Modeling Pollyanna Phenomena in Chinese Sentiment Analysis

Ting-Hao (Kenneth) Huang*, Ho-Cheng Yu**, Hsin-Hsi Chen**
*Carnegie Mellon University, **National Taiwan University

Introduction

- ***Pollyanna Phenomena*** *describes the human tendency to use positive words more frequently than negative words.*
- *This phenomena has been observed in* ***Chinese, English, Italian, German, Spanish, and even across 20 different languages.***
- *Three main contributions of this work:*

1. *Conduct the* ***first detailed survey of the Pollyanna phenomena in various modern Chinese corpora.***
2. *Through quantitative and qualitative analyses, we discover that* ***for the documents with relatively fewer positive words, the intra-polarity document similarity****, of either positive or negative opinion polarity,* ***significantly increases.***
3. *Propose a* ***partition strategy for sentiment classification*** *and improve performance significantly.*

Data Analysis

- *We tag all corpora by an extended version of the* ***NTU Sentiment Dictionary (NTUSD)***
- *The average positive word frequencies in these four corpora are* ***1.83 to 5 times*** *of those of negative words.*
- *We propose an indicator* ***Bias(d)*** *to measure the degree of word-level linguistic bias in a given document d:*

$$\boldsymbol{Bias(d)} = \frac{C_p(d) - C_n(d)}{C_p(d) + C_n(d)}$$

$$\begin{cases} C_p(d) = (\text{Number of positive words in } d) + 1 \\ C_n(d) = (\text{Number of negative words in } d) + 1 \end{cases}$$

- ***The positivity bias is demonstrated to be strong in all four corpora.*** *The last row of TABLE 1 shows the average biases of all four corpora.*
- ***What happens in the documents which are "not that positive"?***
  - ✓ *The majority of them are not always negative documents.*
  - ✓ *Maybe focus on some topics?*

Data Set

		CTB	RECI	iPeen	MOAT
Basic Information	Data Type	Generic Corpus	News Opinion Summary	Restaurant Review	Evaluation Data
	Data Instance Type	Document			Sentence
	#Instance (Inst.)	892	2,389	19,986	4,652
	Avg. Inst. Length (#Word)	532.25	60.36	331.84	16.06
Sentimental Information	Opinion Polarity Label	Untagged	POS, NEG	POS, NEG, NEU	POS, NEG, NEU
	% Pos. Instance	-	59.36%	44.16%	8.88%
	% Neg. Instance	-	40.64%	7.88%	11.11%
	Avg. Pos. WF	0.10	0.11	0.08	0.11
	Avg. Neg. WF	0.02	0.03	0.03	0.06
	Avg. Bias	0.66	0.47	0.48	0.15

TABLE 1: Statistics of the four corpora

Quantitative Analysis

- *Partition the data by a bias value into the* ***upper set*** *and the* ***lower set****.*
- *Measure the average degree of similarity of word use between documents of the* ***same and different opinion polarities*** *in the two split data portion respectively.*

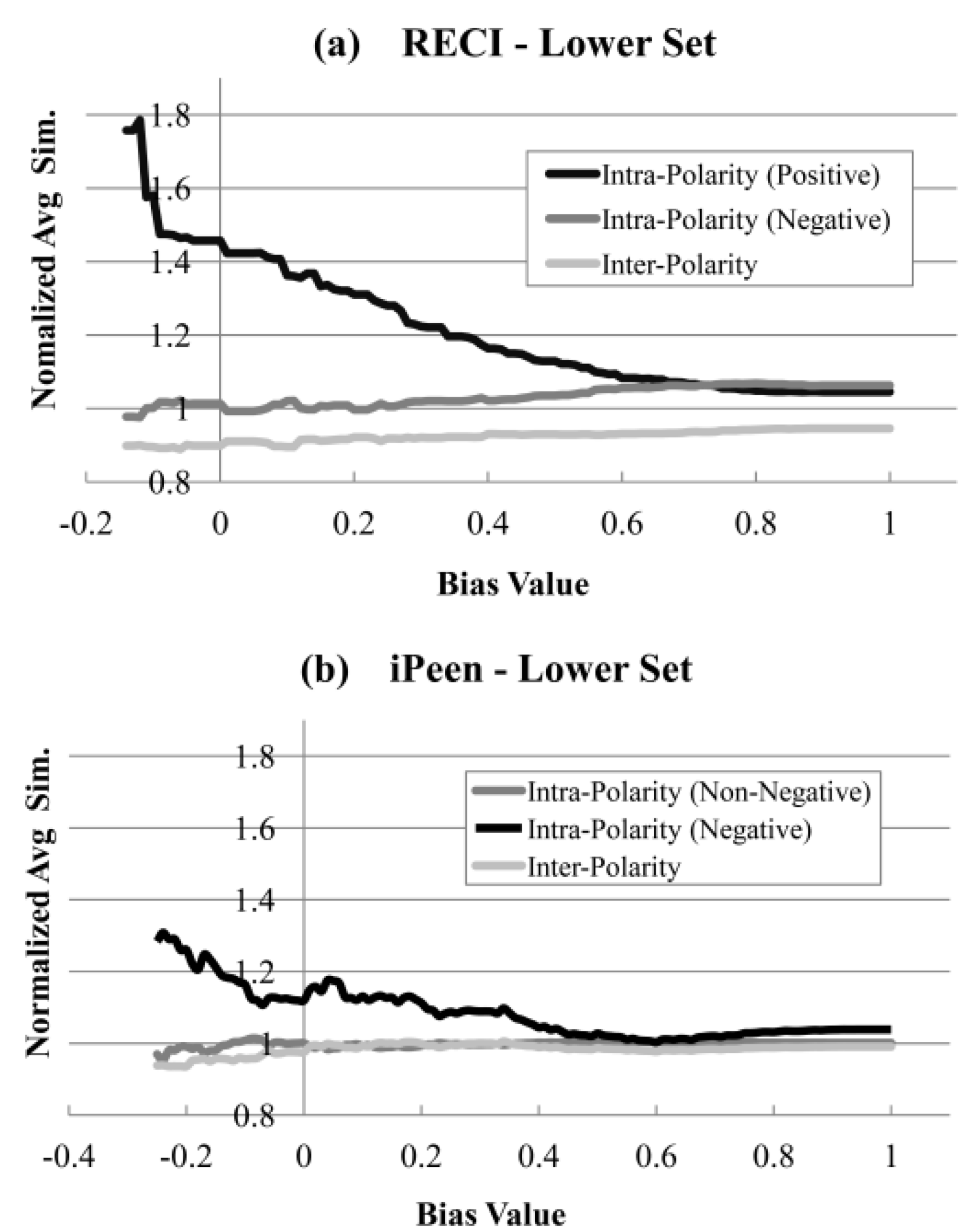

FIGURE 1: The curves of intra- and inter- polarity similarity of the lower sets in (a) RECI (target = Positive) and (b) iPeen (target = Negative).

- *Within the target opinion polarity,* ***the average cosine similarity among documents obviously rises up in the portion of data which has less positivity bias.***
- *In other words,* ***when we look at those outlier documents with lower bias values****, for certain target opinion polarity,* ***people actually tend to use more similar words with each other.***

Partitioning Strategy

Our analyses reveal the increase of intra-polarity similarity in certain part of data, and thus shed light on sentiment classification. We propose a strategy to partition the data sets by bias value, and train another model for the data portion which has lower positivity bias.

	Polarity	Strategy	Lower	Upper	Whole
(a) RECI	Pos.	Original	.646	.846	**.833**
		Partition	.692	.874	**.858****
	Neg.	Original	.789	.692	**.726**
		Partition	.869	.704	**.761****
	#Positive Docs.		191	1,227	1,418
	#Negative Docs.		348	623	971

	Polarity	Strategy	Lower	Upper	Whole
(b) iPeen	Neg.	Original	.333	.342	**.341**
		Partition	.342	.342	**.342****
	non-Neg.	Original	.811	.927	**.921**
		Partition	.870	.929	**.926****
	#Negative Docs.		260	1,314	1,574
	#Non-negative Docs.		932	17,480	18,412

TABLE 3: The F1-Score of Sentiment Classification in (a) RECI (b) iPeen

Qualitative Analysis

In ***iPeen****, negative comments cover a wide range of topics. But the negative comments mainly* ***focus on the "poor service"*** *in the lower set. In* ***RECI****, while the positive news in the upper set covers a wide range of topics, the positive news in the lower set mostly* ***focus on the indicators and rates which would be better if reduced.*** *The increase of word similarity actually reflect* ***the shrinkage of topics in the less positively biased portion of data.***