UGSD:

User Generated Sentiment Dictionaries from Online Customer Reviews

Chun-Hsiang Wang,[†] Kang-Chun Fan,[‡] Chuan-Ju Wang,[‡] Ming-Feng Tsai[†]
[†] National Chengchi University, Taiwan
[‡] Academia Sinica, Taiwan

CFDA-CLIP Labs

Source Code: UGSD

Motivation

- ✓ Constructing domain-specific sentiment dictionaries is important to sentiment analysis.
- ✓ Customer review platforms attain rich information about the ways that people convey their sentiment in certain domains.
- ✓ General sentiment dictionaries or annotated seed words greatly impact the results of dictionary construction.

Contributions

- ✓ Data-driven dictionary: Requiring no additional annotation of seed words or external dictionaries.
- ✓ Domain-specific dictionaries: Applying to a variety of user-generated content from different domains.
- ✓ Application scalability: Producing representations of the learned sentiment words during the dictionary construction.

Experiments

Table 4: Performance on sentiment classification

		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

Table 5: Performance on entity ranking

		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

Our framework: UGSD

Review transformation

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 …

- ✓ Leverage POS information
- ✓ Concatenate adverbs and adjectives
- ✓ Replace entities with ratings

Representation learning

- ✓ Joint probability between words

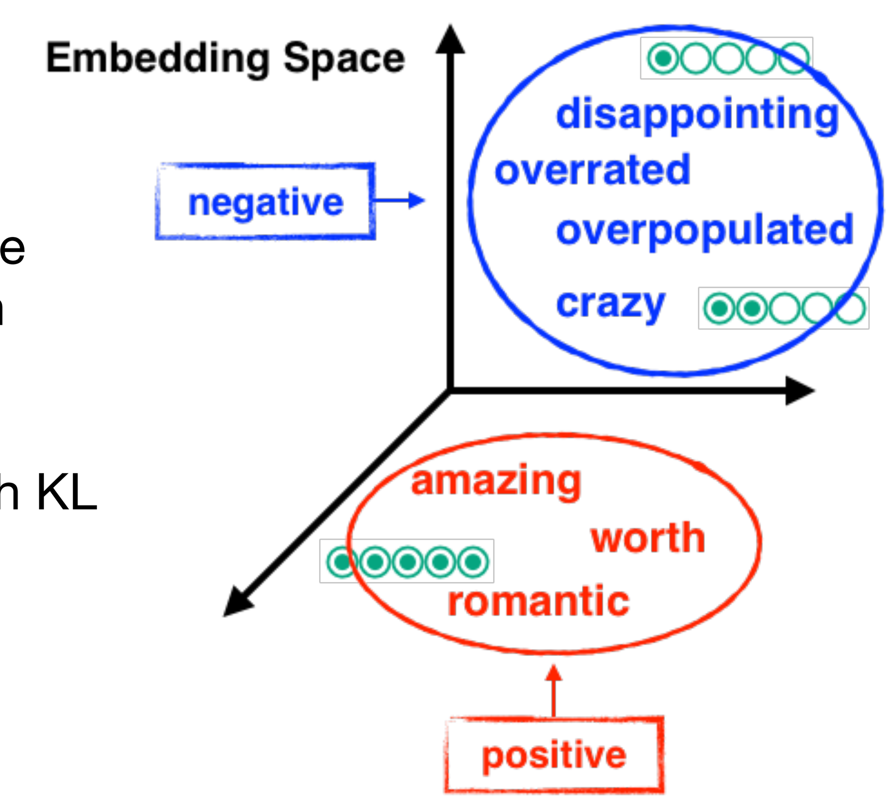
$$p(i,j) = \frac{1}{1+e^{-\vec{v_i}^\intercal \cdot \vec{v_j}}}$$

- ✓ Minimize the distance between the empirical and learned distribution

$$O = dist(\hat{p}(\cdot,\cdot), p(\cdot,\cdot))$$

- ✓ Replace the distance function with KL divergence

$$O = -\sum_{(i,j)\in\mathcal{A}} f_{ij} \log p(i,j)$$



Dictionary construction

$$\begin{array}{c} \\ \text{good} \\ \text{great} \\ \vdots \\ \text{bad} \end{array}
\begin{array}{c} r_5 \quad r_4 \quad \cdots \quad r_1 \\ \begin{pmatrix} 0.7 & 0.5 & \cdots & -0.6 \\ 0.6 & 0.4 & \cdots & -0.8 \\ \vdots & \vdots & & \vdots \\ -0.3 & -0.1 & \cdots & 0.6 \end{pmatrix} \end{array}
\begin{array}{cc} \mathcal{G}_{max} & \mathcal{G}_{z>1.0} \\ \mathcal{L}_{r_5} & \mathcal{L}_{r_5}\mathcal{L}_{r_4} \\ \mathcal{L}_{r_5} & \mathcal{L}_{r_5} \\ \vdots & \vdots \\ \mathcal{L}_{r_1} & \mathcal{L}_{r_1} \end{array}$$

$$a_{ij} = cos(\vec{v}_{s_i}, \vec{v}_{r_j})$$

- ✓ 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:

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

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

Amazon dictionaries

Top	$\mathcal{L}_{r_5}$	θ_s^{r5}	$\mathcal{L}_{r_4}$	θ_s^{r4}	$\mathcal{L}_{r_3}$	θ_s^{r3}	$\mathcal{L}_{r_2}$	θ_s^{r2}	$\mathcal{L}_{r_1}$	θ_s^{r1}
1	wonderful wonderfully	0.599	not_perfect	0.695	okay	0.813	unfortunate unfortunately	0.785	extremely_disappointed	0.769
2	fantastic fantastically	0.538	overall	0.600	ok	0.605	not_good	0.626	worthless	0.740
3	awesome	0.536	standalone	0.525	alright	0.583	disappointed disappointing	0.579	not_new	0.631
4	amazing amazingly	0.532	nice nicely	0.503	not_bad	0.521	not_waterproof	0.542	worse	0.609
5	really_great	0.526	good	0.469	dumb	0.517	really_disappointed really_disappointing	0.516	far_worst	0.594
6	great greatly	0.503	almost_perfect	0.449	not_great	0.418	unreliable	0.508	unacceptable	0.589
7	lovely loving	0.428	lightest	0.312	decent decently	0.399	dissapointed dissapointing	0.508	totally_useless	0.583
8	excellent excellently excelent excellant	0.406	far_satisfied	0.290	temporary	0.386	not_smart	0.480	useless	0.578
9	best	0.369	little	0.284	otherwise	0.375	overrated	0.458	faulty	0.576
10	absolutely_wonderful	0.347	starter	0.281	pretty_decent	0.346	sad sadly	0.409	not_acceptable	0.568
11	exellent	0.319	great greatly	0.265	not_smooth	0.297	not_happy not_happier	0.400	lemon	0.531
12	happy	0.315	pretty_happy	0.257	bland	0.290	unbearable	0.389	dissatisfied	0.527
13	really_loving	0.297	solid solidly	0.256	not_happy not_happier	0.283	not_worst	0.386	not_happy not_happier	0.524
14	smart	0.290	graphically_intense	0.238	not_crazy	0.276	absolutely_terrible	0.370	apparent apparently	0.514
15	ever	0.271	not_primary	0.220	really_annoying	0.271	unhappy	0.367	defective	0.512
16	absolute absolutely absolutly	0.263	uncertain	0.219	beloved	0.265	astonishing	0.359	miserable miserably	0.509
17	totally_satisfied	0.258	not_expensive	0.199	fully_capable	0.264	ongoing	0.351	unusable unused	0.488
18	bought	0.251	still_amazing	0.194	really_excellent	0.248	still_slow	0.349	unhappy	0.487
19	beatiful	0.242	darn darned	0.165	wise	0.247	not_worth	0.342	ashamed	0.483
20	perfect perfectly	0.225	not_smart	0.163	inaccurate	0.236	frustrated frustrating	0.339	completely_dead	0.472