Sentiment Analysis of Twitter Data

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau

Columbia University

June 23, 2011

Ways in which people use Twitter

- Posting real-time sentiments about "everything" (*tweet moment*)

Ways in which people use Twitter

- Posting real-time sentiments about “everything” (*tweet moment*)
- Discussion on various topics (*tweet party*)

Ways in which people use Twitter

- Posting real-time sentiments about “everything” (*tweet moment*)
- Discussion on various topics (*tweet party*)
- Real-time complaints (e.g. flight delays)

Application: Social Media Analysis

- Start-ups: Radian6, Trendr, Twendz
- Want to correlate sentiment of posts to events
- Sales and Marketing: want to know how a product is doing in the market – summarize sentiment of posts along different dimensions (area, time, gender etc.)

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = <u>document</u>, sentence, phrase

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = <u>document</u>, sentence, phrase
- Task:

  #Apple is mega-awesome! :)

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = document, sentence, phrase
- Task:

Entity 1

#Apple *is* *mega-awesome! :)*

Positive

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = document, sentence, phrase
- Task:

Entity 1

#Apple *is* *mega-awesome! :)*

Positive

- **E**: set of entities in the world. **T**: set of tweets

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = <u>document</u>, sentence, phrase
- Task:

Entity 1

#Apple *is* *mega-awesome! :)*

Positive

- **E**: set of entities in the world. **T**: set of tweets
- Ideal function: $f : \mathbf{E} \times \mathbf{T} \rightarrow \{+, -, o\}$

Sentiment Analysis + Twitter Data

- Sentiment Analysis: Find polarity (+, -, o) of opinion in $\mathcal{X}$.
- $\mathcal{X}$ = document, sentence, phrase
- Task:

Entity 1

#Apple *is* *mega-awesome! :)*

Positive

- **E**: set of entities in the world. **T**: set of tweets
- Ideal function: $f : \mathbf{E} \times \mathbf{T} \rightarrow \{+, -, o\}$
- We (and other researchers) learn: $f : \mathbf{T} \rightarrow \{+, -, o\}$

Literature Survey

Paper	Data	Features	Conclusion
Go et. al. 2009	Distant learning, about 200 test examples	unigrams, bigrams, POS	unigram works best, bigrams and POS do not help
Pak and Paroubek 2010	Distant learning, about 200 test examples	unigrams, bigrams, POS	bigrams and POS help
Barbosa and Feng 2010	Distant learning, 1000 for development and 1000 for testing	meta-features, prior polarity of words, unigrams, POS	No feature analysis

Our Data

- 11,875 manually annotated tweets from a commercial source
- Collect a stream of tweets; translate tweets in foreign language using Google translator; randomly select tweets and annotate them for {*junk*, +, -, *o*}
- Ignore *junk* tweets. Leaves us with 8,753 tweets
- Use stratified sampling to get a balanced data-set of 5127 tweets (1709 each of +, -, o)

Resources and Pre-processing

Resources

- Emoticon→polarity dictionary: manually assign polarity to 170 emoticons from Wikipedia

Resources and Pre-processing

Resources

- Emoticon→polarity dictionary: manually assign polarity to 170 emoticons from Wikipedia
- Acronym dictionary (`noslang.com`): *lol* = laugh out loud

Resources and Pre-processing

Resources

- Emoticon→polarity dictionary: manually assign polarity to 170 emoticons from Wikipedia
- Acronym dictionary (`noslang.com`): *lol* = laugh out loud
- Dictionary of Affect in Language (DAL, Whissel 1989): prior polarity lexicon

Resources and Pre-processing

Resources

- Emoticon→polarity dictionary: manually assign polarity to 170 emoticons from Wikipedia
- Acronym dictionary (`noslang.com`): *lol* = laugh out loud
- Dictionary of Affect in Language (DAL, Whissel 1989): prior polarity lexicon
- WordNet: for increasing the coverage of DAL (Agarwal et. al. 2009)

Resources and Pre-processing

Resources

- Emoticon→polarity dictionary: manually assign polarity to 170 emoticons from Wikipedia
- Acronym dictionary (`noslang.com`): *lol* = laugh out loud
- Dictionary of Affect in Language (DAL, Whissel 1989): prior polarity lexicon
- WordNet: for increasing the coverage of DAL (Agarwal et. al. 2009)

Pre-processing

- Convert emoticons to polarity tags: positive emoticons = $||P||$
- Convert all URLS to tag $||U||$
- Convert all target mentions to $||T||$
- Convert *cooooool* to *coool*

Models

Two types of models:

- Tree kernels: represent tweets as trees encoding bag-of-words, POS tags, prior polarity scores of words (minimal feature engineering required)

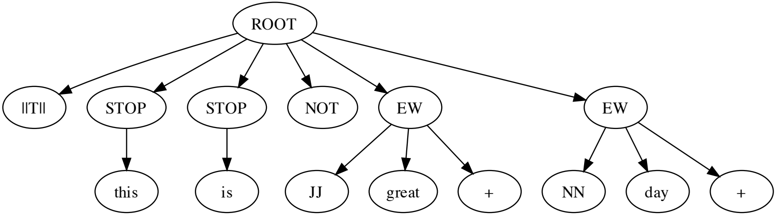
Models

Two types of models:

- Tree kernels: represent tweets as trees encoding bag-of-words, POS tags, prior polarity scores of words (minimal feature engineering required)
- Explicit feature engineering based model (Senti-feature model)

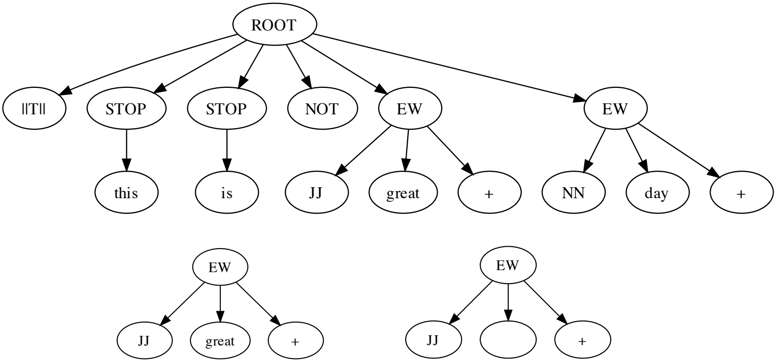
Model 1: Design of Tree Kernel

@Fernando this isn't a great day



Model 1: Design of Tree Kernel

@Fernando this isn't a great day



Model 2: Senti-features

- Non-polar: count of POS tags, acronyms, dictionary words, URLs, hashtags, newlines, % capitalized words, presence of exclamation marks and capitalization

Model 2: Senti-features

- Non-polar: count of POS tags, acronyms, dictionary words, URLs, hashtags, newlines, % capitalized words, presence of exclamation marks and capitalization
- Polar POS: For each POS (JJ, RB, VB, NN)
  - # of +/- words with that POS
  - Summation of prior polarity scores of words with that POS

Model 2: Senti-features

- Non-polar: count of POS tags, acronyms, dictionary words, URLs, hashtags, newlines, % capitalized words, presence of exclamation marks and capitalization
- Polar POS: For each POS (JJ, RB, VB, NN)
  - # of +/- words with that POS
  - Summation of prior polarity scores of words with that POS
- Polar Other: # of negation words, +/- words, +/- emoticons, +/- hashtags, summation of prior polarity scores of all the words

(Use Stanford tokenizer and POS tagger)

Experimental Set-up

- 5-fold cross-validation with SVM classifier
- For selecting “c” parameter for SVM we perform cross-validation on training set and report accuracy for the held-out test set

Results

Model	2-way (% Acc)	3-way (% Acc)
Majority class baseline	50	33
Unigram	71.35	56.58
Senti-features	71.27	56.31
Kernel	73.93	**60.60**
Unigram + Senti-features	**75.39**	60.50

Feature Analysis

Features	2-way (% Acc)	3-way (% Acc)
Unigram baseline	71.35	56.58
+ Non-polar	70.1	56.91
+ Polar POS	74.84	59.86
+ Polar Non-POS	**75.39**	**60.50**

Learning curve

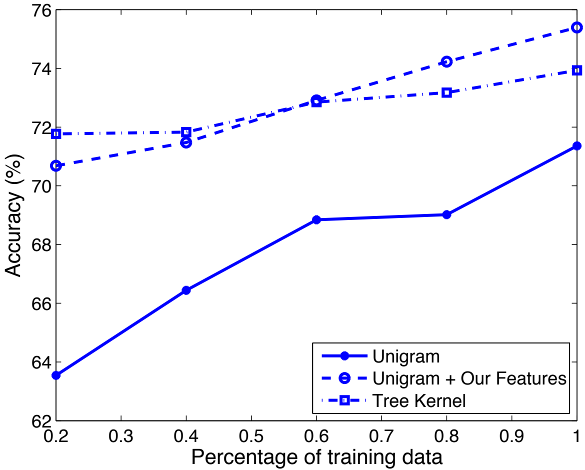


Figure: Learning curve for two-way classification task.

Future Work

- Study the affect of using different dictionaries (acronym, emoticon, DAL)

Future Work

- Study the affect of using different dictionaries (acronym, emoticon, DAL)
- Explore other linguistically rich features: dependency trees, FrameNet, Verbnet

Future Work

- Study the affect of using different dictionaries (acronym, emoticon, DAL)
- Explore other linguistically rich features: dependency trees, FrameNet, Verbnet
- Tree kernels seem to encode many more features but their performance is a little less than feature extraction methods (same observation in relation extraction)

Conclusion

- Proposed two models for sentiment analysis of twitter

Conclusion

- Proposed two models for sentiment analysis of twitter
- On both 2-way and 3-way classification tasks beat the baseline model by over 4%

Conclusion

- Proposed two models for sentiment analysis of twitter
- On both 2-way and 3-way classification tasks beat the baseline model by over 4%
- Make available two new resources for the task: Acronym dictionary, Emoticon to polarity dictionary

Conclusion

- Proposed two models for sentiment analysis of twitter
- On both 2-way and 3-way classification tasks beat the baseline model by over 4%
- Make available two new resources for the task: Acronym dictionary, Emoticon to polarity dictionary
- No matter how noisy and non-standard English the data might seem at first, there is hope for being able to use linguistically rich resources/features