EXPERIMENTS -->

1. that your new system is better than local systems; proving your global approach helps.
2. choose a number of (recognized) NER systems and show that you beat them.
3. experiment on stream vs. non-stream setting
4. testing with unified classifier vs separate classifier for syntax and semantic features
5. show that all the components of your system are needed: you will need to remove each and show accuracy degradation.
6. need to discuss the effort of tuning the system, i.e., labeling and training.

Model Used: Finetuned BERTweet (RoBERTa model) for NER

Unified--> Classifier I (syntactic + contextual embeddings, average pooling)

Separate--> Classifier II (only contextual embeddings, average pooling)

Single--> ClassifierAlt (only contextual embeddings, learned pooling function)

Dataset Type	Dataset Name	Entity Embedding Size	Dataset Size	Entity Classifier Type	Once-and-done				Collective				Improvement		
					P	R	F1		P	R	F1				
Streaming	1K	768	1104	Unified (I)	0.65	0.47	0.55		0.79	0.61	0.69		25.45454545		
				Separate(II)											
		300		Unified(I)	0.66	0.49	0.56		0.81	0.65	0.72		28.57142857		15.53164878
				Separate(II)					0.8	0.64	0.71		26.78571429		20.65586873
				Single(Alt)				33.16	0.84	0.66	0.74	1.16	32.14285714		
Streaming	2K (covid)	768		Unified	0.56	0.51	0.53								
				Separate											
		300		Unified	0.56	0.51	0.53		0.65	0.6	0.62				
				Separate					0.65	0.6	0.62		16.98113208		
				Single					0.64	0.64	0.64		20.75471698		
Streaming	3K	768	3067	Unified	0.77	0.63	0.69		0.86	0.81	0.84		21.73913043		
				Separate											
		300		Unified	0.77	0.63	0.69		0.86	0.81	0.84		21.73913043		
				Separate					0.85	0.83	0.84				
				Single					0.84	0.82	0.83		20.28985507		
Streaming	6K														
	roevwade	768		Unified	0.66	0.42	0.51		0.81	0.69	0.74				
				Separate											
		300		Unified	0.68	0.41	0.52		0.89	0.73	0.8		53.84615385		
				Separate					0.87	0.73	0.79		51.92307692		
				Single					0.87	0.73	0.79		51.92307692		
												1.6			
	billdeblasio	768		Unified	0.5	0.71	0.59		0.85	0.79	0.82				
				Separate											
		300		Unified	0.5	0.72	0.59		0.87	0.82	0.84		42.37288136		
				Separate					0.78	0.8	0.79		33.89830508		
				Single					0.79	0.8	0.8		35.59322034		
												1.8			
	pikapika	768		Unified	0.69	0.58	0.64		0.72	0.67	0.69				
				Separate											
		300		Unified	0.67	0.59	0.63		0.8	0.7	0.75		19.04761905		
				Separate											
				Single					0.75	0.76	0.76	2.3	20.63492063		
	ripcity	768		Unified	0.85	0.53	0.65		0.91	0.63	0.74				
				Separate											
		300		Unified	0.81	0.52	0.64		0.75	0.7	0.72		12.5		
				Separate											
				Single					0.74	0.71	0.72	1.48	12.5		
	billnye	768		Unified	0.78	0.45	0.57		0.81	0.49	0.61		7.01754386		
				Separate											
		300		Unified	0.78	0.45	0.57		0.81	0.49	0.61		7.01754386		
				Separate											
				Single	0.83	0.7	0.76		0.91	0.78	0.84	1.6	10.52631579		
	WHOLE	768		Unified	0.696	0.538	0.6068849271		0.814	0.7166666667	0.762238676		25.59855122		
				Separate											
		300		Unified	0.688	0.538	0.59		0.824	0.688	0.744		26.10169492		
				Separate											
				Single	0.698	0.588	0.628	210	0.812	0.756	0.782	8.78	24.52229299	24.82228103	
Non - Streaming	WNUT	768	1287	Unified	0.61	0.43	0.51		0.74	0.47	0.58		13.7254902		
				Separate											
		300		Unified	0.61	0.43	0.51		0.73	0.46	0.57				
				Separate					0.72	0.45	0.56		9.803921569		
				Single					0.73	0.52	0.60736		19.09019608		
Non - Streaming	BTC	768	9551 (-200)	Unified	0.64	0.52	0.57								
				Separate											
		300		Unified	0.63	0.5	0.56		0.55	0.61	0.58		3.571428571		
				Single					0.69	0.58	0.63		12.5		
									0.7	0.56	0.62		10.71428571	21.51560098	
														14.9022409	
						0.5545									
														12.21988796	

Model Used: Finetuned Gaguilar et al. for NER

** NB: Gaguilar input files are sentence level, conversion done during pre-processing prior to extracting POS Tags

Dataset Type	Dataset Name	Entity Embedding Size	Dataset Size	Entity Classifier Type	Once-and-done				Collective				Improvement	
					P	R	F1		P	R	F1			
Streaming	1K	100	1000		0.75	0.57	0.65		0.77	0.62	0.69		6.153846154	
					0.76	0.58	0.66		0.79	0.64	0.71		7.575757576	
					0.76	0.55	0.64		0.87	0.66	0.7505882353		17.27941176	
Streaming	2K (covid)	100	2000		0.63	0.57	0.6		0.69	0.67	0.68		13.33333333	
Streaming	3K	100	3000		0.83	0.42	0.56		0.78	0.66	0.72			
					0.77	0.64	0.7		0.82	0.77	0.794		13.42857143	
Streaming	6K		6000											
	roevwade	100			0.89	0.62	0.73		0.9	0.77	0.83		13.69863014	
					0.87	0.62	0.72	-	0.89	0.82	0.85	0.98	18.05555556	
	billdeblasio	100			0.83	0.73	0.78	-	0.91	0.81	0.86	1.02	10.25641026	
	pikapika	100			0.83	0.57	0.67	-	0.83	0.67	0.74	1.4	10.44776119	
	ripcity	100			0.86	0.67	0.75	-	0.86	0.71	0.78	1.01	4	
	billnye	100			0.72	0.46	0.56	-	0.88	0.74	0.8	1.3	42.85714286	
	WHOLE				0.822	0.61	0.696		0.874	0.75	0.806		15.8045977	15.35815522
							0.69		0.88	0.75	0.81		17.39130435	11.91775629
Non - Streaming	WNUT	100	1287		0.68	0.47	0.56		0.72	0.48	0.58		3.571428571	
					0.68	0.47	0.56		0.72	0.5	0.5901639344	1.5	5.386416862	
Non - Streaming	BTC	100			0.75	0.56	0.64		0.77	0.59	0.67	10.2	4.6875	5.036958431

Model Used: Finetuned NP Chunker for NER

Dataset Type	Dataset Name	Entity Embedding Size	Dataset Size	Entity Classifier Type	Once-and-done			Collective			Time	Improvement			
					P	R	F1	P	R	F1				F1 Gain	
														77.50%	
	1K				0.3	0.58	0.4	0.81	0.63	0.71	1.2	77.5	36.6990374	36.40%	
														17.30%	25.61%
													46.63509456	32.10%	
	2K				0.4	0.47	0.43	0.59	0.62	0.6	2.09	39.53488372		39.50%	
														51.20%	
														13.30%	
	3K				0.59	0.54	0.56	0.71	0.66	0.68	3.04	21.42857143		20.80%	
														21.40%	
														13.04%	
														13.60%	
														20.30%	
	6K				0.47	0.59	0.52	0.83	0.73	0.77		48.07692308		48.10%	
														42.30%	
														17.40%	
	roevwade							0.92	0.6	0.72	1.92			26.10%	
	billdeblasio							0.88	0.82	0.85	1.73			12.80%	
	pikapika													48.70%	
	ripcity													5.40%	
	billnye													13.70%	
	WHOLE													20.80%	
														5.70%	
														4.70%	
														12.50%	
	WNUT				0.42	0.35	0.39	0.63	0.35	0.44		12.82051282			
													16.82692308		
	BTC				0.46	0.51	0.48	0.66	0.52	0.58	14.54	20.83333333			

Model Used: Finetuned TwitterNLP for NER

Dataset Type	Dataset Name	Entity Embedding Size	Dataset Size	Entity Classifier Type	Once-and-done			Collective			Time	Improvement	
					P	R	F1	P	R	F1			
	1K				0.65	0.47	0.55	0.8	0.66	0.72	1.4	30.90909091	31.06916747
													34.36994342
	2K				0.33	0.52	0.41	0.71	0.55	0.62		51.2195122	
	3K				0.75	0.64	0.69	0.88	0.71	0.78	3.19	13.04347826	
	6K				0.67	0.41	0.52	0.89	0.64	0.74		42.30769231	
	roevwade							0.93	0.77	0.84	1.36		
	billdeblasio							0.9	0.76	0.82	1.62		
	pikapika												

ripcity									
billnye									
WHOLE									
WNUT	0.35	0.42	0.39		0.65	0.49	0.5587719298	**43.2748538**	
									24.46761558
BTC	0.69	0.43	0.53	77.15	0.74	0.45	0.56	**5.660377358**	

Time Overhead

	D1		D2		D3		D4		WNUT		BTC		
NP Chunker	1.2	1.195219124	2.09	1.690664941	2.6	1.483171706	5.4	0.979	2.34	1.930374526	14.04	2.2357	**1.58577888**
TwitterNLP	0.96	13.57850071	1.66	8.778424114	2.9	19.20529801	5.82	16.51	2.47	17.33333333	10.65	13.804	**14.86919284**
Aguilar	1.27	1.017628205	1.7	0.5743243243	3.14	1.052984574	5.98	1.069	1.72	0.7500436072	10.3	1.5030	**0.9946315005**
BERTweet	1.16	3.498190591	2.35	5.841411882	3.58	6.109215017	6.78	2.938	1.75	7.172131148	10.69	5.5159	**5.179198227**

Comparison with Document EMD (Global EMD Baselines)

Dataset	Doc-NER-System	P	R	F1	Best EMD Globalizer F1			
1K	HIRE-NER	0.65	0.62	0.6346456693				
	DOCL-NER	0.61	0.64	0.62464	0.74			
2K	HIRE-NER	0.46	0.56	0.51				
	DOCL-NER	0.52	0.62	0.56	0.68			
3K	HIRE-NER	0.75	0.73	0.74				
	DOCL-NER	0.74	0.72	0.73	0.83			
6K								
roevwade		0.74	0.75	0.75				
billdeblasio	HIRE-NER	0.72	0.74	0.73				
pikapika		0.26	0.61	0.36	0.81			
ripcity		0.73	0.77	0.75		0.9	0.84	0.87
billnye		0.43	0.52	0.47		0.85	0.79	0.82
WHOLE		0.576	0.678	0.612	0.6228516746	0.76	0.82	0.79
						0.77	0.71	0.74
						0.91	0.78	0.84
roevwade		0.74	0.85	0.79				
billdeblasio	DOCL-NER	0.55	0.83	0.66	0.81			
pikapika		0.22	0.39	0.29				
ripcity		0.77	0.83	0.8				
billnye		0.37	0.49	0.42				
WHOLE		0.53	0.678	0.592	0.5949337748			
WNUT	HIRE-NER	0.5	0.5	0.5				
	DOCL-NER	0.49	0.47	0.48	0.58			
BTC	HIRE-NER	0.6	0.49	0.54				
	DOCL-NER	0.56	0.48	0.52	0.67			

Other Experiments

BERTweet			Phase1NotInPhase2		Mentions	Entities missed	Totally Missed Mentions	Annotated Entities		mislabeled candidates	mentions missed by mislabeling
	1K		1/1297	1	1297	143	393	274		9	24
	2K		4/761	4	761	195	321	467		13	68
	3K		86/3324	86	3324	367	484	889		30	195
	6K	roevwade	1/1824	1	1824	61	596	107		4	6
		billdeblasio	13/1198	13	1198	65	214	129		8	42
		pikapika	8/1093	8	1093	60	296	143		6	24
		ripcity	2/1118	2	1118	63	317	173		5	8
		billnye	14/797	14	797	64	387	124		6	102
				129	11412	1018	3008	2306	26.35821942	81	469
											4.109709078
						1.130389064		1288	0		
									0.06288819876		