Detecting Sarcasm is Extremely Easy ;) (Parde & Nielson 2018)

Gist

- Doain general sarcasm detection system
- Applied to twitter and amazon product reviews
- Contains error breakdown

Intro

- Sarcasm is difficult even for humans
  - Primariy indicated using prosodic rather than syntactic cues
- Previous approaches have been largely domain specific, this is an attempt at a general purpose sarcasm detection system

Background

- Tweets may be expecially challenging because the text limit may encourage brief coments that require more contextual information
  - The example of saying "Great" just after an election may be understandable to others at that point in time but for an automatic system that is not aware of such events, it becomes very difficult.
- Rajadesingan et al 2015 "developed behavioral models of sarcasm usage specific to individual users" (p. 22)
- Sarcastic tweets are sampled using hashtags indicating sarcasm, Amazon reviews are sampled using star ratings

- The prior work (Parde and Nielson 2017) created a domain adaption system that was used prior to training the model, this achieved better performance "in predicting sarcasm in Amazon product reviews over models that trained on reviews alone or on a a simple combination of reviews and tweets" (p. 22)

Sarcasm detection methods

Data source

- Train
  - 3998 tweets, 1003 Amazon product reviews
- Test
  - 1000 tweets (609 non-sarcastic and 391 sarcastic)
  - 251 amazon reviews (87 sarcastic and 164 non-sarcastic)

Features

- Contains Twitter Indicator
  - "Multiple binary features indicating whether the instance contains one of th esarcasm-related has-tags, emoticons, and/or indicator phrases learned by Maynard and Greenwood (2014)" (p 23)
- "Twitter-Based predicates and situations
  - "Multiple binary features indicating whether the instance contains a positive predicate, a positive sentiment and/or negative situation phrase learned by Riloff et al. (2013) from a corpus of tweets. Includes an additional binary feature that indicates whether one ofo those positive preedicates or sentiments precedes one of those negative situation phrases by <= 5 tokens"
- Star Rating
  - "Number of stars associated with the review" (p 23) left blank for tweets
- Laughter and interjections
  - "Multiple binary features indicatingi whether the instance contains: hahahaa, haha, hehehe, hehe,jajaja, jaja, lol, lmao, rofl, wow, ugh, and/or huh" (p 23)
- Specific characters
  - "Multiple binary features indicating whether the instance contains an ellipsis, an exclamation mark and/or a question mark" (p 23)
- Polarity
  - "Multiple features indicating the most polar (positive or negative) unigram in the

instance, the polarity score (-5 to +5) associated with that unigram, the average polarity of the instance, the overall (sum) polarity for the instance, the largest difference in polarity between any two words in the instance, and the percentages of positive and negative words in the instance" (p 23)
- Subjectivity
  - "The percentages of strongly subjective positive words, strongly subjective negative words, weakly subjective positive words, and weakly subjective negative words in the instance" (p. 23)
- PMI
  - "Multiple features indicating the highest number of consecutive repeated characters in the instance (e.g., Sooooo => 5) and the higehest number of consecutive punctuation characters in the instance" (p 23)
- All-Caps
  - "Multiple features indicating the number and percentage of all-caps words in the instance" (p. 23)
- Bag of words
  - Features for words most closely associated with the different training pairs (e.g. Amazon - Sarcastic, Amazon non-sarcastic, twitter sarcastic etc.)
  - Features for most common words in each of these different class source pairings.

Classification Algorithm

Naive bayes using Daume III (2007)'s method for domain adaptation. to generate source, target and general feature mappings.

Results

.59 F-score on twitter data, 1% over previous literature (not really meaningful) Recall of system is much higher (.68 vs .62) at the cost of some precision (53 vs 55). .78 F-score on Amazon reviews, much higher than previous results (Buschmeier et al 2014) (78 to 74). Once again, much higher recall (82 to 69) at the cost of precision (75 to 85)

Error analysis

- Many did not convey sarcasm once the sarcastic hash tags were removed (23)
- 8 only had sarcastic content in the hashtags
- 13 tweets were discovered not to be sarcastic upon manual inspection
- 63 Required world knowledge to know that it was sarcastic.

- Highly negative
- Reviews also had story-like passages that were sarcastic. E.g. a narrative where the thing being reviewed is doing things that are impossible.