

Multiclass Classification



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.



Entity Linking

Although he originally won the event, the United States Anti-Doping Agency announced in August 2012 that they had disqualified **Armstrong** from his seven consecutive Tour de France wins from 1999–2005.



Lance Edward Armstrong is an American former professional road cyclist



Armstrong County is a county in Pennsylvania...

- ▶ 4,500,000 classes (all articles in Wikipedia)
- ▶ Needs a very different structure for $f(\mathbf{x}, \mathbf{y})$ to accommodate so many classes



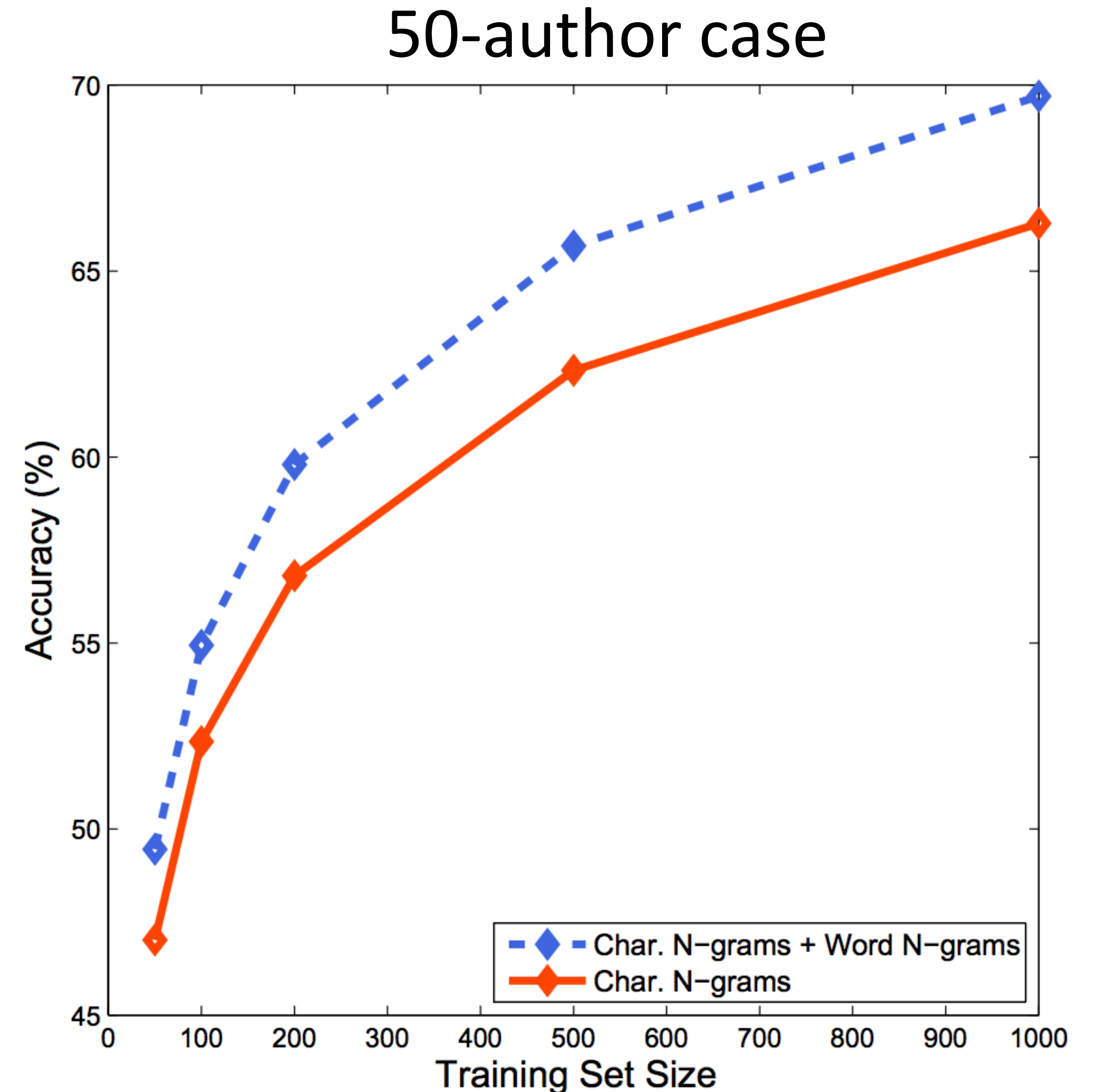
Authorship Attribution

- ▶ Statistical methods date back to 1930s and 1940s
 - ▶ Based on handcrafted heuristics like stopword frequencies
- ▶ Early work: Shakespeare's plays, Federalist papers (Hamilton vs. 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, word 2-grams through 5-grams
- ▶ 1000 authors, 200 tweets per author => 30% accuracy
- ▶ 50 authors, 200 tweets per author => 71.2% accuracy





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 <u>^ _ ^</u> Glad you found it!
		Hope everyone is having a good afternoon <u>^ _ ^</u>
		REF Smirnoff lol keeping the goose in the freezer <u>^ _ ^</u>
	‘yew ’	gurl <u>yew</u> serving me tea nooch
		REF about wen <u>yew</u> and ronnie see each other
		REF lol so <u>yew</u> goin to check out tini’s tonight huh???



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

Word n-grams	.. lal	REF aww those are cool where u get those.. how do ppl react.. <u>lal</u>
		Ludas album is gone be hott.. <u>lal</u>
		Dayum refs don't get injury timeouts.. <u>lal</u> .. get him off the field..
	smoochies , e3	I'm just back after takin' a very long, icy cold shower.....Shivering <u>smoochies,E3</u> http://bit.ly/4CzzP9
		A blue stout or two would be nice as well, Purr!Blue smooth <u>smoochies,E3</u> http://bit.ly/75D4fO
		That is soooooooooooooooooooooo unfair!Double <u>smoochies,E3</u> http://bit.ly/07sXRGX