



Sentiment analysis is not solved!

Assessing and probing sentiment classifiers

Jeremy Barnes, Lilja Øvrelid, Erik Velldal
University of Oslo

Research questions

R1: Are there any sentences that all current sentiment classifiers still get wrong?

R2: If so, are there any patterns (linguistic or paralinguistic) that these sentences present?

4 sentiment classifiers

BOW - SVM
BiLSTM
ELMo
BERT

6 datasets

MPQA
OpeNER
SemEval 2013
Stanford Sentiment Treebank
Täckström
Thelwall

Code available

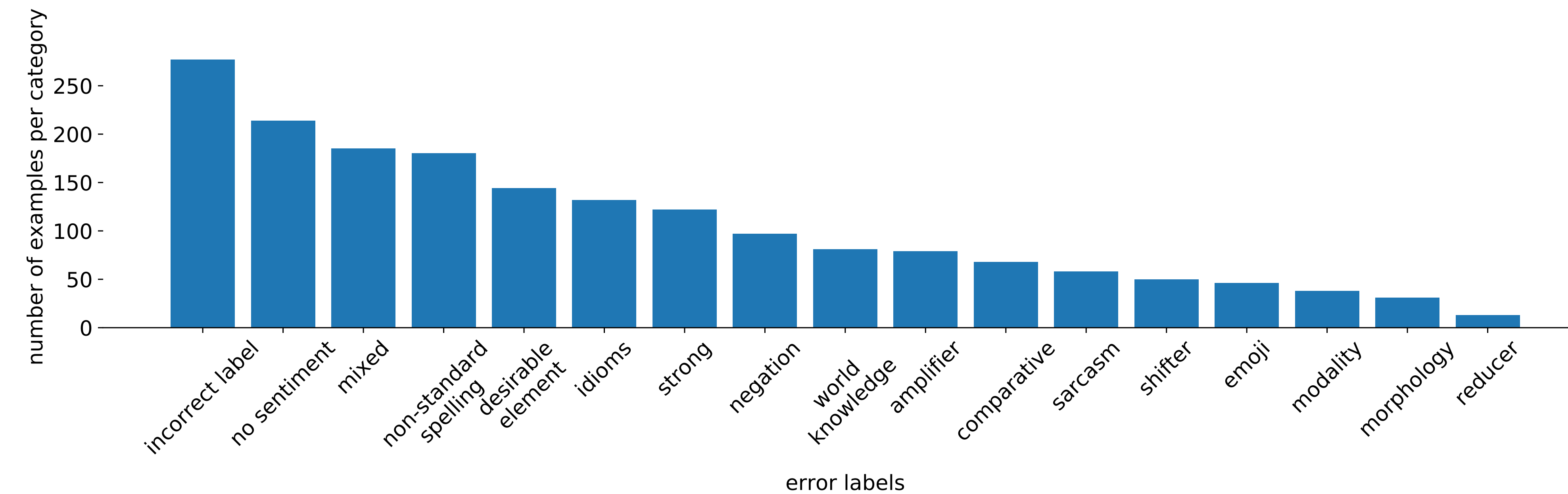
https://github.com/litgoslo/assessing_and_probing_sentiment



Contact: jeremycb@ifi.uio.no

Results

We collected a total of 836 sentences that all models misclassified. The negative class contributed the most (250), while strong positive has the fewest (95).



Error analysis and labeling scheme

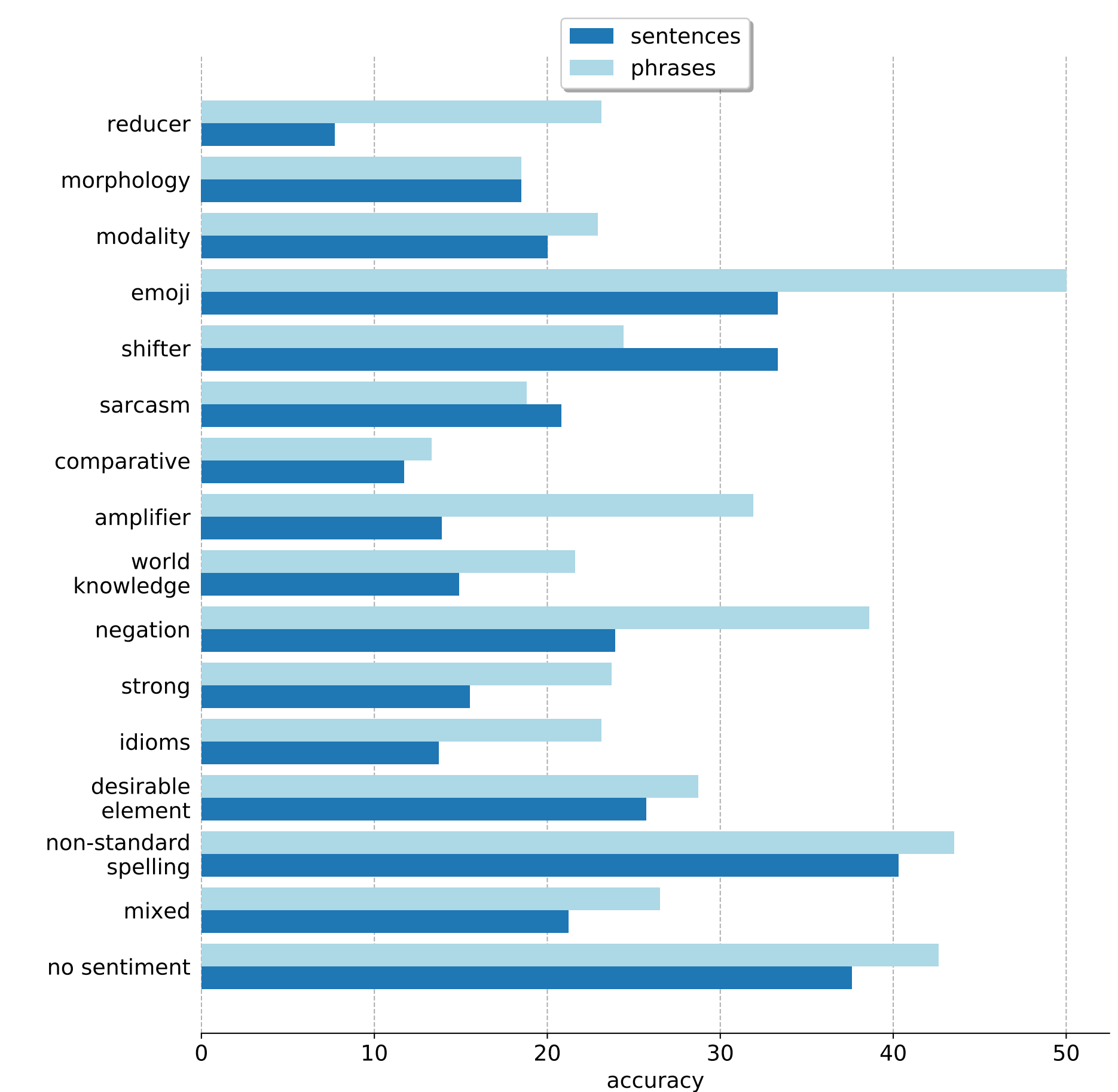
negation	strong	amplifier	reducer	desirable element	comparative
shifter	modality	world knowledge	morphology	spelling	idioms
sarcasm	emoji	no sentiment	mixed	incorrect label	

We label the collected sentences with **17 linguistic and paralinguistic labels** (shown above), taken from previous analyses and from the error analysis itself.

gold	sentence	error labels
neg.	And what they have said about him has not been pretty .	positive::negated
str. pos.	No problems whatsoever .	negative::negated::amplified
pos	At my final IWD event. Bring on the bubbles! #IWD	idioms::positive::non-standard spelling
neg.	I have to take the TAKS test tomorrow..... #what	non-standard spelling::world-knowledge
pos.	Newcastle Utd upping the ante in the 2nd half.	idioms::positive
str. pos.	Tender yet lacerating and darkly funny fable .	difficult vocab::positive::amplified::strong
str. neg.	The ending is a cop-out .	idioms::negative::strong

Case study

The BERT model trained on SST-phrases (each phrase in a constituency tree is labeled with sentiment) is better able to model compositional phenomena than the model trained only on sentence labels.



Take aways

Many of the phenomena found in our challenge test set are well known in the literature but not easily learnable from sentiment corpora

To move forward we can either:

1. build new sentiment datasets in such a way that we explicitly incorporate these phenomena.
2. use external sources of knowledge (other annotated datasets, lexicons, etc) to improve sentiment classifiers ability to predict these phenomena.