



TwitterHawk: A Feature Bucket Approach to Sentiment Analysis

William Boag, Peter Potash, Anna Rumshisky

Text Machine Lab for Natural Language Processing, UMass Lowell

The Data

SemEval 2015

Task A: phrase-level

100517814976266242 324748107 1 3 negative
Eerily, shockingly quiet day at work today,

Task B: message-level

264259009274576896 261890798 positive
Huge day of NBA tomorrow. Whoop. League pass is awesome.

Task C: topic-based

255732290246815744 billy cundiff negative
Billy Cundiff may be leaving Washington.
Hopefully he won't miss the door on the way out.

Train Data

	positive	negative	neutral
phrase	4832	2549	384
tweet	3640	1458	4586

Test Data

	positive	negative	neutral
Task A	1899	1008	190
Task B	1040	365	987
Task B - Sarcasm	33	40	13
Task C	870	260	1256

SemEval 2015 Task 10

TwitterHawk is a system for sentiment analysis of tweets that participated in SemEval 2015 Task 10 - Subtasks A through D. The system performed competitively, most notably placing 1st in topic-based sentiment classification (Subtask C) and ranking 4th out of 40 in identifying the sentiment of sarcastic tweets (for Subtask B). In addition, it placed 5th/11 for Subtask A (sentiment of a tweet's sub-phrase), 10th/40 for Subtask B (sentiment of a full tweet), and 3rd/6 for Subtask D (summarization of Subtask C).

TwitterHawk used a supervised learning approach to perform three-way classification to assign positive, negative, or neutral labels. System development efforts focused on text pre-processing and feature engineering, with a particular focus on handling negation, integrating sentiment lexicons, parsing hashtags, and handling expressive word modifications and emoticons. Two separate classifiers were developed, one for phrase-level and the other for tweet-level sentiment classification

Features

We built two classifiers, one for phrase-level sentiment and one for tweet-level. Both classifiers contained the following features. For a text span (phrase or tweet)

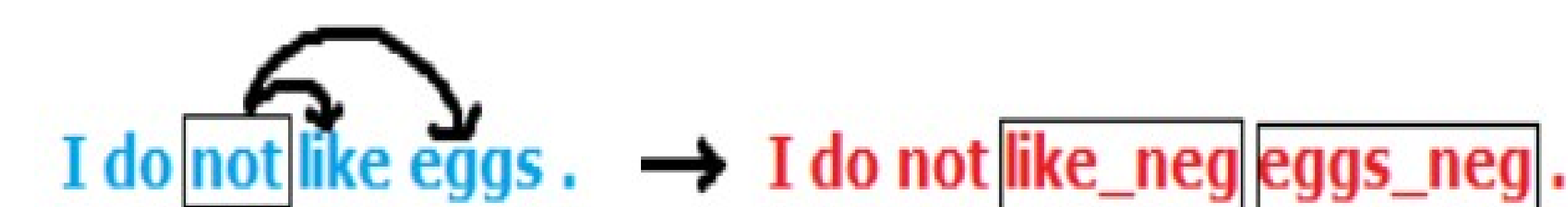
- presence or absence of:
 - raw bag-of-words (BOW) unigrams
 - normalized/stemmed BOW unigrams
 - stemmed segmented hashtag BOW
 - user mentions
 - URLs
 - hashtags;
- number of question marks and number of exclamation points
- number of positive, negative, and neutral emoticons
- whether the text span contains and elongated words

In addition, the phrase-level classifier was supplemented with these additional features.

- normalized BOW unigram features derived from 3 tokens preceding the target phrase;
- normalized BOW unigram features derived from 3 tokens following the target phrase;
- length 2, 3, and 4 character prefixes and suffixes for each token in the target phrase;
- whether the phrase was in all caps;
- whether phrase contained only stop words;
- whether a phrase contained only punctuation;
- whether the phrase contained a word whose length is eight or more

Text Preprocessing

- Tokenization and POS-Tagging using ARK Tweet NLP
- Spell Correction
- Hashtag Segmentation
- Remove Stop Words
- Lowercase All Text
- Negation Contexts



Lexicons

We used the features derived from the following lexicons: AFINN, Opinion Lexicon, Brown Clusters,, Hashtag Emotion, Sentiment140, Hashtag Sentiment, Subjectivity, and General Inquirer.

We used the following features for lexicons that provided scores:

- the average sentiment score for the text span
- the total number of positively scored words in the span
- the maximum score (or zero if no words had a sentiment score)
- the score of the last positively scored word
- three most influential (most positive or most negative) scores for the text span; this was only used by the phrase-level system

For lexicons that labeled words as members of given classes, our features were the number of words that belonged to each class.

Spelling Correction

- Non-prose
 - URL, mention, number, emoticon, proper noun
- Abbreviations
 - “2moro”, “omg”
- Elongated Words
 - “heyyy”
- Colloquial
 - “lol”, “haha”, “hahaha”, ...
- Else: PyEnchant spell checker

Hashtag Segmentation

“On the bright side we have school today... Tomorrow and the day after ! #killmenow”

“On the bright side we have school today... Tomorrow and the day after ! #killmenow”

	Live Journal 2014	SMS 2013	Twitter 2013	Twitter 2014	Twitter Sarcasm 2014
nBOW	58.87	57.22	59.19	60.27	46.76
nBOW +hashtag	58.94	57.81	60.09	61.38	53.00

Official Results

	micro-average F1	rank
Task A	82.32	5 th / 11
Task B	61.99	10 th / 40
Task B - Sarcasm	61.24	4 th / 40
Task C	50.51	1 st / 7