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.

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. The Avengers was good but the plot was just simple minded and predictable.
4. The Avengers was good but 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

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:
  $$Pr_{c_i,c_j} = Pr^p(c_i)Pr^p(c_j) + Pr^n(c_i)Pr^n(c_j)$$

- Inconsistency probability – the probability that $C_i$ and $C_j$ have different polarities:
  $$Pr_{i,n} = Pr^p(c_i)Pr^n(c_j) + Pr^n(c_i)Pr^p(c_j)$$

Objective Function:

$$\text{minimize} \sum \sum (w^{long} (1 - Pr_{c_i,c_j}) + w^{very} (1 - Pr_{c_i,c_j}))$$

where $w^{long}$ and $w^{very}$ 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 movies targets
2) 1,500 tweets labeled with sentiment expressions and overall polarities for the persons 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: Assign 0.5 to all the candidates as their initial P Probabilities.
- COM-gexl: Initialize the candidates' polarities according to the top 500 word pairs.

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-gexl | 0.5164 | 0.5576 | 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-gexl | 0.4599 | 0.5507 | 0.5012

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