2012

Extracting Diverse Sentiment Expressions With Target-dependent Polarity from Twitter

Lu Chen
Wright State University - Main Campus

Wenbo Wang
Wright State University - Main Campus

Meenakshi Nagarajan
Wright State University - Main Campus

Shaojun Wang
Wright State University - Main Campus, shaojun.wang@wright.edu

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Extracting Diverse Sentiment Expressions With Target-dependent Polarity from Twitter

Lu Chen[1], Wenbo Wang[1], Meenakshi Nagarajan[2], Shaojun Wang[1], and Amit P. Sheth[1]
[1]Kno.e.sis Center, Wright State University, Dayton, OH USA
[2]IBM Almaden Research Center, San Jose, CA USA

1. Overview

This study focuses on ***automatic extraction of sentiment expressions associated with given targets from Twitter***.

One of the key challenges: Wide **diversity** and **informal** nature of sentiment expressions that cannot be trivially enumerated or captured using predefined lexical patterns

Contributions:

- Extracting a **diverse** and **richer** set of sentiment-bearing expressions, including formal and slang words/phrases, not limited to pre-specified syntactic patterns
- Assessing the **target-dependent** polarity of each sentiment expression
- A novel formulation of assigning polarity to a sentiment expression as a **constrained optimization problem** over the tweet corpus

Target-dependent Sentiment Expressions

positive expression
negative expression

Just saw **The Avengers**! It lived up to the hype.

Just got out of **the Avengers** movie. It was all storyboard, no story. What a clumsy, clunky mess. Got kind of bored 90 min in. B-

Went saw **The Avengers** tonight and man it was epic....a definite must see!!!

President Obama watched **the Avengers**.

The Avengers movie was bloody amazing! Mmm looking good Robert Downey Jr and Captain America ;)

If I can be half as cool as Scarlett Johansson in **The Avengers** when I grow up, my life will be accomplished.

Saw **the avengers** last night. Mad overrated. Cheesy lines and horrible writing. Very predictable.

My best friend wants to see **The Avengers** today, so of course I'll go. Now to decide which Batman shirt to wear...

Okay, **the Avengers** was worth the hype. Best superhero movie of all time. Blown away, here.

Distributions of N-grams and Part-of-speech of the Sentiment Expressions in the Gold Standard Data Set

N-gram	1	2	3	4	5	>5
Movie Domain	54.24%	21.02%	10.17%	6.44%	4.74%	3.39%
Person Domain	71.38%	17.75%	7.25%	1.81%	1.45%	0.36%

Part-of-speech	Adj.	Verb	Noun	Others
Movie Domain	57.63%	26.10%	13.22%	3.05%
Person Domain	45.29%	31.52%	21.02%	2.17%

2. Extracting Candidate Expressions

Root word: a word that is considered sentiment-bearing in general sense.

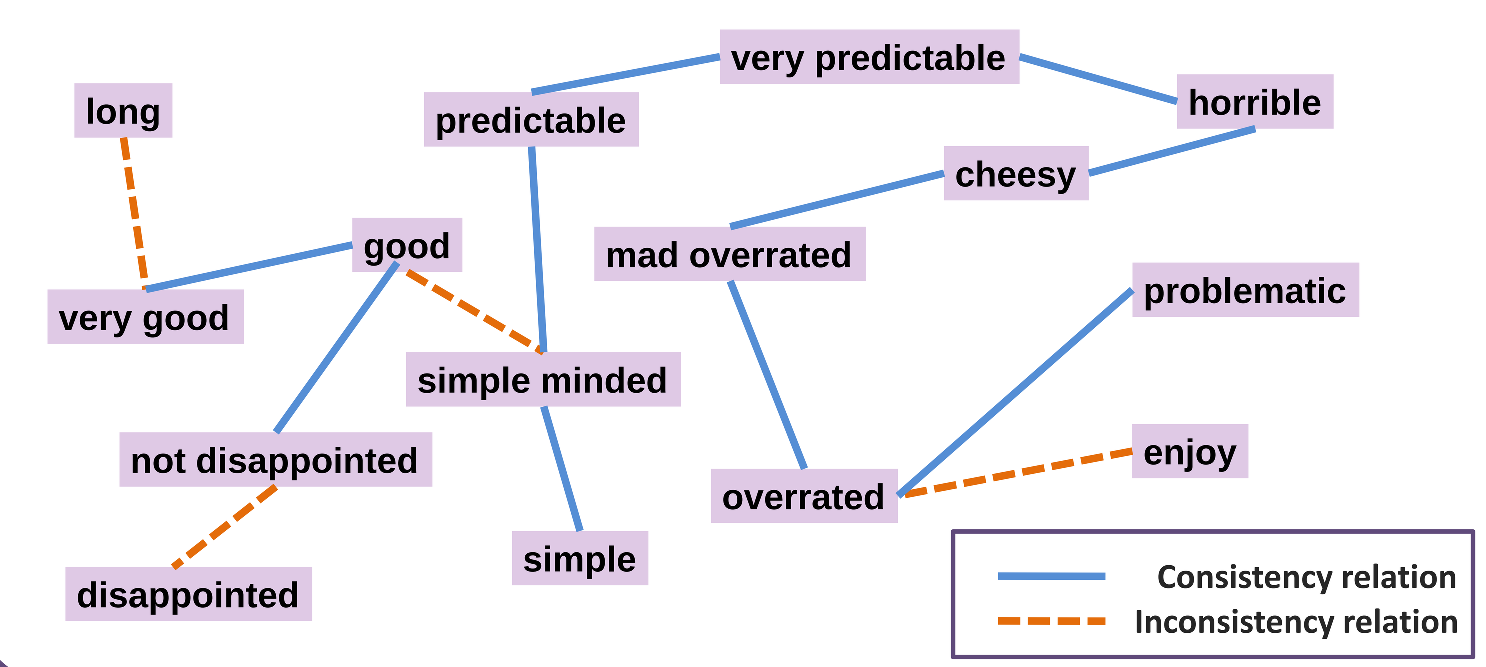
- Collecting root words from
  - MPQA, General Inquirer, and SentiWordNet (general-purpose sentiment lexicons)
  - Urban Dictionary (slang dictionary)
- For each tweet, selecting the "**on-target**" root words, and extracting all the n-grams that contain at least one selected root word as candidates

3. Identifying Inter-Expression Relations

Connecting the candidate expressions via two types of inter-expression relations – **consistency relation** and **inconsistency relation**

Example:

1. I saw The Avengers yesterday evening. It was long **but** it was very good!
2. I do enjoy The Avengers, **but** it's both overrated and problematic.
3. Saw the avengers last night. Mad overrated. Cheesy lines and horrible writing. Very predictable.
4. The avengers was good **but** the plot was just simple minded and predictable.
5. The Avengers was good. I was not disappointed.



4. An Optimization Model

For each candidate expression c_i,

- **P-Probability** $\Pr^P(c_i)$ – the probability that c_i indicates *positive* sentiment
- **N-Probability** $\Pr^N(c_i)$ – the probability that c_i indicates *negative* sentiment

$$\Pr^P(c_i)+\Pr^N(c_i)=1$$

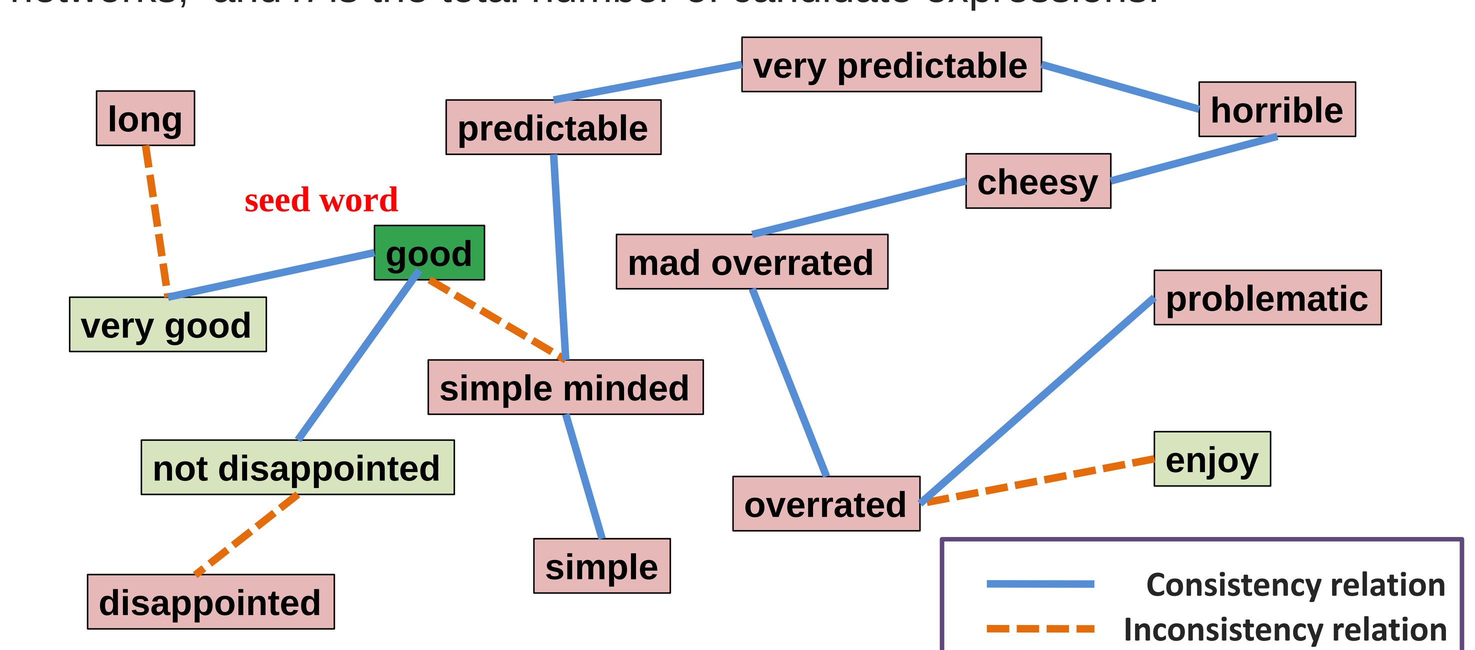
For each pair of candidate expressions c_i and c_j,

- **Consistency probability** – the probability that c_i and c_j have the same polarity: $\Pr^{cons}(c_i,c_j)=\Pr^P(c_i)\Pr^P(c_j)+\Pr^N(c_i)\Pr^N(c_j)$
- **Inconsistency probability** – the probability that c_i and c_j have different polarities: $\Pr^{incons}(c_i,c_j)=\Pr^P(c_i)\Pr^N(c_j)+\Pr^N(c_i)\Pr^P(c_j)$

Objective Function:

$$\text{minimize}\left\{\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(w_{ij}^{cons}\left(1-\Pr^{cons}(c_i,c_j)\right)^2+w_{ij}^{incons}\left(1-\Pr^{incons}(c_i,c_j)\right)^2\right)\right\}$$

where w_{ij}^{cons} and w_{ij}^{incons} are the weights of the edges (the frequency of the relations) between c_i and c_j in the **consistency** and **inconsistency** relation networks, and *n* is the total number of candidate expressions.



5. Experimental Setup

Datasets:

1) 168,005 tweets about ***movies***
2) 258,655 tweets about ***persons***

Gold standard:

1) 1,500 tweets labeled with sentiment expressions and overall polarities for the ***movie*** targets
2) 1,500 tweets labeled with sentiment expressions and overall polarities for the ***person*** targets

Baseline methods:

- **MPQA**, **GI**, **SWN**: For each extracted root word regarding the target, simply look up its polarity in MPQA, General Inquirer and SentiWordNet, respectively.
- **PROP**: A propagation approach proposed by Qiu et al.
- **COM-const**: Assign 0.5 to all the candidates as their initial P-Probabilities.
- **COM-gelex**: Initialize the candidates' polarities according to the root word set.

Reference: Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2009. Expanding domain sentiment lexicon through double propagation. In Proc. of IJCAI.

Quality of the Extracted Sentiment Expressions

Method	Precision	Recall	F-measure
Movie Domain			
MPQA	0.3542	0.5136	0.4193
GI	0.3318	0.4320	0.3753
SWN	0.2876	0.4898	0.3624
PROP	0.4742	0.5034	0.4884
COM-const	**0.6433**	0.5170	**0.5733**
COM-gelex	0.5164	**0.5578**	0.5363
Person Domain			
MPQA	0.3523	0.4746	0.4045
GI	0.2949	0.4058	0.3416
SWN	0.2161	0.3659	0.2718
PROP	0.5352	0.3696	0.4372
COM-const	**0.5879**	0.4710	**0.5230**
COM-gelex	0.4599	**0.5507**	0.5012

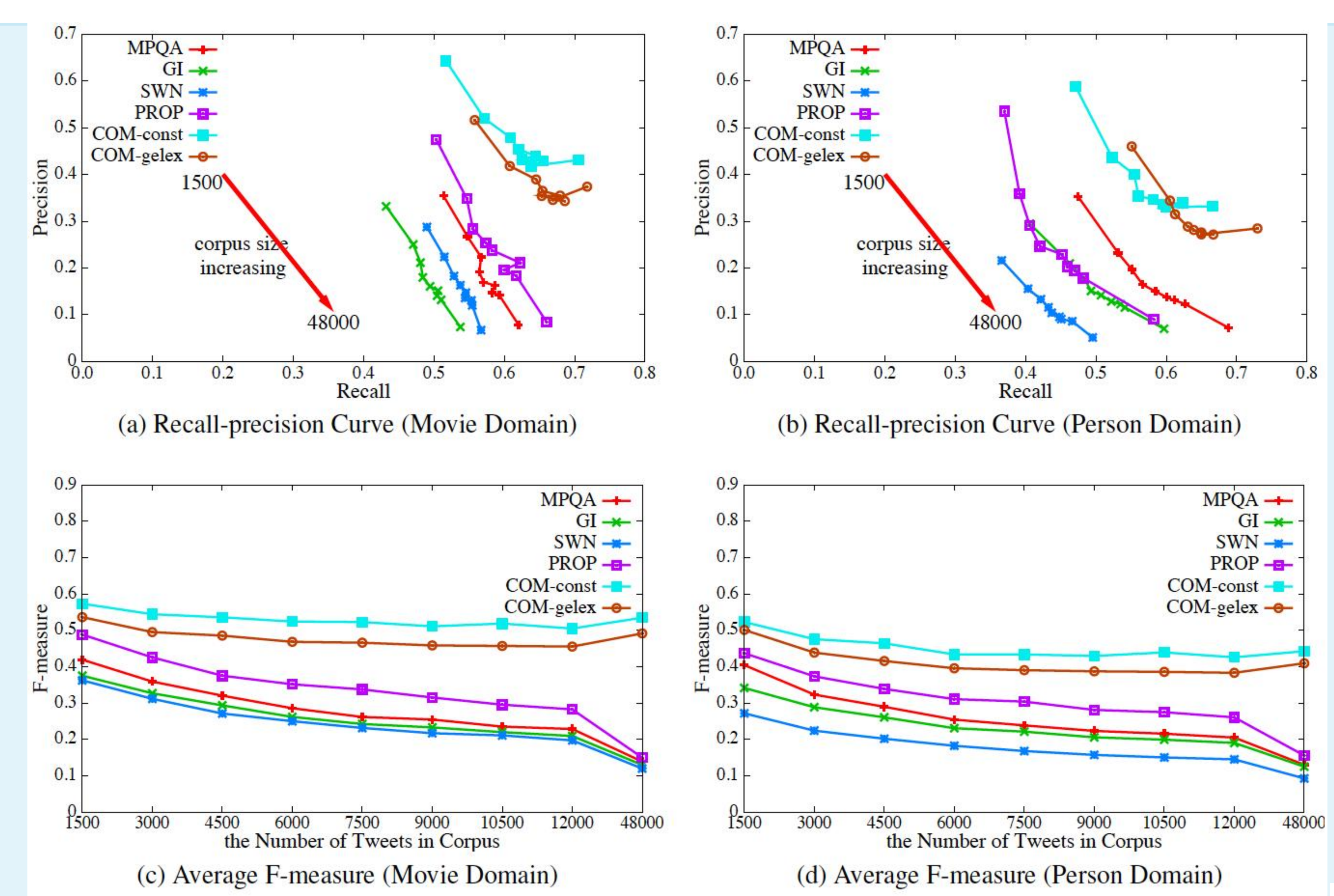


Figure 1: Results of Sentiment Expression Extraction with Various Corpora Sizes

Acknowledgement: This work is sponsored by NSF Award #IIS-1111182, "SoCS: Social Media Enhanced Organizational Sensemaking in Emergency Response"