UGSD: User Generated Sentiment Dictionaries from Online Customer Reviews

Chun-Hsiang Wang
National Chengchi University

Joint work with Kang-Chun Fan, Prof. Chuan-Ju Wang,
Prof. Ming-Feng Tsai

January 27, 2019
Honolulu, Hawaii, USA

Eiffel Tower

User-generated Reviews

Eiffel Tower

●●●●●

Eiffel Tower is an amazing place to spend at Paris. A must see through out the day …

●●●●●

Romantic Eiffel Tower. Well worth paying the extra to get to the top for …

●○○○○

The Eiffel Tower is an overrated land mark and was overpopulated with tourists …

●●○○○

Very disappointing. Lines were crazy, people trying to get you to buy …

User-generated Reviews

Eiffel Tower

●●●●●

Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

●●●●●

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

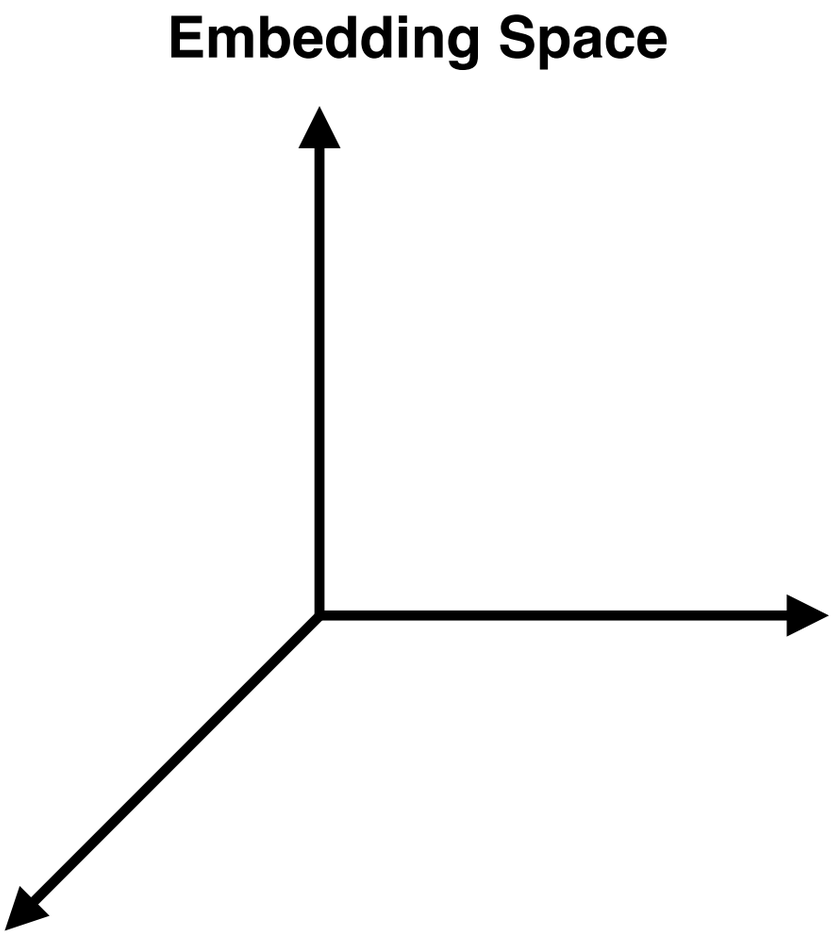
●○○○○

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

●●○○○

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Embedding Space



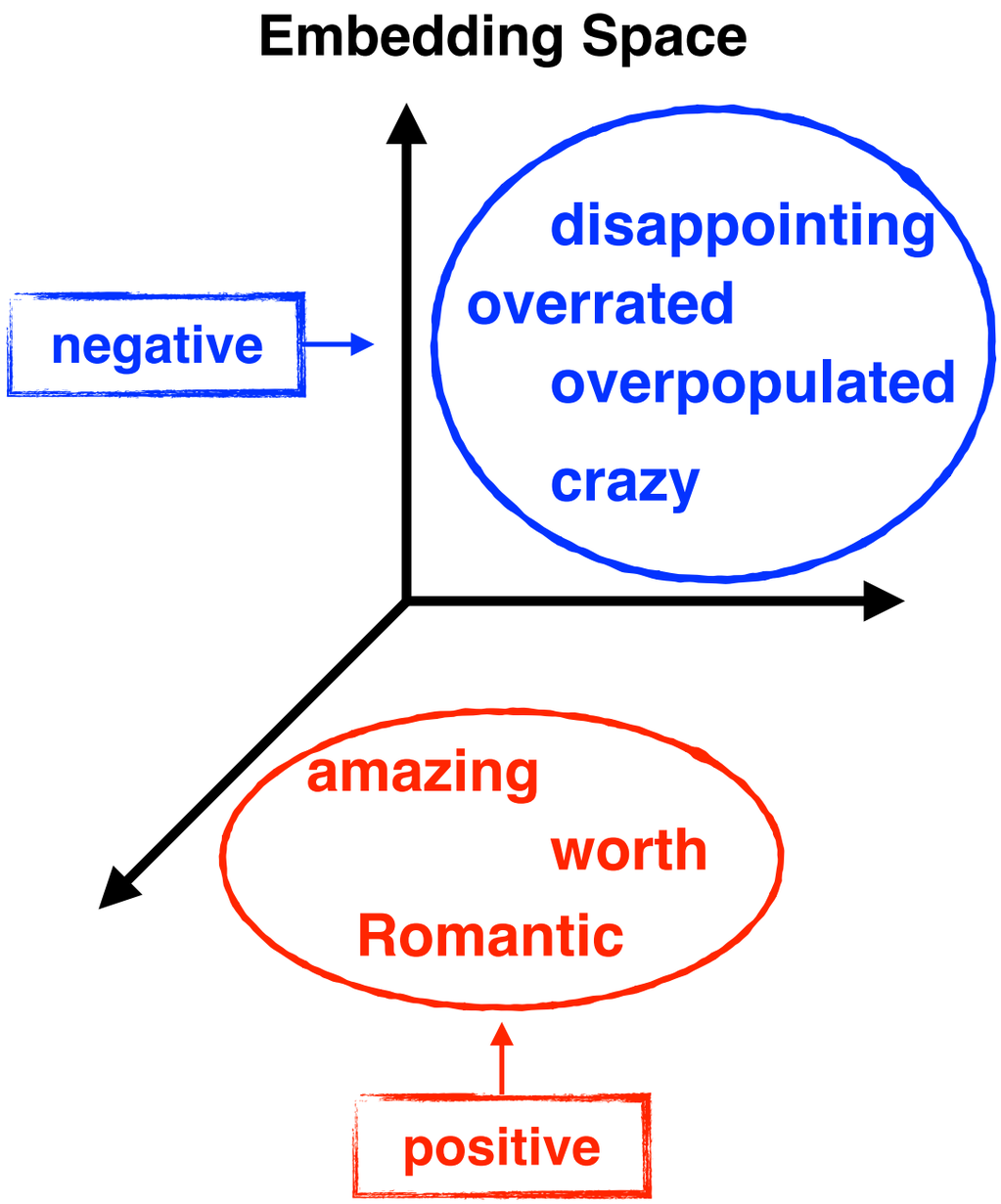
Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Embedding Space



Eiffel Tower is an **amazing** place to spend at Paris. A must see through out the day …

Romantic Eiffel Tower. Well **worth** paying the extra to get to the top for …

The Eiffel Tower is an **overrated** land mark and was **overpopulated** with tourists …

Very **disappointing**. Lines were **crazy**, people trying to get you to buy …

Importance of Sentiment Lexicons

- Sentiment analysis and opinion mining
- Sentiment words are domain-specific

TripAdvisor

The hotels in this city are usually too **small** for the whole family to stay overnight.

Amazon

The cellphone is **small** and therefore convenient for people to use it with a single hand.

Related Work

- Calculate the scores of new words via the proximity to one or more seed words (Chetviorkin, et al., 2014; Qiu, et al., 2009)
- Apply a general-purpose sentiment lexicon, a synonym-antonym dictionary, and linguistic heuristics (Lu, et al., 2011)
- Build semantic representations and propagate the polarity score of each word from a seed set with random walks (Hamilton, et al., 2016)

Our Framework: UGSD

- Construct sentiment lexicons from user-generated reviews
- Features:
  1. Data-driven: require no seed words or external lexicons
  2. Domain-specific: construct domain-specific sentiment lexicons with reviews from different domains
  3. Application scalability: produce representations of the learned sentiment words

Problem Definition

Eiffel Tower

●●●●●
Eiffel Tower is an amazing place to …

●●●●●
Romantic Eiffel Tower. Well worth …

●○○○○
The Eiffel Tower is an overrated land …

●●○○○
Very disappointing. Lines were crazy …

A set of reviews of a certain domain $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$

A rating $r \in \mathcal{R}$ corresponds to each of reviews

A set of entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$

Generate a set of words $\mathcal{L}_r$ corresponding to the rating $r \in \mathcal{R}$

Candidate Word Selection

- Extract adjectives and adverbs as candidates $\mathcal{S} = \{s_1, s_2, \ldots, s_G\}$
- Combine consecutive adverbs and adjectives

Eiffel Tower

●●●●●
Eiffel Towel is an **amazing** place to spend at Paris. A must see through out the day …

●●●●●
Romantic Eiffel Towel. **Well_worth** paying the extra to get to the top for …

●○○○○
The Eiffel Towel is an **overrated** land mark and was **overpopulated** with tourists …

●●○○○
Very_disappointing. Lines were **crazy**, people trying to get you to buy …

Entity Substitution

- Replace entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ with the rating $r \in \mathcal{R}$

Eiffel Tower

Eiffel Tower is an amazing place to spend at Paris. A must see through out the day …

Romantic **Eiffel Tower** Well_worth paying the extra to get to the top for …

The **Eiffel Tower** is an overrated land mark and was overpopulated with tourists …

Very_disappointing. Lines were crazy, people trying to get you to buy …

Entity Substitution

- Replace entities $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ with the rating $r \in \mathcal{R}$

Eiffel Tower

★★★★★ is an amazing place to spend at Paris. A must see through out the day …

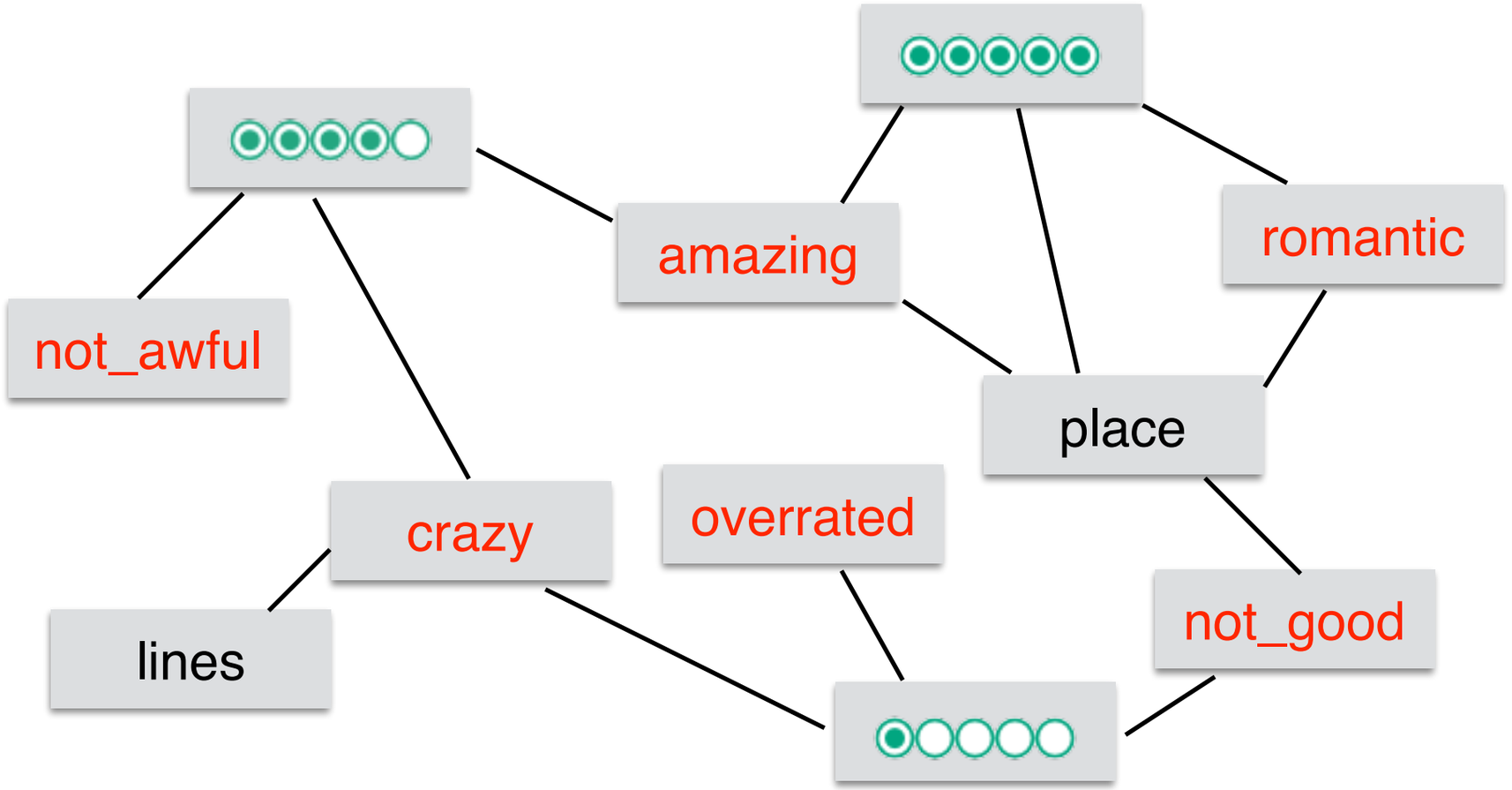
★★★★★ Romantic ★★★★★ . Well_worth paying the extra to get to the top for …

★☆☆☆☆ The ★☆☆☆☆ is an overrated land mark and was overpopulated with tourists …

★★☆☆☆ Very_disappointing. Lines were crazy, people trying to get you to buy …

Co-occurrence Proximity Learning

- Construct a *k* co-occurrence network with a predefined window size *k*



Dictionary Construction

- Calculate cosine similarity to build the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1|\mathcal{R}|} \\ a_{21} & a_{22} & \cdots & a_{2|\mathcal{R}|} \\ \vdots & \vdots & \ddots & \vdots \\ a_{|\mathcal{S}|1} & a_{|\mathcal{S}|2} & \cdots & a_{|\mathcal{S}||\mathcal{R}|} \end{pmatrix} \quad \text{where} \quad a_{ij} = \cos(\vec{v}_{s_i}, \vec{v}_{r_j})$$

- Define an element-wise function $\mathcal{G}(\cdot)$

$$\mathcal{G}(A) = (b_{ij}) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|} \quad \text{where} \quad b_{ij} \in \{0, 1\}$$

$$\mathcal{L}_{r_j} = \left\{ (s_i, \vec{v}_{s_i}, \theta_{s_i}^{r_j} = a_{ij}) \,|\, b_{ij} = 1 \right\}$$

Dictionary Construction

- Maximum-cosine-similarity scheme

$$m_i = \max_{1 \le j \le |\mathcal{R}|} a_{ij}$$

$$\mathcal{G}_{\max}(A) = (b_{ij}) = \left(1_{\{a_{ij} \ge m_i\}}\right) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$$

- Z-score scheme

$$z_{ij} = (a_{ij} - \mu_j)/\sigma_j$$

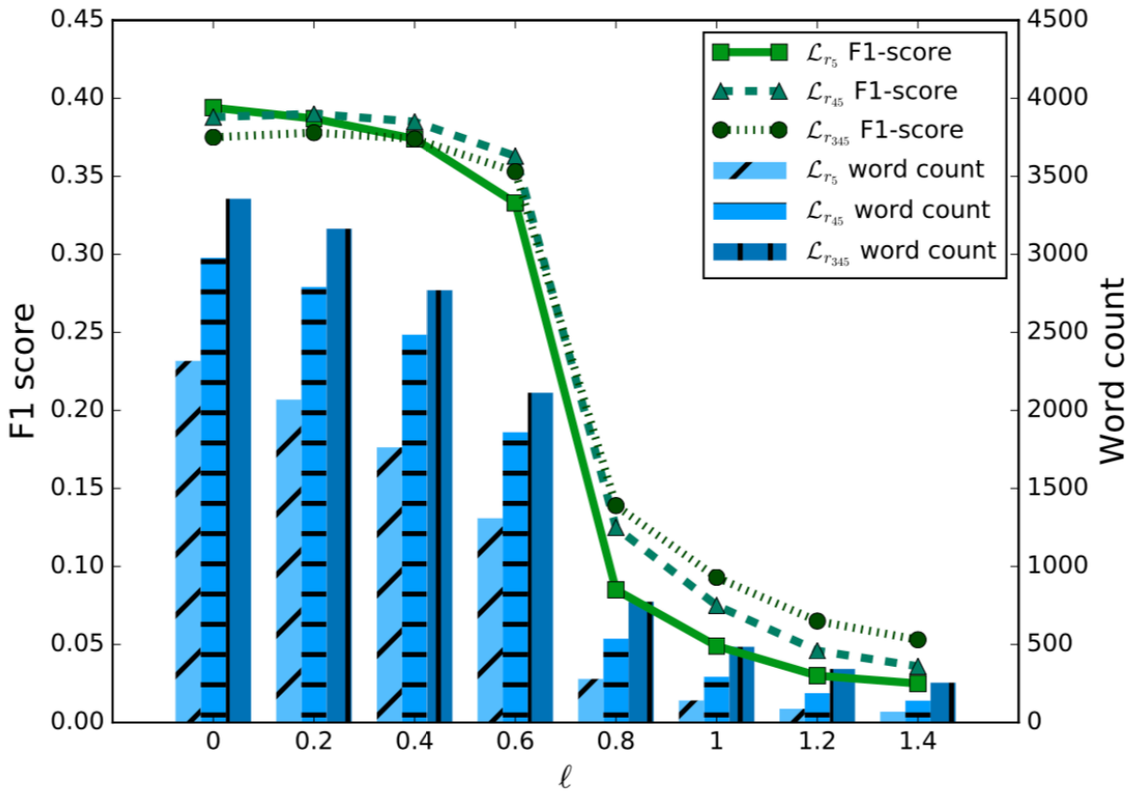
$$\mathcal{G}_{z>\ell}(A) = (b_{ij}) = \left(1_{\{z_{ij} > \ell\}}\right) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$$

Real-World Datasets

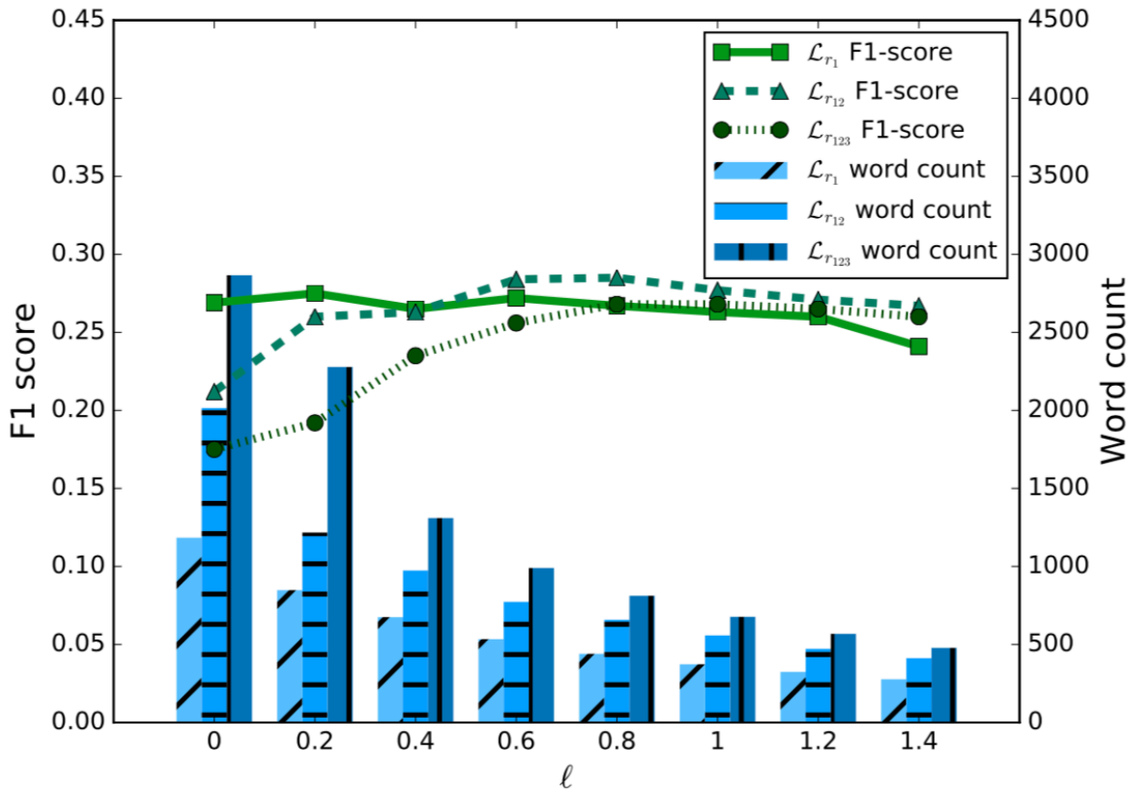
- Yelp:
  - Round 9 of Yelp dataset challenge
- TripAdvisor dataset:
  - Top 25 cities in 2016 and top 20 attractions or tours of each city
- Amazon dataset: (Wang, et al., 2010)
  - 6 categories of electronic supplies and top 20 products of each category

Comparison with Yelp Dictionary

- Compare Yelp dictionaries with the state-of-the-art Yelp dictionaries (Reschke, et al., 2013)



(a) Positive dictionary



(b) Negative dictionary

Comparison with Yelp Dictionary

- Compare Yelp dictionaries with the state-of-the-art Yelp dictionaries (Reschke, et al., 2013)

		Positive				Negative			
		# word	P	R	F1	# word	P	R	F1
NLTK		2,006	0.196	0.275	0.229	4,783	0.072	**0.607**	0.129
MPQA		2,304	0.198	0.318	0.244	4,152	0.079	0.579	0.139
SentiWordNet		14,712	0.039	0.395	0.071	10,751	0.015	0.288	0.029
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5}$	594	**0.352**	0.146	0.206	$\mathcal{L}_{r_1}$	1,112	0.161	0.314	0.213
	$\mathcal{L}_{r_{45}}$	1,125	0.332	0.260	0.292	$\mathcal{L}_{r_{12}}$	1,901	0.140	0.467	0.215
	$\mathcal{L}_{r_{345}}$	1,685	0.315	0.369	0.340	$\mathcal{L}_{r_{123}}$	2,461	0.119	0.512	0.193
$\mathcal{G}_{z>0.6}(\cdot)$	$\mathcal{L}_{r_5}$	1,309	0.349	0.318	0.333	$\mathcal{L}_{r_1}$	534	**0.281**	0.263	0.272
	$\mathcal{L}_{r_{45}}$	1,860	0.322	0.417	**0.363**	$\mathcal{L}_{r_{12}}$	773	0.247	0.335	**0.284**
	$\mathcal{L}_{r_{345}}$	2,113	0.296	**0.436**	0.353	$\mathcal{L}_{r_{123}}$	990	0.202	0.351	0.256

Sentiment Classification

- Conduct binary sentiment classification on reviews for three datasets

		Yelp			TripAdvisor			Amazon		
		# word	F1	Acc	# word	F1	Acc	# word	F1	Acc
	NLTK	6,787	0.762	0.697	6,787	**0.759**	0.699	6,787	0.766	0.707
	MPQA	6,450	0.708	0.601	6,450	0.701	0.608	6,450	0.716	0.616
	SentiWordNet	24,123	0.675	0.534	24,123	0.670	0.520	24,123	0.685	0.551
	Stanford Yelp	2,005	0.682	0.534	2,005	0.686	0.544	2,005	0.679	0.530
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5} \cup \mathcal{L}_{r_1}$	1,524	0.733	0.755	1,888	0.664	0.679	717	0.744	0.727
	$\mathcal{L}_{r_{45}} \cup \mathcal{L}_{r_{12}}$	2,692	0.771	**0.777**	3,428	0.746	**0.753**	1,566	0.763	0.755
$\mathcal{G}_{z>1.2}(\cdot)$	$\mathcal{L}_{r_5} \cup \mathcal{L}_{r_1}$	364	0.784	0.758	710	0.726	0.630	189	**0.801**	**0.782**
	$\mathcal{L}_{r_{45}} \cup \mathcal{L}_{r_{12}}$	451	**0.792**	0.762	1,060	0.736	0.650	346	0.800	0.772

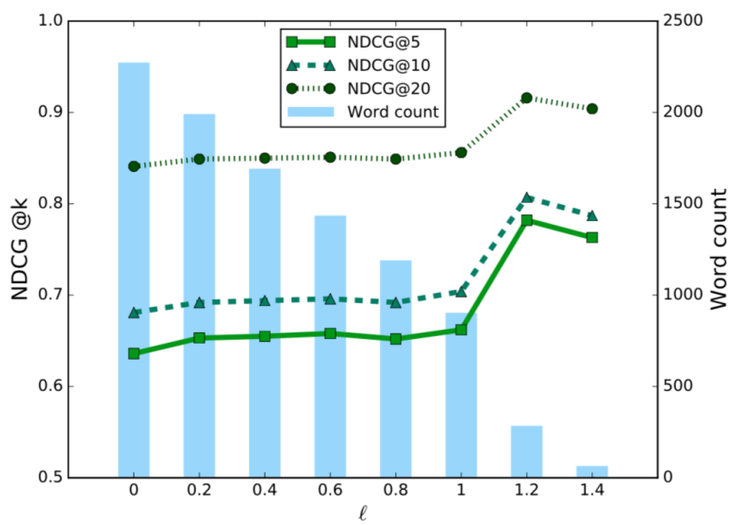
Entity Ranking

- Conduct entity ranking by the scoring function: (Chao, et al., 2017)

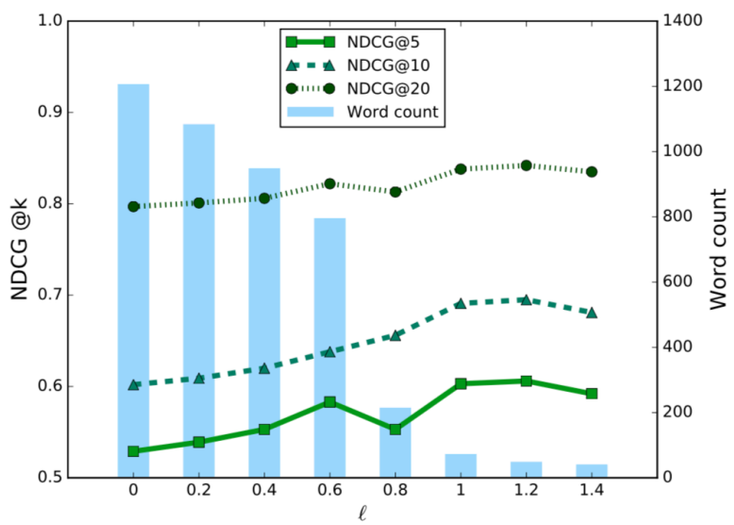
$$Score(y_i) = \sum_{j=1}^{n} f_{s_j} \cdot \cos(\vec{v}_{y_i}, \vec{v}_{s_j}) 1_{\{cos(\vec{v}_{y_i}, \vec{v}_{s_j})>0\}}$$

strength of sentiment words on entities

- Measure the ranking performance with normalized discounted cumulative gain (NDCG)



(a) TripAdvisor



(b) Amazon

Entity Ranking

- Entity ranking performance

		TripAdvisor			Amazon		
		# word	NDCG@5	NDCG@10	# word	NGCG@5	NDCG@10
	Frequency	-	0.610	0.664	-	0.494	0.623
	NLTK	1,071	0.556	0.632	595	0.603	0.659
	MPQA	1,294	0.562	0.641	710	0.571	0.654
	SentiWordNet	4,522	0.442	0.530	2,207	0.543	0.574
$\mathcal{G}_{\max}(\cdot)$	$\mathcal{L}_{r_5}$	207	**0.794**	**0.818**	258	**0.635**	**0.712**
	$\mathcal{L}_{r_{45}}$	745	0.669	0.724	493	0.549	0.641
	$\mathcal{L}_{r_{345}}$	1,626	0.654	0.698	995	0.574	0.655
$\mathcal{G}_{z>1.2}(\cdot)$	$\mathcal{L}_{r_5}$	288	0.782	0.807	51	0.606	0.695
	$\mathcal{L}_{r_{45}}$	569	0.735	0.770	114	0.515	0.631
	$\mathcal{L}_{r_{345}}$	895	0.719	0.751	221	0.515	0.627

Amazon Lexicons

Top	$\mathcal{L}_{r_5}$	θ_s^{r5}	$\mathcal{L}_{r_4}$	θ_s^{r4}
1	wonderful wonderfully	0.599	not_perfect	0.695
2	fantastic fantastically	0.538	overall	0.600
3	awesome	0.536	standalone	0.525
4	amazing amazingly	0.532	nice nicely	0.503
5	really_great	0.526	good	0.469
6	great greatly	0.503	almost_perfect	0.449
7	lovely loving	0.428	lightest	0.312
8	excellent excellently excelent excellant	0.406	far_satisfied	0.290
9	best	0.369	little	0.284
10	absolutely_wonderful	0.347	starter	0.281
11	exellent	0.319	great greatly	0.265
12	happy	0.315	pretty_happy	0.257
13	really_loving	0.297	solid solidly	0.256
14	smart	0.290	graphically_intense	0.238
15	ever	0.271	not_primary	0.220
16	absolute absolutely absolutly	0.263	uncertain	0.219
17	totally_satisfied	0.258	not_expensive	0.199
18	bought	0.251	still_amazing	0.194
19	beatiful	0.242	darn darned	0.165
20	perfect perfectly	0.225	not_smart	0.163

Amazon Lexicons

Disappointed. The phone is **not new**, it is a used phone.

$\mathcal{L}_{r_3}$	θ_s^{r3}	$\mathcal{L}_{r_2}$	θ_s^{r2}	$\mathcal{L}_{r_1}$	θ_s^{r1}
okay	0.813	unfortunate unfortunately	0.785	extremely_disappointed	0.769
ok	0.605	not_good	0.626	worthless	0.740
alright	0.583	disappointed disappointing	0.579	not_new	0.631
not_bad	0.521	not_waterproof	0.542	worse	0.609
dumb	0.517	really_disappointed really_disappointing	0.516	far_worst	0.594
not_great	0.418	unreliable	0.508	unacceptable	0.589
decent decently	0.399	dissapointed dissapointing	0.508	totally_useless	0.583
temporary	0.386	not_smart	0.480	useless	0.578
otherwise	0.375	overrated	0.458	faulty	0.576
pretty_decent	0.346	sad sadly	0.409	not_acceptable	0.568
not_smooth	0.297	not_happy not_happier	0.400	lemon	0.531
bland	0.290	unbearable	0.389	dissatisfied	0.527
not_happy not_happier	0.283	not_worst	0.386	not_happy not_happier	0.524
not_crazy	0.276	absolutely_terrible	0.370	apparent apparently	0.514
really_annoying	0.271	unhappy	0.367	defective	0.512
beloved	0.265	astonishing	0.359	miserable miserably	0.509
fully_capable	0.264	ongoing	0.351	unusable unused	0.488
really_excellent	0.248	still_slow	0.349	unhappy	0.487
wise	0.247	not_worth	0.342	ashamed	0.483
inaccurate	0.236	frustrated frustrating	0.339	completely_dead	0.472

Conclusions

- Propose a representation learning framework for constructing sentiment dictionaries from user reviews
  - Data-driven
  - Domain-specific
  - Application scalability
- Code & Datasets: github.com/cnclabs/UGSD