4122
White Female	0.9619	0.8900	0.9555	0.9714	0.7862	0.4503
White Male	0.9541	0.8930	0.9504	0.9589	0.7234	0.6388
OOV Name	0.7405	0.8962	0.8720	0.8374	0.1003	0.0381

- White male names have the highest accuracy across models while black female names have the lowest
- For ELMo model muslim female names have the lowest accuracy, while white female names have the highest accuracy

Mishra, S., He, S., & Belli, L. (2020). Assessing Demographic Bias in Named Entity Recognition. *ArXiv, abs/2008.03415*.

Thank You

More details:

- https://socialmediaie.github.io/tutorials/
- https://socialmediaie.github.io/
- Contact: https://twitter.com/TheShubhanshu