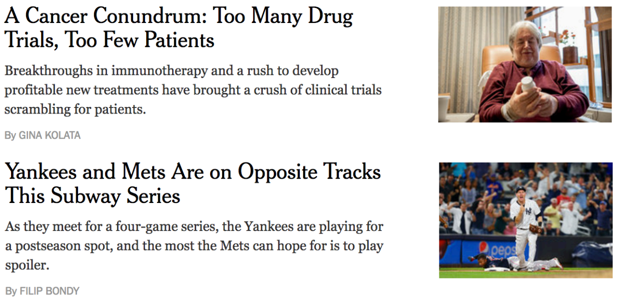
Text Classification



A Cancer Conundrum: Too Many Drug Trials, Too Few Patients

Breakthroughs in immunotherapy and a rush to develop profitable new treatments have brought a crush of clinical trials scrambling for patients.

By GINA KOLATA

→ Health

Yankees and Mets Are on Opposite Tracks This Subway Series

As they meet for a four-game series, the Yankees are playing for a postseason spot, and the most the Mets can hope for is to play spoiler.

By FILIP BONDY

→ Sports

~20 classes

- 20 Newsgroups, Reuters, Yahoo! Answers, …

Entailment

- Three-class task over sentence pairs
- Not clear how to do this with simple bag-of-words features

A soccer game with multiple males playing.
ENTAILS
Some men are playing a sport.

A black race car starts up in front of a crowd of people.
CONTRADICTS
A man is driving down a lonely road

A smiling costumed woman is holding an umbrella.
NEUTRAL
A happy woman in a fairy costume holds an umbrella.

Bowman et al. (2015)

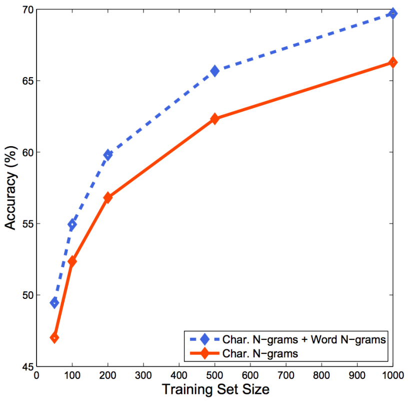
Authorship Attribution

- Statistical methods date back to 1930s and 1940s
  - Based on handcrafted heuristics like stopword frequencies
  - Early work: Shakespeare's plays, Federalist papers (Hamilton v. Madison)
- Twitter: given a bunch of tweets, can we figure out who wrote them?
  - Schwartz et al. EMNLP 2013: 500M tweets, take 1000 users with at least 1000 tweets each
- Task: given a held-out tweet by one of the 1000 authors, who wrote it?

Authorship Attribution

- SVM with character 4-grams, words 2-grams through 5-grams
- 1000 authors, 200 tweets per author => 30% accuracy
- 50 authors, 200 tweets per author => 71.2% accuracy

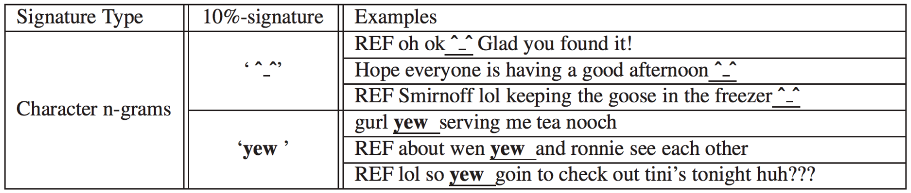
50-author case



Schwartz et al. (2013)

Authorship Attribution

- k-signature: n-gram that appears in k% of the authors tweets but not appearing for anyone else — suggests why these are so effective



Signature Type	10%-signature	Examples
Character n-grams	'^_^'	REF oh ok ^_^ Glad you found it!
		Hope everyone is having a good afternoon ^_^
		REF Smirnoff lol keeping the goose in the freezer ^_^
	'yew '	gurl **yew** serving me tea nooch
		REF about wen **yew** and ronnie see each other
		REF lol so **yew** goin to check out tini's tonight huh???

Schwartz et al. (2013)

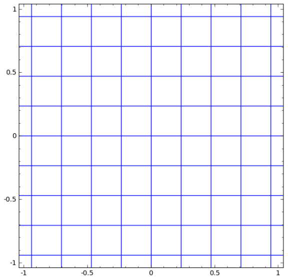
Neural Networks

$$\mathbf{z} = g(V f(\mathbf{x}) + \mathbf{b})$$

Nonlinear transformation ↗ ↖ Warp space ↑ Shift

$$y_{\text{pred}} = \text{argmax}_y \mathbf{w}_y^\top \mathbf{z}$$

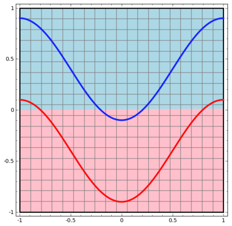
- Ignore shift / +**b** term for the rest of the course



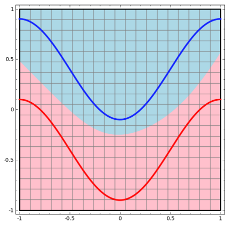
Taken from http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Neural Networks

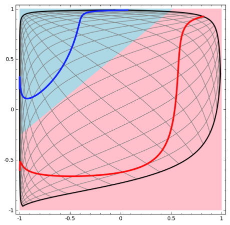
Linear classifier



Neural network



Linear classification in the transformed space!



Taken from http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

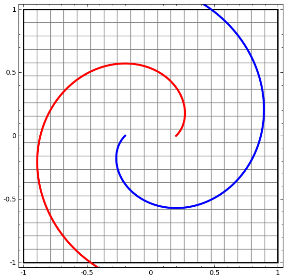
Deep Neural Networks

$$\mathbf{z}_1 = g(V_1 f(\mathbf{x}))$$

$$\mathbf{z}_2 = g(V_2 \mathbf{z}_1)$$

$$\ldots$$

$$y_{\text{pred}} = \text{argmax}_y \mathbf{w}_y^\top \mathbf{z}_n$$



Taken from http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/