Informal sentiment analysis in multiple domains for English and Spanish

Tadej Štajner
Inna Novalija
Dunja Mladenić

Jožef Stefan Institute
Introduction - sentiment analysis

- Computational study of opinions, sentiment, evaluations, attitudes, views, emotions, subjectivity, etc. in text
- Also known as ‘opinion mining’
Motivation

- “Opinions” are important influencers of human behavior:
- To a large extent, our perception of reality is condition on how others see the world
- When we are making decisions, we often look for opinions of others
Domains

- Where can we find these opinions?
  - On the web, via word of mouth
    - Social media
    - Product, movie reviews
  - News
  - Internal data (customer feedback)

- Do different domains exhibit different properties?
Related work

- Early work focused on predicting movie review polarity as a text mining task [Pang & Lee, 2004]
  - Only positive vs. negative
- In some domains, separating subjective from objective is an important subproblem [Wiebe & Riloff, 2005]
Related work

- An interesting ground for testing various machine learning approaches, such as domain adaptation [Mejova & Srinivasan, 2012] or deep learning [Glorot et al, 2011].

- Integration of external and domain knowledge using sentiment lexicons
  - SentiWordNet [Esuli & Sebastiani, 2006]
  - SenticNet [Cambria et al., 2012]
Problem formulation

- General definition of opinion:
  - Opinion =
    (Holder, Target, Aspect, Orientation, Time)

- Some of these can be interesting sub-problems:
  - Holder, Target - named entity extraction
  - Aspect – target property extraction
  - Orientation – what is the strength and orientation of the opinion, if any (positive, negative, objective)?
This work focuses mainly on orientation, determining whether the opinion is positive, negative or objective.
Goals

- Do external sources of information increase performance?
- What is the best way to model this additional knowledge?
- Which lexicon resources work best?
- What are the differences across domains and languages?
Data description

- 5 datasets (2 Spanish, 3 English)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Language</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-ES</td>
<td>news</td>
<td>Spanish (translated from english)</td>
<td>1281 examples (pos, neg, obj)</td>
</tr>
<tr>
<td>Balahur et al. 2010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RenderES</td>
<td>social media</td>
<td>Spanish</td>
<td>891 examples (pos, neg, obj)</td>
</tr>
<tr>
<td>PangLee</td>
<td>reviews</td>
<td>English</td>
<td>2000 examples (pos, neg)</td>
</tr>
<tr>
<td>Pang and Lee, 2002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JRC-EN</td>
<td>news</td>
<td>English</td>
<td>1281 examples (pos, neg, obj)</td>
</tr>
<tr>
<td>Balahur et al. 2010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RenderEN</td>
<td>social media</td>
<td>English</td>
<td>134 examples (pos, neg)</td>
</tr>
</tbody>
</table>
Feature representation

Three main sources of knowledge:

- **Content**
  - counting preprocessed word tokens

- **Sentiment lexicons**
  - is there a global sentiment score assigned to a particular word?

- **Surface patterns**
  - How is the text phrased, written, expressed?
Content features

- Goal: bag of words representation
- Preprocessing steps:
  - Tokenization (preserving punctuation)
  - Target masking
  - Number masking
  - URL masking
  - Lower-casing
  - ASCII-normalization
  - Stopword filtering
  - Stemming
  - TF-IDF weighing
Lexicon features

- Sentiment lexicons have a numerical score attached to each word
- We calculate:
  - Sum of scores
  - Sum of absolute scores
  - Ratio of positive to negative words
  - + all of the above for every simplified part of speech – noun, verb, adjective, adverb
Lexicons

- Existing resources
  - SentiWordNet (en) [Esuli and Sebastian, 2006]
  - SenticNet (en) [Cambria et al., 2012]
  - UNTFull, UNTMedium (es) [Perez-Rosas et al. 2012]

- Novel resources: developed using a bootstrapping approach and a corpus of text
  - RenderLex (es, en)
  - RenderLexLinks (en)
    - Also contains the positive and negative link counts – the positive link count is the number of times a word co-occurs with a positive word, or is contrasted with a negative word.
Features (surface)

- count of fully capitalized words
- count of question-indicating words
- count of words that start with a capital letter
- count of repeated exclamation marks
- count of repeated same vowel
- count of repeated same character
- proportion of capital letters
- proportion of vowels
- count of negation words
- count of contrast words
- count of positive emoticons
- count of negative emoticons
- count of punctuation
- count of profanity words
Modeling hypotheses

- Given the different distribution properties of the BoW space, should we separate the model?

Concatenation model:

Two-layer words-features (W+F) model:
Experimental setup

- Varying feature representation:
  - Combinations of Surface, Lexicon, BoW

- Model combinations:
  - Two-layer [W+F-\*] vs concatenation
  - FeatureScaling [*Sc]
Results on JRC-ES

- SVM
- MNB
- WF-SVM
- WF-SVMSc

F1 scores for different methods:
- Surface
- BoW
- Bow+Surf.
- BoW+Lex.
- BoW+Lex.+Surf.
- Lexicons
- Lex.+Surf.
- FullUNT+Surf.
- MedUNT+Surf.
- RenLex+Surf.
Results on RenderES

- SVM
- MNB
- WF-SVM
- WF-SVMSc

The graph shows different combinations of features such as Surface, BoW, Bow+Surf., BoW+Lex., BoW+Lex.+Surf., Lexicons, Lex.+Surf., FullUNT+Surf., MedUNT+Surf., and RenLex+Surf., plotted against $F_1$ scores.
Results on Spanish data

- News domain: no improvement over the SVM BoW baseline.
- Social media: W+F-SVMSc with BoW+L+S significantly outperforms the SVM BoW baseline.
Results on PangLee

The graph shows the performance of different models and feature sets on the PangLee dataset, measured by precision-recall (PR) curves. The models compared include SVM, MNB, WF-SVM, and WF-SVMSc. The features include Surface, BoW, Bow+Surf, BoW+Lex, BoW+Lex+Surf, Lexicons, Lex.+Surf, SenticNet+Surf, RenLex+Surf, RenLexLinks+Surf, and SWN+Surf. The F1 scores are represented on the y-axis, ranging from 0.3 to 0.9.
Results on JRC-EN

- SVM
- MNB
- WF-SVM
- WF-SVMSc

F1 scores for various methods:
- Surface
- BoW
- BoW+Surf.
- BoW+Lex.
- BoW+Lex.+Surf.
- Lexicons
- Lex.+Surf.
- SenticNet+Surf
- RenLex+Surf
- RenLexLinks+S...
- SWN+Surf.
Results on RenderEN

- SVM
- MNB
- WF-SVM
- WF-SVMSc

F1 scores for different features:
- Surface
- BoW
- BoW+Surf.
- BoW+Lex.
- BoW+Lex.+Surf.
- Lexicons
- Lex.+Surf.
- SenticNet+Surf
- RenLex+Surf
- RenLexLinks+Surf
- SWN+Surf
Results on English datasets

- On reviews, none of the additions beat the baseline.
- On news data, two-layer models help a lot, especially with surface features.
- On social media, adding lexicons and surface feature helps a lot, especially in two-layer models (W+F-SVMSc).
- No benefit from using positive/negative link counts.
Model analysis [JRC-ES]

- Lexicon features!
- Nouns bear the most sentiment
- Capitalization
- Question phrases

full_unt_pos > 0.0
---yes: [OBJ] [88.0]: 161
---no: renderlex_noun_sum_neg > 0.0
   ---yes: [SUBJ/NEG] [4.0]: 4
   ---no: numcaps > 0.0386
      ---yes: renderlex_adjective_abs > 0.4069
         ---yes: h1w5 > 0.0312
         ---no: [SUBJ/POS] [4.0]: 5
         ---no: [OBJ] [5.0]: 6
      ---no: renderlex_all_sum > 3.866
         ---yes: [OBJ] [21.0]: 32
         ---no: h1w5 > 0.0833
            ---yes: [OBJ] [10.0]: 17
            ---no: full_unt_neg > 0.0
               ---yes: [OBJ] [4.0]: 8
               ---no: repeat_vowel > 0.0244
                  ---yes: [SUBJ/POS] [2.0]: 4
                  ---no: numvowel > 0.3429
                     ---yes: [OBJ] [113.0]: 129
                     ---no: renderlex_all_abs > 2.1249
                        ---yes: renderlex_all_sum > 2.7152
                           ---yes: [OBJ] [14.0]: 16
                           ---no: [SUBJ/NEG] [9.0]: 14
                           ---no: [OBJ] [43.0]: 47
            ---no: full_unt_neg > 0.0
               ---no: [OBJ] [399.0]: 601
Model analysis [RenderES]

numvowel > 0.3246
  +--yes: numcaps > 0.8462
    |    +--yes: [SUBJ/POS] [13.0]: 15
    |    +--no: renderlex_all_sum_neg > 0.2682
    |       +--yes: [SUBJ/POS] [7.0]: 9
    |       +--no: numvowel > 0.3566
    |          +--yes: [SUBJ/NEG] [177.0]: 257
    |          +--no: renderlex_adverb_sum_neg > 0.4899
    |             +--yes: [SUBJ/POS] [22.0]: 29
    |             +--no: repeat_letter > 0.0588
    |                +--yes: [SUBJ/POS] [20.0]: 32
    |                +--no: [SUBJ/NEG] [112.0]: 178
  +--no: renderlex_adverb_abs > 0.52
       +--yes: renderlex_adverb_abs > 0.5964
        |    +--yes: [SUBJ/POS] [10.0]: 19
        |    +--no: [SUBJ/NEG] [8.0]: 8
       +--no: negation > 0.0
        +--yes: repeat_letter > 0.0357
          |    +--yes: [SUBJ/NEG] [11.0]: 13
          |    +--no: [SUBJ/POS] [12.0]: 17
          +--no: full_unt_neg > 0.0
               +--yes: [SUBJ/NEG] [8.0]: 10
               +--no: length > 27.0
                +--yes: renderlex_noun_abs > 4.4911
                  |    +--yes: sad_face > 0.0
                  |    |    +--yes: [SUBJ/POS] [9.0]: 9
                  |    |    +--no: [SUBJ/NEG] [2.0]: 2
                  |    +--no: [OBJ] [15.0]: 22
                  +--no: [SUBJ/POS] [75.0]: 102

• Expression of sentiment through writing form
• Capitalization, vowels, repetition
• Negation
• Adverbs bear most sentiment
Model analysis [PangLee]

renderlex_adjective_sum > 0.1096
   --yes: senticnet > 15.509
   |   --yes: renderlex_adverb_abs > 8.1989
   |   |   --yes: swn_posneg_ratio > 5.2202
   |   |   |   --yes: [SUBJ/POS] [146.0]: 207
   |   |   --no: numpunc > 0.0313
   |   |   |   --yes: renderlex_pos_links > 8025.0
   |   |   |   |   --yes: renderlex_adjective_sum > 1.1693
   |   |   |   |   |   --yes: [SUBJ/POS] [20.0]: 25
   |   |   |   |   |   --no: [SUBJ/NEG] [28.0]: 53
   |   |   |   |   |   --no: [SUBJ/NEG] [61.0]: 80
   |   |   --no: [SUBJ/POS] [111.0]: 181
   |   --no: [SUBJ/POS] [126.0]: 164
   --no: numvowel > 0.2808
   --yes: renderlex_adjective_abs > 0.3998
   |   --yes: [SUBJ/NEG] [90.0]: 164
   |   --no: [SUBJ/POS] [15.0]: 17
   --no: swn_total_pos > 17.0
   --yes: [SUBJ/NEG] [35.0]: 37
   |   --no: renderlex_noun_sum > 7.8051
   |   --yes: [SUBJ/POS] [4.0]: 4
   |   --no: [SUBJ/NEG] [6.0]: 8
   --no: senticnet > 27.085
   --yes: [SUBJ/POS] [98.0]: 182
   --no: repeat_letter > 0.1193
   --yes: senticnet > 13.511
   |   --yes: [SUBJ/POS] [13.0]: 14
   |   --no: [SUBJ/NEG] [6.0]: 9
   --no: ... (continues)

- Lexicon features dominate
- Minor role of vowel and letter repetition
Model analysis [JRC-EN]

numcaps > 0.0345
+-yes: senticnet_neg > 1.113
|   ++-yes: [SUBJ/NEG] [4.0]: 4
|   |   |   ++-no: renderlex_adjective_sum_neg > 0.2178
|   |   |   |   +--yes: [SUBJ/POS] [5.0]: 10
|   |   |   |   |   --no: senticnet_neg > 0.084
|   |   |   |   |   |   ++-yes: swn_total_neg > 3.0
|   |   |   |   |   |   |   +--yes: [SUBJ/POS] [2.0]: 2
|   |   |   |   |   |   |   |   --no: numcaps > 0.037
|   |   |   |   |   |   |   |   |   ++-yes: [OBJ] [120.0]: 135
|   |   |   |   |   |   |   |   |   |   --no: [SUBJ/NEG] [3.0]: 7
|   |   |   |   |   |   |   |   |   |   |   ++-no: renderlex_all_abs > 1.5025
|   |   |   |   |   |   |   |   |   |   |   |   +--yes: senticnet_abs > 0.816
|   |   |   |   |   |   |   |   |   |   |   |   |   +--yes: renderlex_adverb_sum > 0.8143
|   |   |   |   |   |   |   |   |   |   |   |   |   |   +--yes: [SUBJ/POS] [1.0]: 2
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: swn_total_neg > 4.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   ++-yes: renderlex_adjective_sum > 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--yes: [SUBJ/NEG] [3.0]: 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: [OBJ] [5.0]: 5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: [OBJ] [70.0]: 74
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: [SUBJ/NEG] [3.0]: 3
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: [OBJ] [200.0]: 289
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   +--no: [OBJ] [302.0]: 512

• Similar to JRC-ES – important lexicon features, followed by sufrace features
• More focus on adjectives and adverbs as opposed to nouns
Model analysis [RenderEN]

- As opposed to Spanish social media, lexicons play a bigger role than surface features, but still a mix of both.
- Quality of lexicons?
- Writing style less indicative of sentiment?
Conclusions

- Across domains and languages, a two-layer model works better.
- Hierarchical representation did not give better results in any domain
- Feature scaling recommended
Conclusions

- We perform below state of the art on the reviews data, but improve performance on the news data compared to the dataset authors’ approach.
- Model analysis shows different feature importance in different domains.
- Comparing languages, some possible cultural differences in expression are apparent in social media.
Questions?