

Semi-Supervised Priors for Microblog Language Identification

Task

Identify the language in which a microblog post is written.

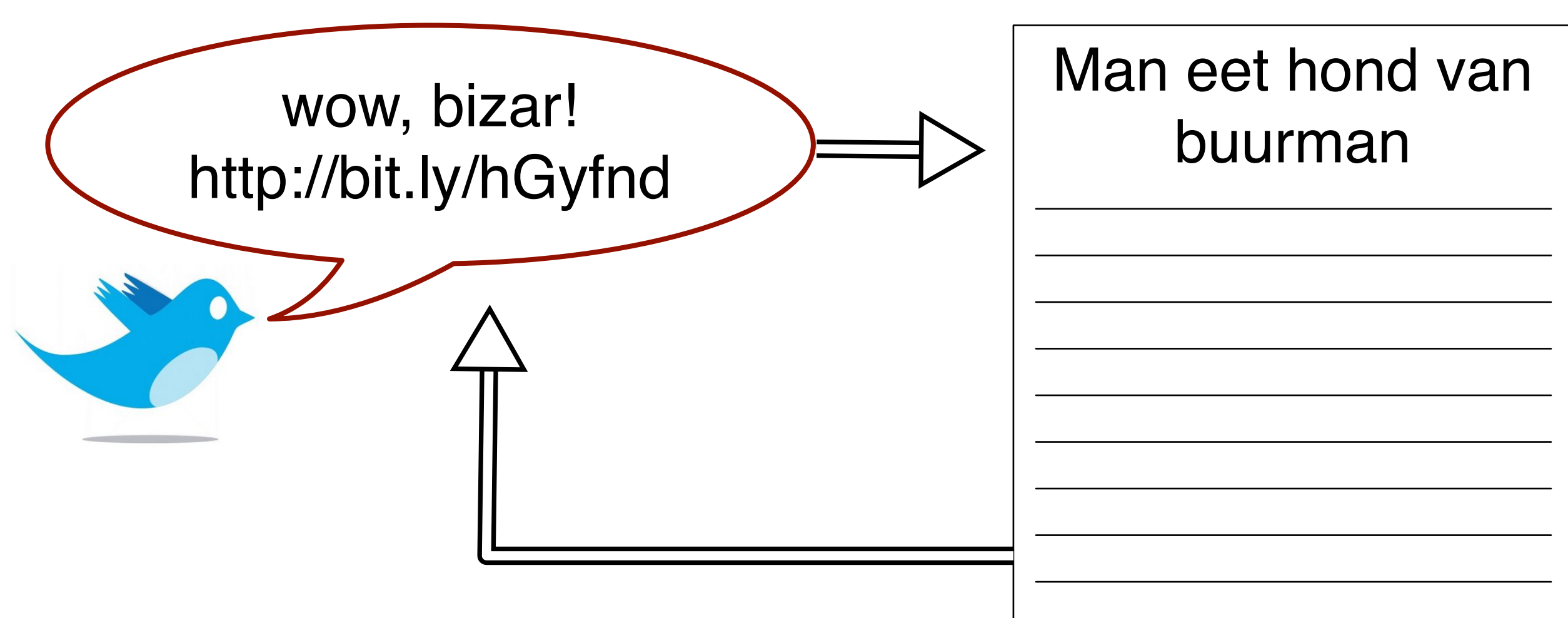
Challenges

1. Posts in microblogs are very short (~140 characters)
2. Language usage in posts is ambiguous

To counter these challenges we introduce two **semi-supervised priors**

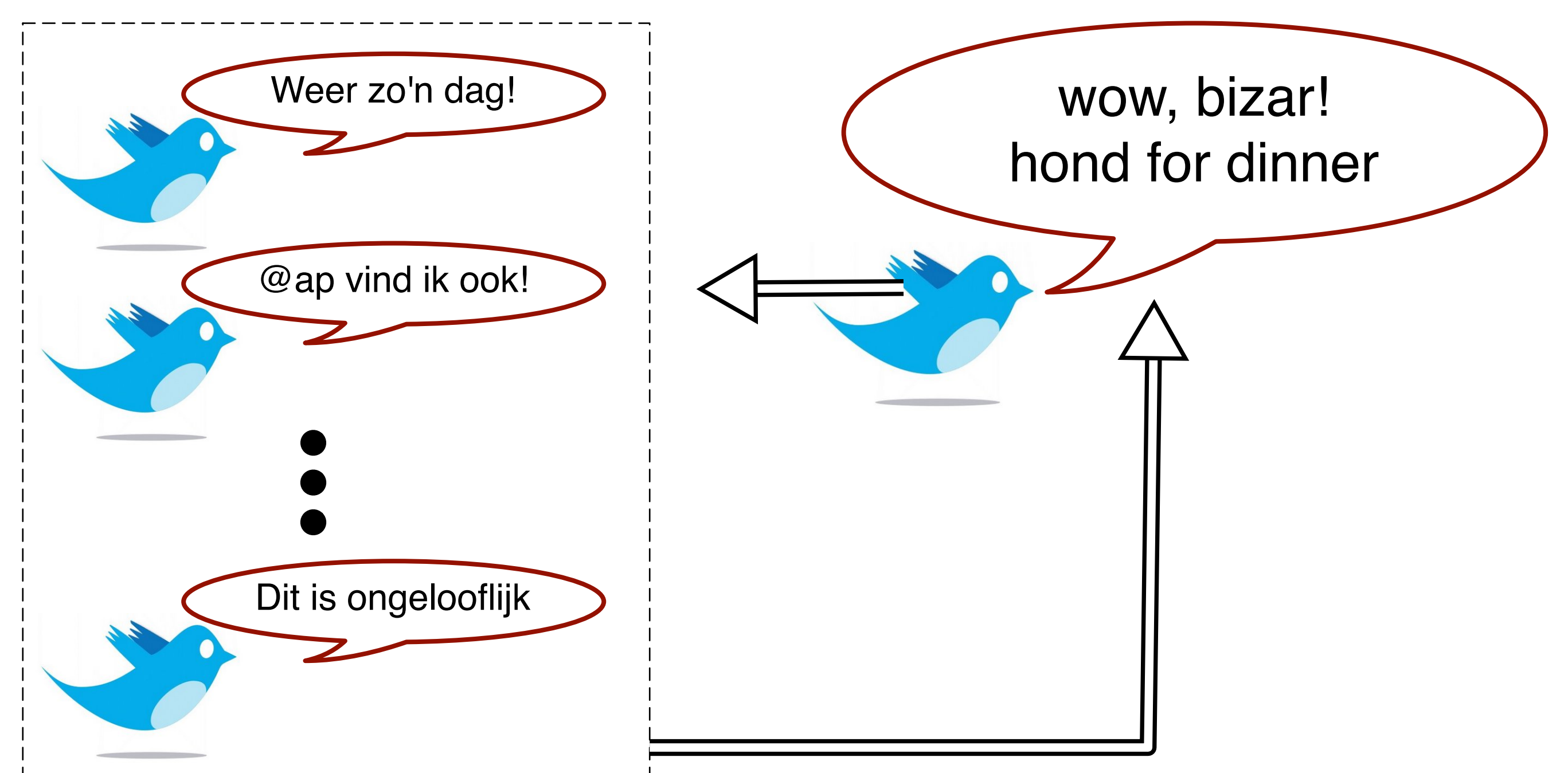
Link prior

Guess language of each of the pages linked to in this post; combine into a language prior for this post's links.



Blogger prior

Guess language of previous posts by the same blogger; combine into one language prior for this blogger.

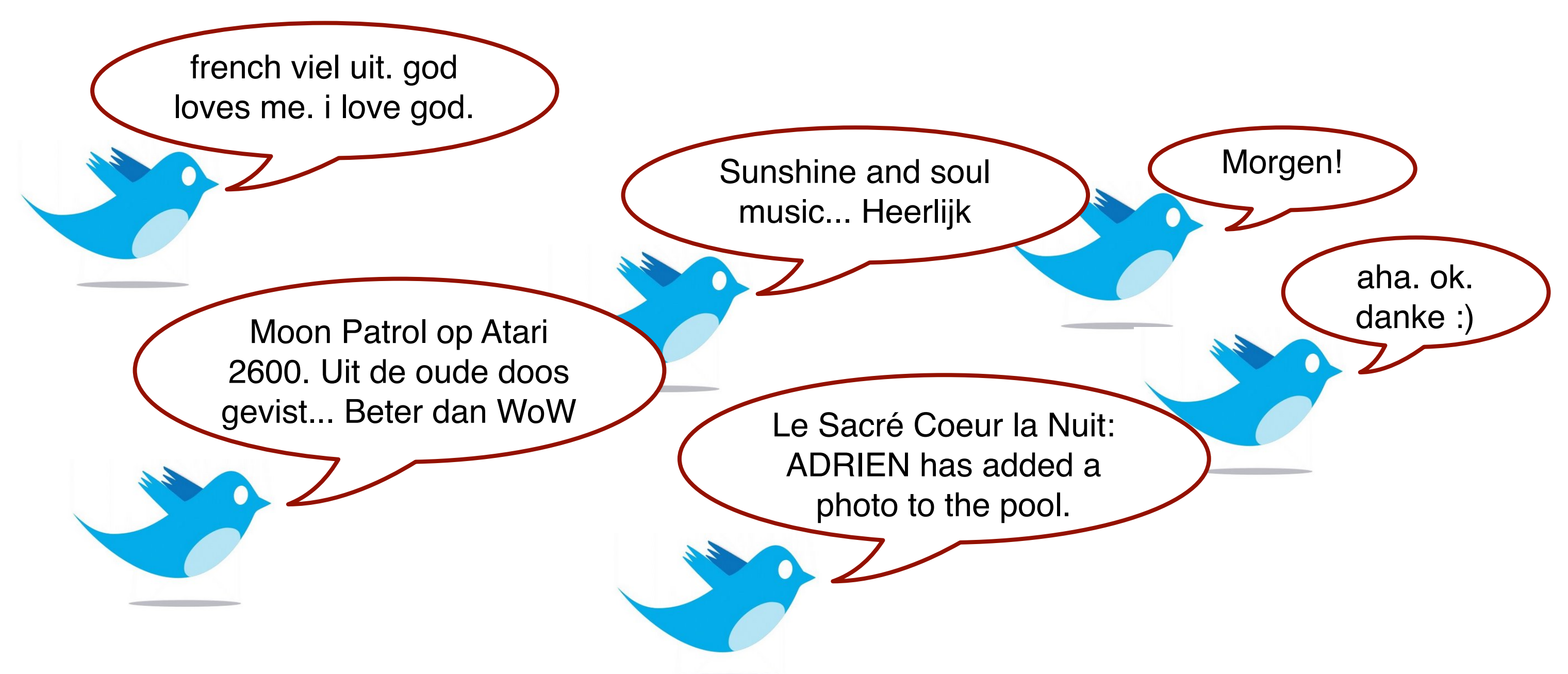


Data and Experiments

- > Microblog platform: Twitter
- > Five languages (Dutch, English, French, German, and Spanish)
- > Per language: 1,000 labeled tweets (50/50 training and test set)
- > Baseline: n-gram-based model (TextCat).

Results and Errors

Run	Dutch	English	French	German	Spanish	Overall
Baseline	90.4%	91.6%	92.2%	95.4%	85.2%	91.0%
<i>Baseline + priors</i>						
Blogger-based	94.6%	93.8%	94.8%	96.4%	84.6%	92.8%
Link-based	92.0%	90.6%	92.6%	92.8%	83.0%	90.2%
Both priors	94.4%	95.0%	94.0%	97.2%	85.4%	93.2%



Conclusions and Future work

Accuracy for all languages improves using at least one prior.

Further improvement can come from:

1. Additional priors (e.g., tag prior, reply prior)
2. Post-dependent weighting of priors