IN SEARCH OF STYLES IN LANGUAGE

Identifying
- ✓ Deceptive Product Reviews
- ✓ Wikipedia Vandalism
- ✓ The Gender of Authors

via **Statistical Stylometric Analysis**

Yejin Choi

Stony Brook University

Styles in Language

Research Papers?   New York Times?   Blogs?

"So how can you spot a fake review? Unfortunately, it's difficult, but with some technology, there are a few warning signs:"

Styles in Language

Research Papers?   New York Times?   Blogs?

"So how can you spot a fake review? Unfortunately, it's difficult, but with some technology, there are a few warning signs:"

"To obtain a deeper understanding of the nature of deceptive reviews, we examine the relative utility of three potentially complementary framings of our problem."

Styles in Language

Research Papers?   New York Times?   Blogs?

"<u>So how can you</u> spot a fake review? Unfortunately, it's difficult, but <u>with some technology</u>, there are a few warning signs:"

"<u>To obtain a deeper understanding of the nature of deceptive reviews, we examine the relative utility of three potentially complementary framings of our problem</u>."

"As online retailers increasingly depend on reviews as <u>a sales tool</u>, <u>an industry of fibbers and promoters</u> has <u>sprung up</u> to buy and sell <u>raves for a pittance</u>."

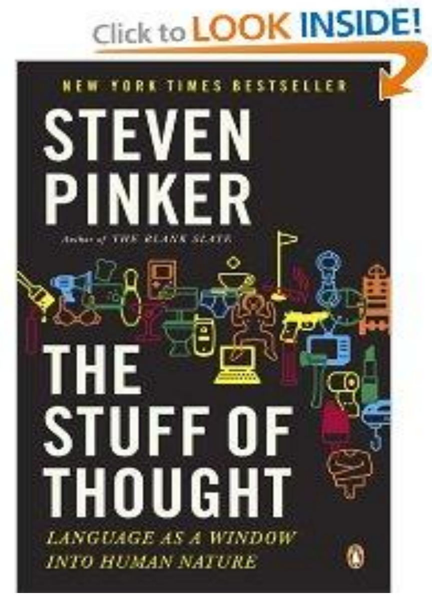
Why different Styles in Language?

Influencing factors:

- Convention / customary style of certain genres
- Expected audience
- Intent of the author
- Personal traits of the author

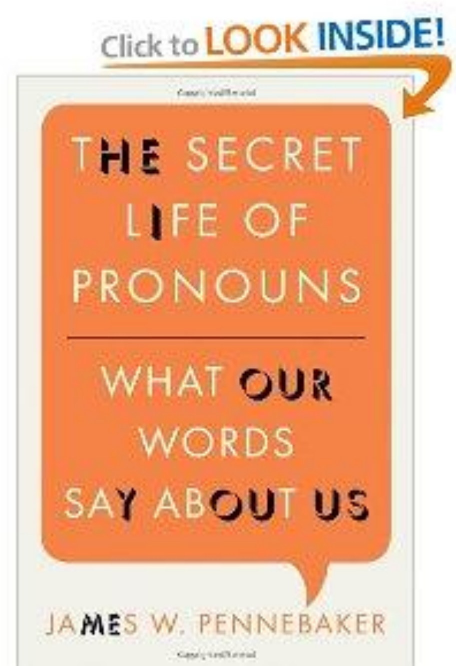
The Stuff of Thought:
Language as a Window into Human Nature

-- Steven Pinker



The Secret Life of Pronouns:
What our Words Say about Us

-- James W. Pennebaker



What constitute Styles in language?

- Lexical Choice
- Grammar / Syntactic Choice
- Cohesion / Discourse Structure
- Narrator / Point of View
- Tone (formal, informal, intimate, playful, serious, ironic, condescending)
- Imagery, Allegory, Punctuation, and more

Computational analysis of styles

Mostly limited to

lexical choices

shallow syntactic choices (part of speech)

--- notable exception: Raghavan et al. (2010)

Previous Research in NLP

Genre Detection

- Petrenz and Webber, 2011
- Sharoff et al., 2010
- Wu et al., 2010
- Feldman et al., 2009
- Finn et al., 2006
- Argamon et al., 2003
- Dewdney et al., 2001
- Stamatatos et al., 2000
- Kessler et al., 1997

Authorship Attribution

- Escalante et al., 2011
- Seroussi et al., 2011
- Raghavan et al., 2010
- Luyckx and Daelemans, 2008
- Koppel and Shler, 2004
- Gamon, 2004
- van Halteren, 2004
- Spracklin et al., 2008
- Stamatatos et al., 1999

In this talk: three case studies of **stylometric analysis**

- ✓ Deceptive Product Reviews
- ✓ Wikipedia Vandalism
- ✓ The Gender of Authors

In this talk: three case studies of **stylometric analysis**

- ✓ Deceptive Product Reviews
- ✓ Wikipedia Vandalism
- ✓ The Gender of Authors

Underlying themes:

A. Discovering "language styles" in a broader range of real-world NLP tasks
B. Learning (statistical) stylistic cues beyond shallow lexico-syntactic patterns.

In this talk: three case studies of **stylometric analysis**

- ✓ **Deceptive Product Reviews**
- ✓ Wikipedia Vandalism
- ✓ The Gender of Authors

Motivation

- Consumers increasingly rate, review and research products online
- Potential for opinion spam
  - Disruptive opinion spam
  - Deceptive opinion spam

Motivation

- Consumers increasingly rate, review and research products online
- Potential for opinion spam
  - Disruptive opinion spam
  - Deceptive opinion spam

Motivation

- Consumers increasingly rate, review and research products online
- Potential for opinion spam
  - Disruptive opinion spam
  - Deceptive opinion spam

"My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn't ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago."

Deceptive or Truthful?

"My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn't ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago."

"I have stayed at many hotels traveling for both business and pleasure and I can honestly say that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples."

"My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn't ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago."

deceptive

Gathering Data

- Label existing reviews?
  - Can’t manually do this

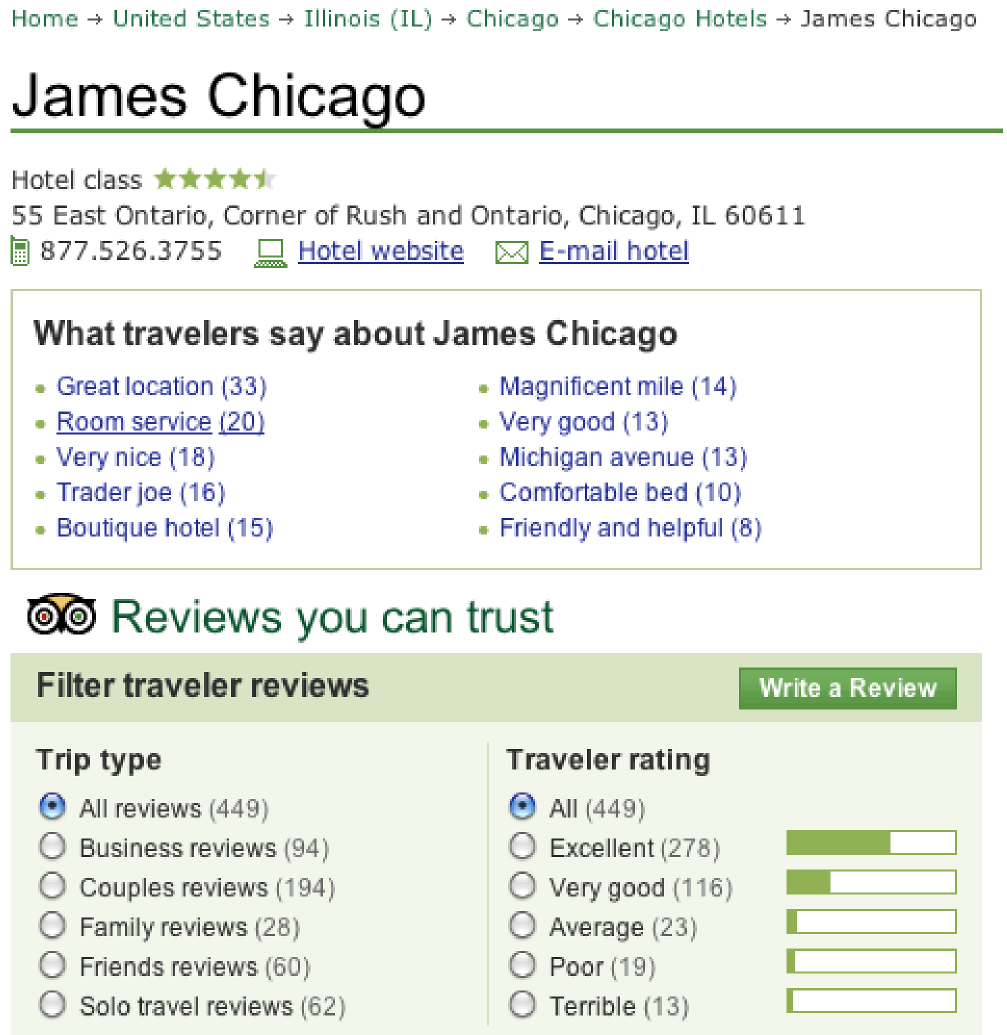
Gathering Data

- ~~Label existing reviews?~~
  - Can't manually do this

- Create new reviews
  - By hiring people to write fake POSITIVE reviews
  - Using Amazon Mechanical Turk

Gathering Data

- Mechanical Turk
  - 20 hotels
  - 20 reviews / hotel
  - Offer $1 / review
  - 400 reviews



Gathering Data

- Mechanical Turk
  - 20 hotels
  - 20 reviews / hotel
  - Offer $1 / review
  - 400 reviews

Gathering Data

- Mechanical Turk
  - 20 hotels
  - 20 reviews / hotel
  - Offer $1 / review
  - 400 reviews

Gathering Data

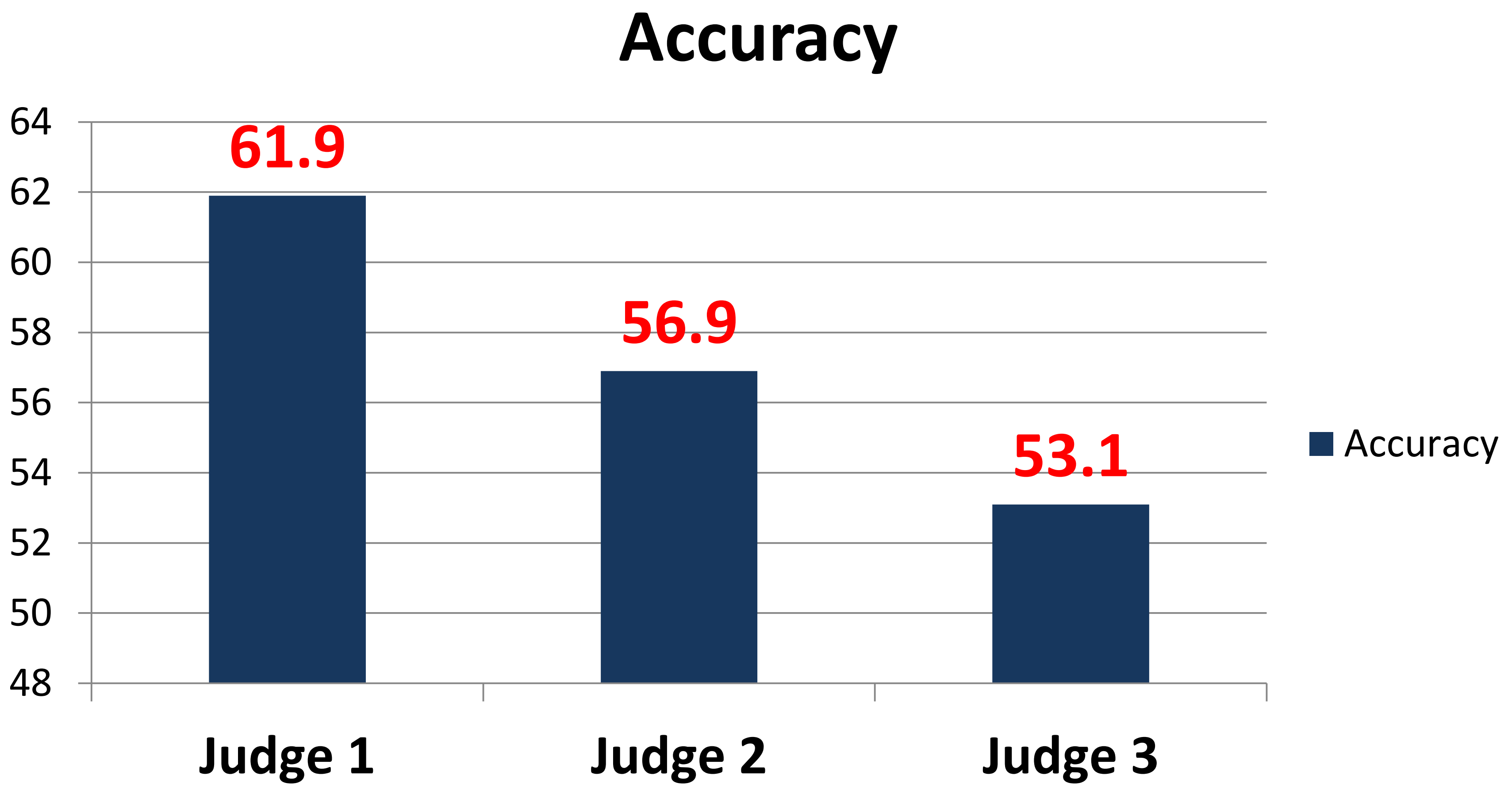
- Mechanical Turk
  - 20 hotels
  - 20 reviews / hotel
  - Offer $1 / review
  - 400 reviews
- Average time spent:
  > 8 minutes
- Average length:
  > 115 words

Human Performance

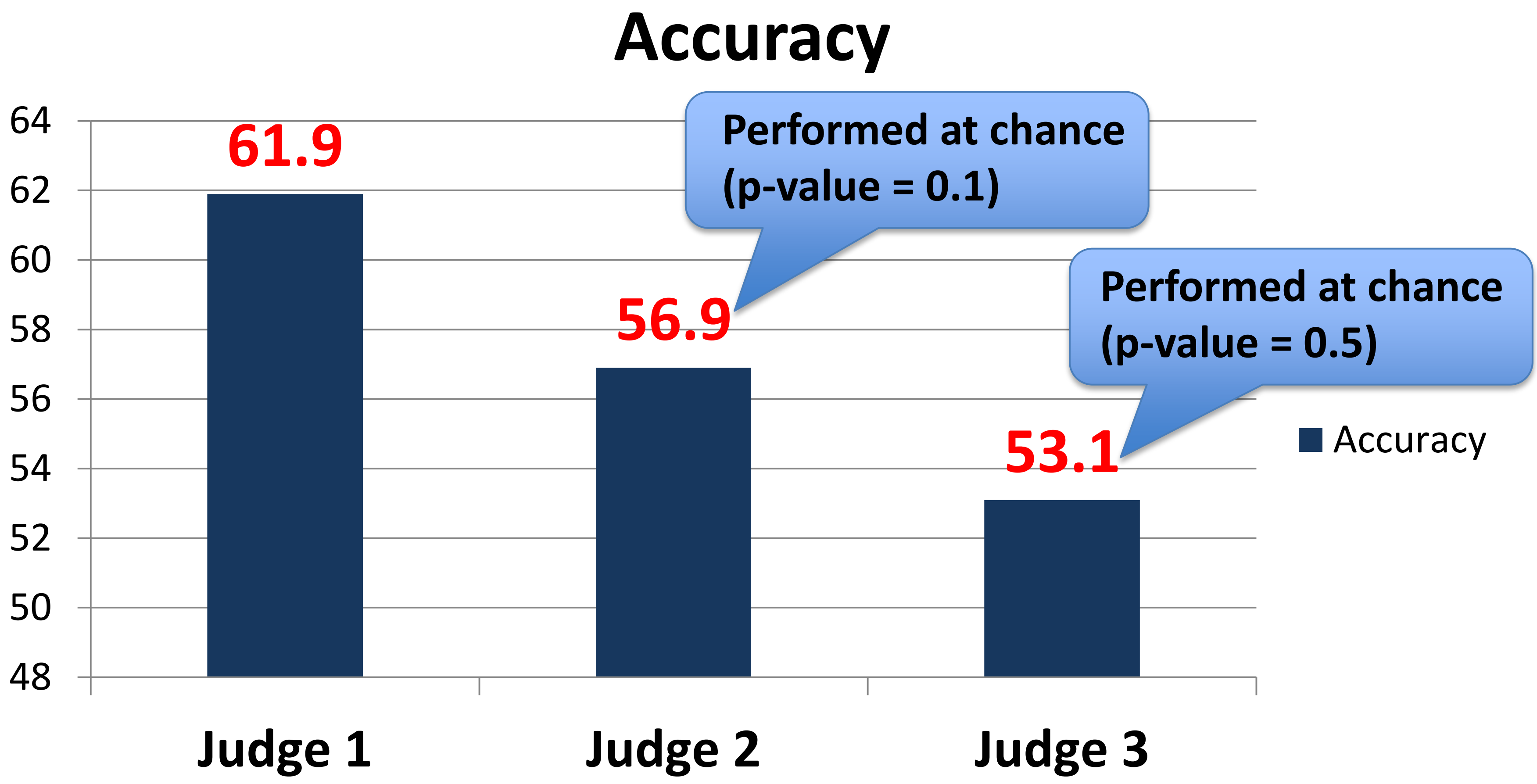
- Why bother?
  - Validates deceptive opinions
  - Baseline to compare other approaches

- 80 truthful and 80 deceptive reviews
- 3 undergraduate judges

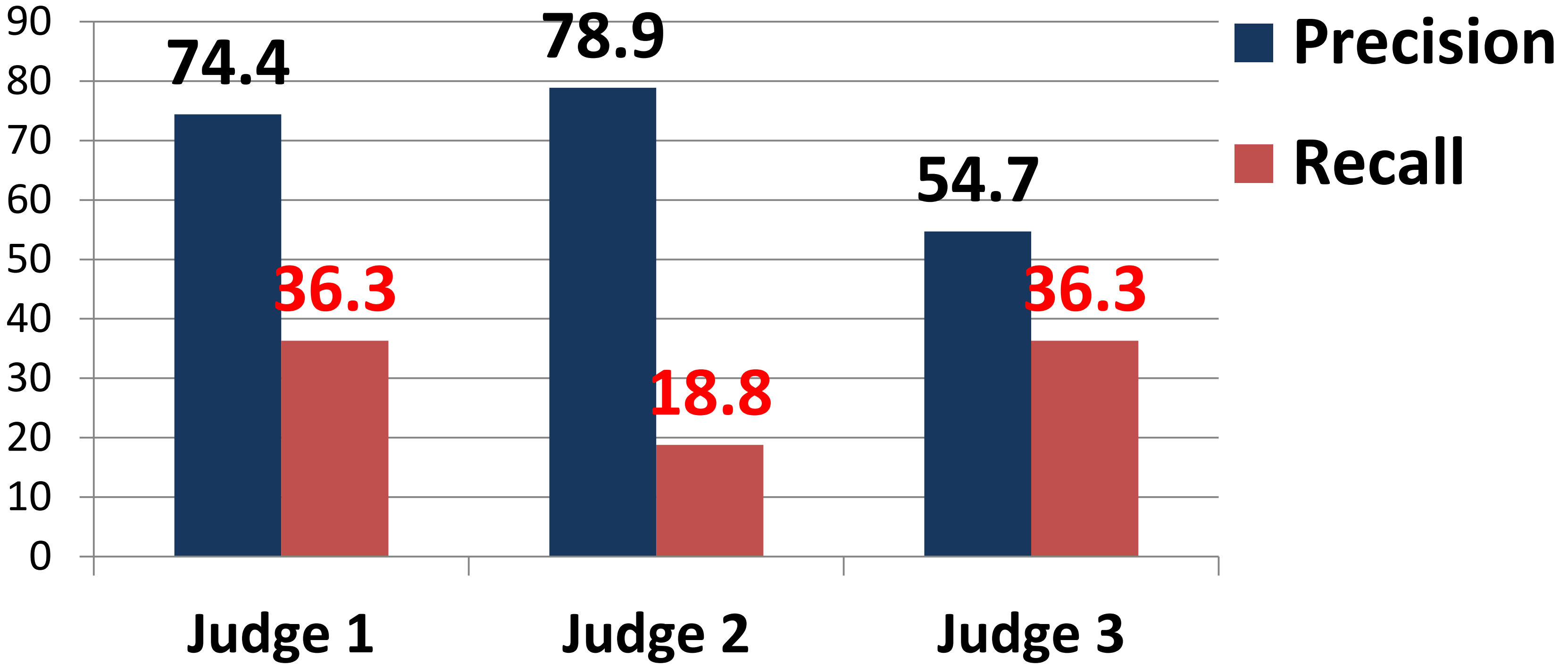
Human Performance



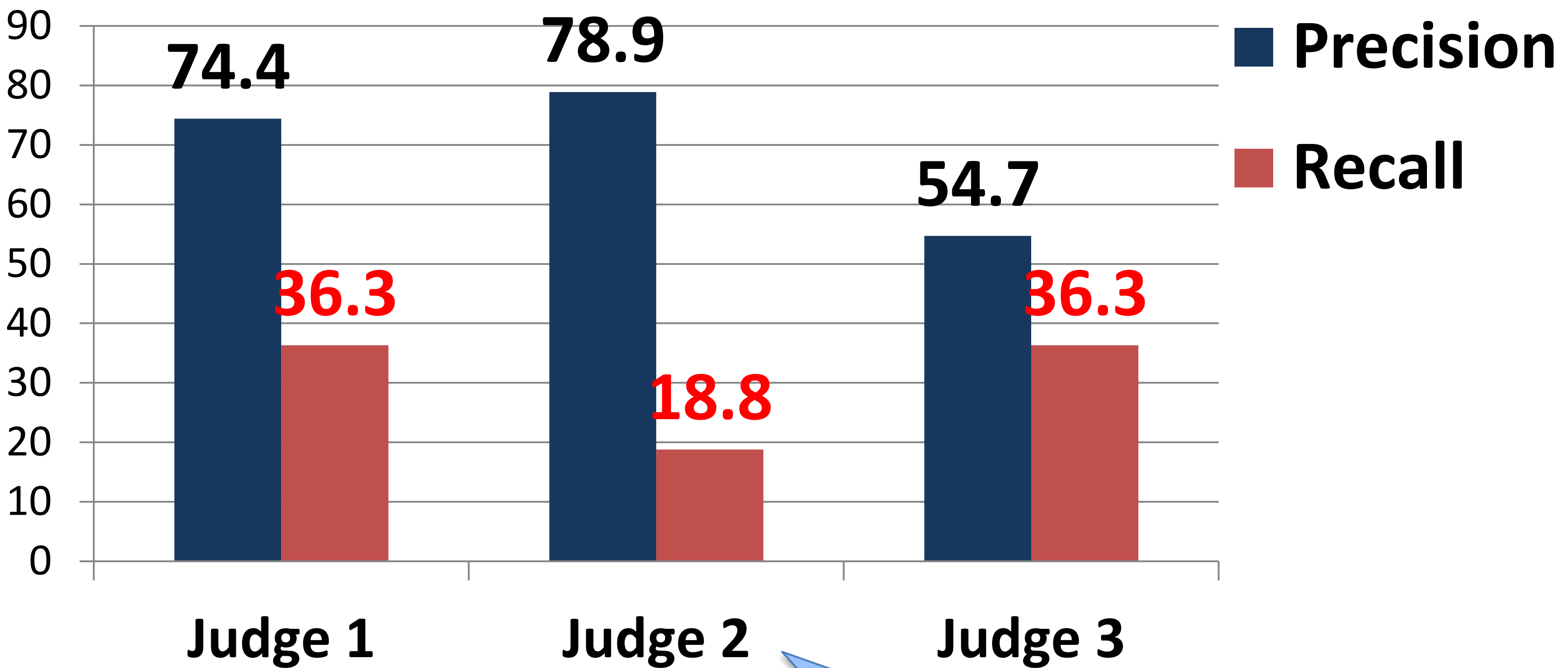
Human Performance



Human Performance



Human Performance



Low Recall → Truth Bias

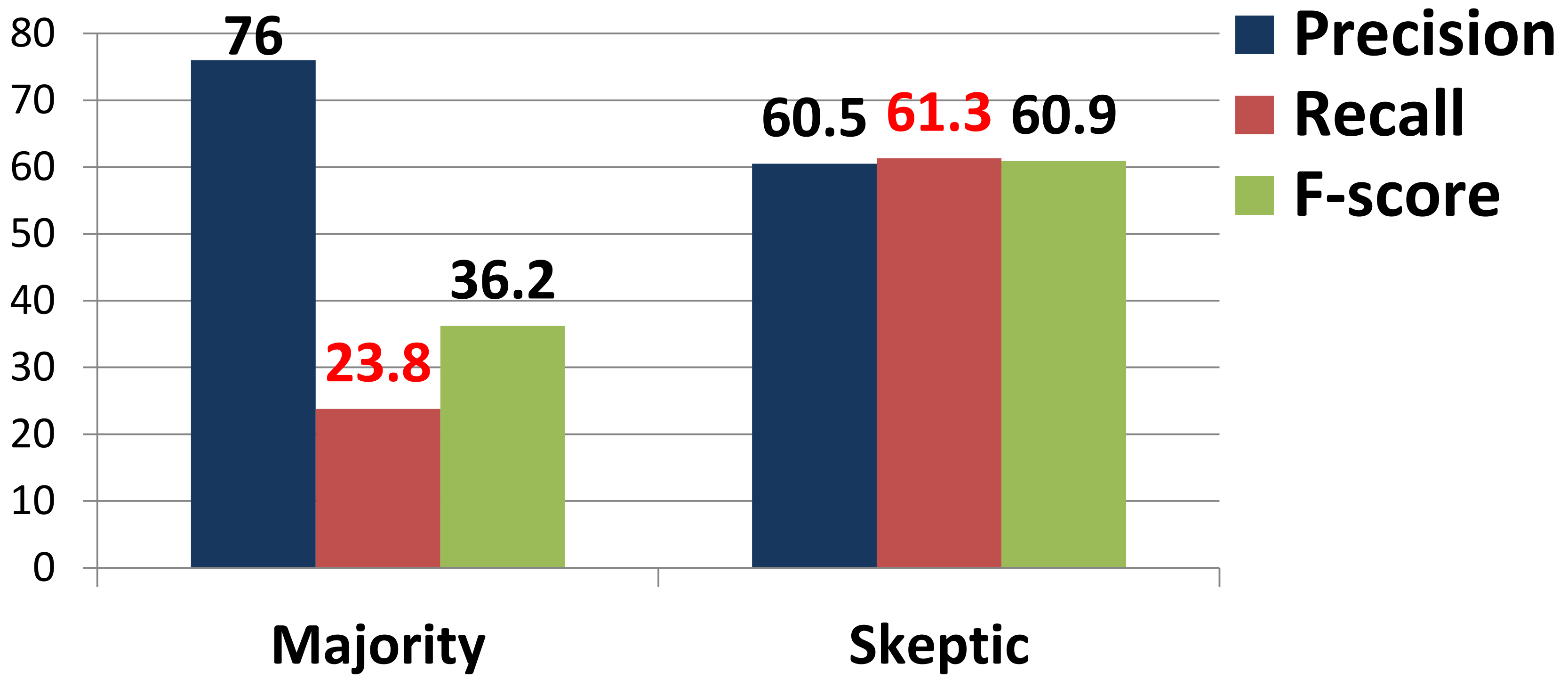
Human Performance

Meta Judges

1. **Majority**
2. **Skeptic**

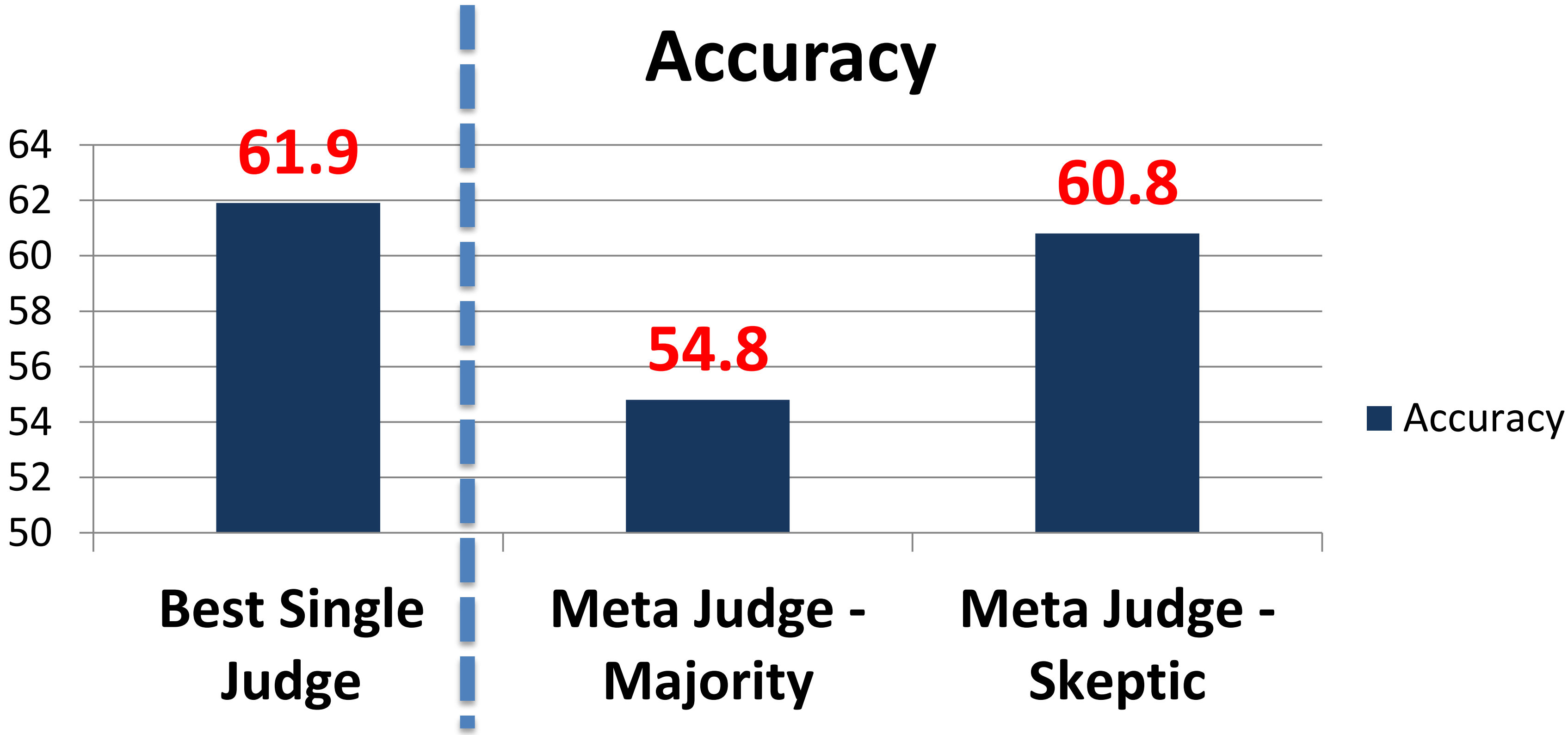
Human Performance

being skeptical helps with recall…



Human Performance

but not the accuracy



Classifier Performance

- Feature sets
  - POS (Part-of-Speech Tags)
  - Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)
  - Unigram, Bigram, Trigram
- Classifiers: SVM & Naïve Bayes

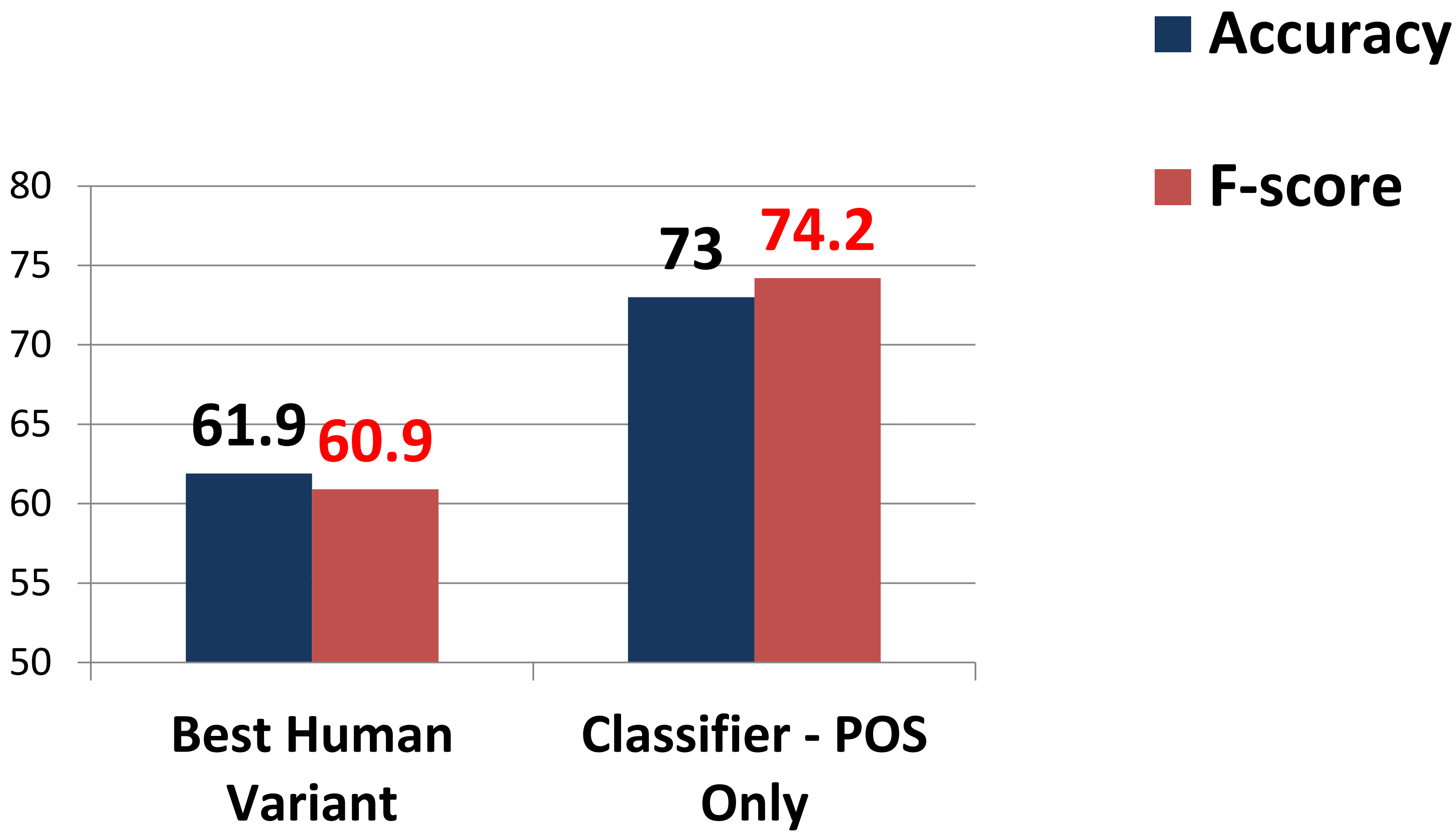
Classifier Performance

- Feature sets
  - **POS (Part-of-Speech Tags)**
  - Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)
  - Unigram, Bigram, Trigram
- Classifiers: SVM & Naïve Bayes

Classifier Performance

- Viewed as ***genre*** identification
  - 48 part-of-speech (POS) features
  - Baseline automated approach
- Expectations
  - Truth similar to ***informative*** writing
  - Deception similar to ***imaginative*** writing

Classifier Performance



TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008	VERBS	Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	0.091		Singular, present	-0.031
ADJECTIVES	General	0.002		Third person singular, present	**0.026**
	Comparative	0.058			
	Superlative	**-0.164**		Modal	-0.063
PREPOSITIONS	General	0.064	ADVERBS	General	**0.001**
DETERMINERS	General	0.009		Comparative	-0.035
COORD. CONJ.	General	0.094	PRONOUNS	Personal	-0.098
VERBS	Past participle	0.053		Possessive	-0.303
ADVERBS	Superlative	**-0.094**	PRE-DETERMINERS	General	**0.017**

Informative writing (left) --- nouns, adjectives, prepositions
Imaginative writing (right) --- verbs, adverbs, pronouns

Rayson et. al. (2001)

TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008		Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	[illegible]		Singular, present	-0.031
ADJECTIVES	General	[illegible]		Third person singular, present	**0.026**
	Comparative	0.058		Modal	-0.063
	Superlative	**-0.164**	ADVERBS	General	**0.001**
PREPOSITIONS	General	0.064		Comparative	-0.035
DETERMINERS	General	[illegible]	[illegible]	Personal	-0.098
COORD. CONJ.	General	[illegible]		Possessive	-0.303
VERBS	Past participle	0.053	PRE-DETERMINERS	General	**0.017**
ADVERBS	Superlative	**-0.094**			

e.g., best, finest

e.g., most

deceptive reviews -- superlatives, exaggerations

TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008	VERBS	Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	0.091		Singular, present	-0.031
ADJECTIVES	General	0.002		Third person singular, present	**0.026**
	Comparative	0.058			
	Superlative	**-0.164**		[illegible]	[illegible]
PREPOSITIONS	General	0.064	ADVERBS	[illegible]	[illegible]
DETERMINERS	General	0.009		Com[illegible]ive	-0.035
COORD. CONJ.	General	0.094	PRONOUNS	Personal	-0.098
VERBS	Past participle	0.053		Possessive	-0.303
ADVERBS	Superlative	**-0.094**	PRE-DETERMINERS	General	**0.017**

e.g., I, my, mine

deceptive reviews -- first person singular pronouns

➔ in contrast to “self-distancing” reported by previous psycholinguistics studies of deception (Newman et al., 2003)

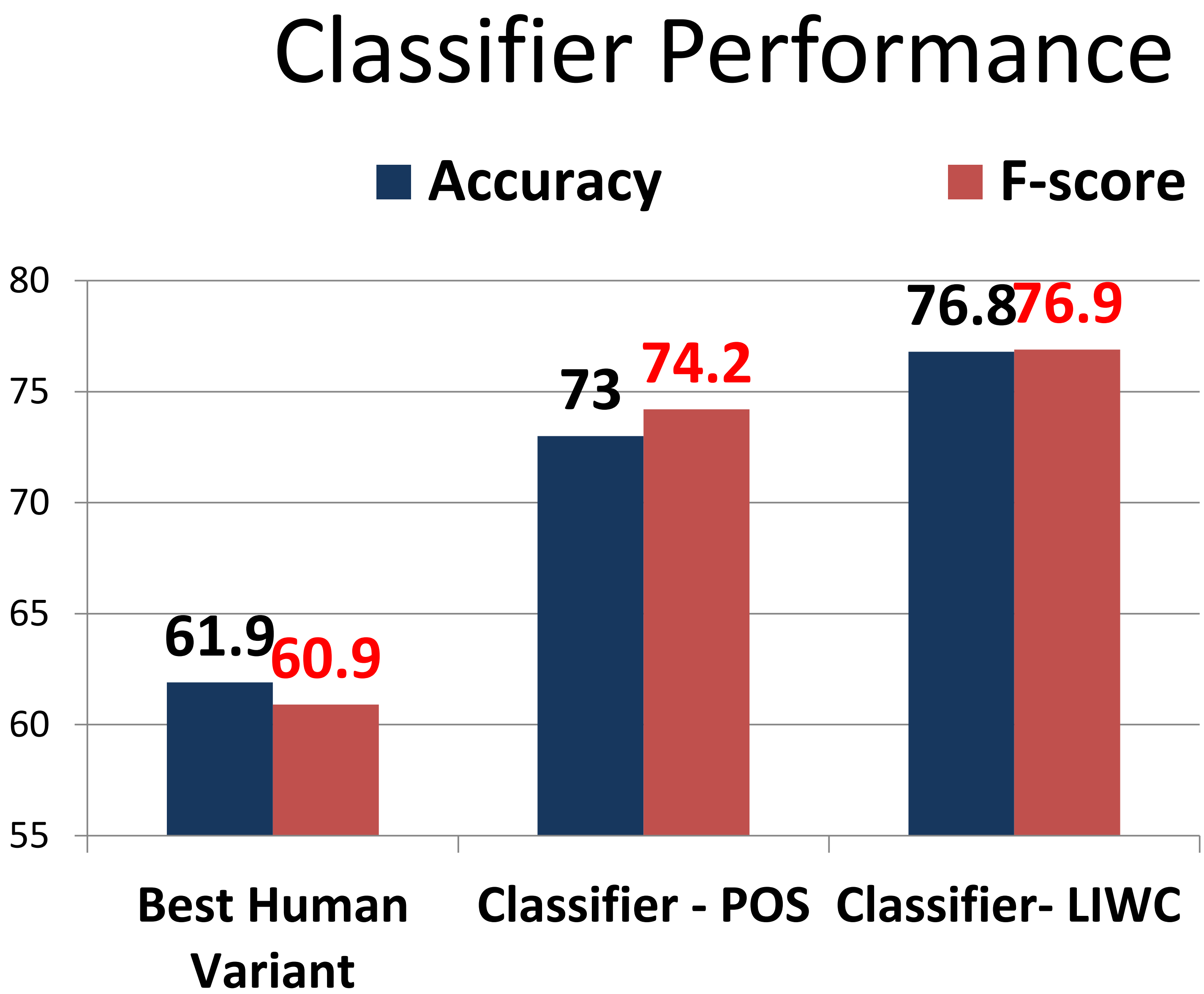
➔ deception cues are domain dependent

Classifier Performance

- Feature sets
  - POS (Part-of-Speech Tags)
  - **Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001, 2007)**
  - Unigram, Bigram, Trigram
- Classifiers: SVM & Naïve Bayes

Classifier Performance

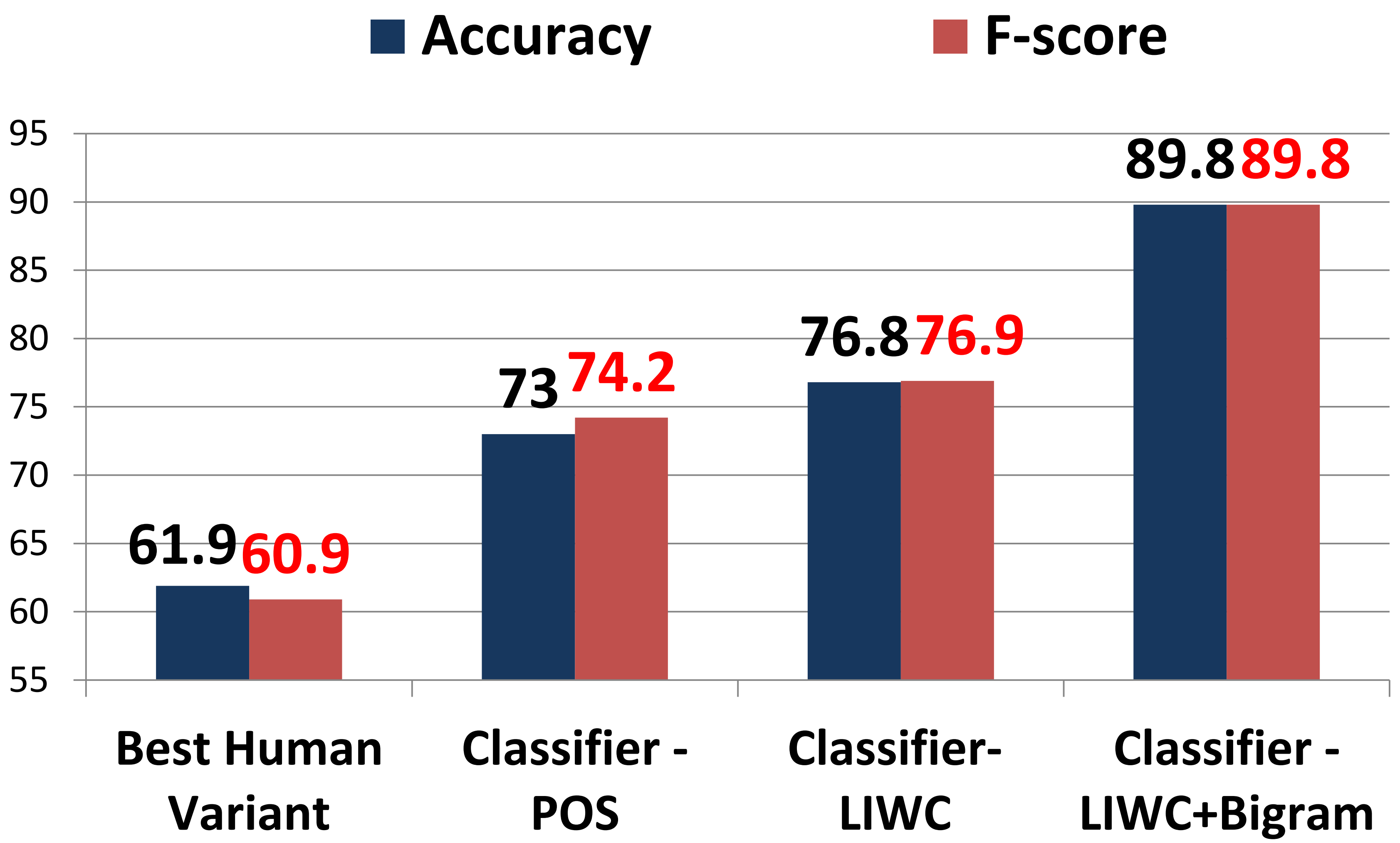
- **L**inguistic **I**nquire and **W**ord **C**ount (LIWC) (Pennebaker et al., 2001, 2007)
  - Widely popular tool for research in social science, psychology, etc
  - Counts instances of ~4,500 keywords
    - Regular expressions, actually
  - Keywords are divided into 80 dimensions across 4 broad groups
    - Linguistic processes, Psychological processes, Personal concerns, Spoken categories


Classifier Performance
Accuracy
F-score
80
75
70
65
60
55
61.9
60.9
73
74.2
76.8
76.9
Best Human Variant
Classifier - POS
Classifier- LIWC

Classifier Performance

- Feature sets
  - POS (Part-of-Speech Tags)
  - Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)
  - **Unigram, Bigram, Trigram**
- Classifiers: SVM & Naïve Bayes

Classifier Performance



Classifier Performance

LIWC+BIGRAMS	
TRUTHFUL	DECEPTIVE
-	chicago
...	my
on	hotel
location	,_and
)	luxury
$\text{allpunct}_{\text{LIWC}}$	experience
floor	hilton
(	business
the_hotel	vacation
bathroom	i
small	spa
helpful	looking
$	while
hotel_.	husband
other	my_husband

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)

Classifier Performance

LIWC+BIGRAMS	
TRUTHFUL	DECEPTIVE
-	chicago
...	my
on	hotel
location	,_and
)	luxury
$\text{allpunct}_{\text{LIWC}}$	experience
floor	hilton
(	business
the_hotel	vacation
bathroom	i
small	spa
helpful	looking
$	while
hotel_.	husband
other	my_husband

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)

Classifier Performance

LIWC+BIGRAMS	
TRUTHFUL	DECEPTIVE
-	chicago
...	my
on	hotel
location	,_and
)	luxury
allpunct$_{\mathrm{LIWC}}$	experience
floor	hilton
(	★ business
the_hotel	★ vacation
bathroom	i
small	spa
helpful	looking
$	while
hotel_.	★ husband
other	★ my_husband

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)

Classifier Performance

LIWC+BIGRAMS	
TRUTHFUL	DECEPTIVE
-	chicago
...	my
on	hotel
location	,_and
)	luxury
allpunct$_{LIWC}$	experience
floor	hilton
(	business
the_hotel	vacation
bathroom	i
small	spa
helpful	looking
$	while
hotel_.	husband
other	my_husband

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)

Media Coverage

- ABC News
- New York Times
- Seattle Times
- Bloomberg / BusinessWeek
- NPR (National Public Radio)
- NHPR (New Hampshire Public Radio)

Conclusion (Case Study I)

- First large-scale gold-standard deception dataset
- Evaluated human deception detection performance
- Developed automated classifiers capable of nearly 90% accuracy
  - Relationship between deceptive and imaginative text
  - Importance of moving beyond universal deception cues

In this talk: three case studies of **stylometric analysis**

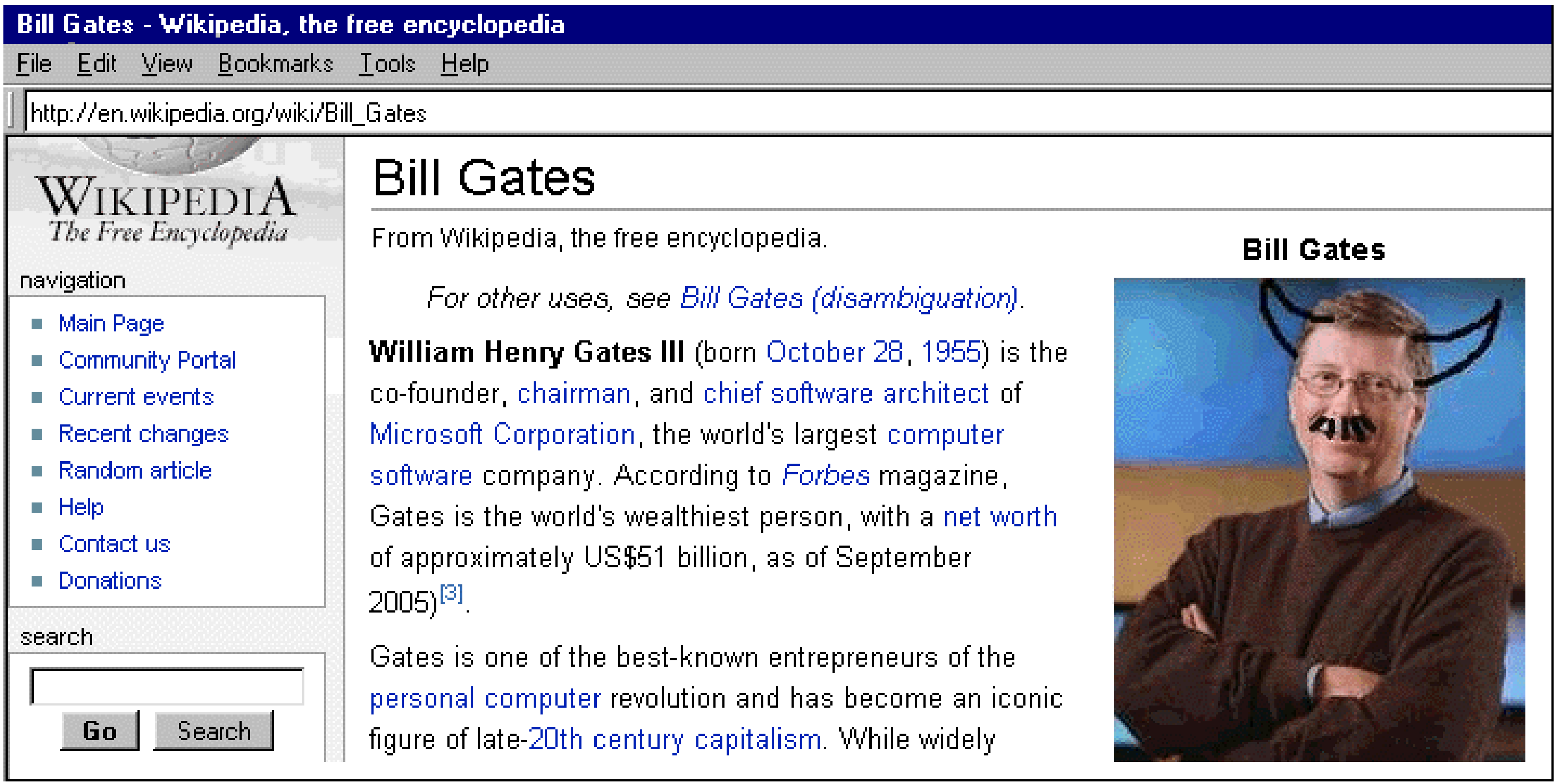
- ✓ Deceptive Product Reviews
- ✓ **Wikipedia Vandalism**
- ✓ The Gender of Authors

Wikipedia

- Community-based knowledge forums (collective intelligence)
- anybody can edit
- susceptible to vandalism --- 7% are vandal edits

- Vandalism **–** ill-intentioned edits to compromise the integrity of Wikipedia.
  - E.g., irrelevant obscenities, humor, or obvious nonsense.

Example of Vandalism



Example of Textual Vandalism

<Edit Title: *Harry Potter*>

- Harry Potter is a teenage boy who likes to smoke crack with his buds. They also run an illegal smuggling business to their headmaster dumbledore. He is dumb!

Example of Textual Vandalism

<Edit Title: *Harry Potter>*

- Harry Potter is a teenage boy who likes to smoke crack with his buds. They also run an illegal smuggling business to their headmaster dumbledore. He is dumb!

<Edit Title: *Global Warming>*

- Another popular theory involving global warming is the concept that global warming is not caused by greenhouse gases. The theory is that Carlos Boozer is the one preventing the infrared heat from escaping the atmosphere. Therefore, the Golden State Warriors will win next season.

Vandalism Detection

- Challenge:
  - Wikipedia covers a wide range of topics (and so does vandalism)
    - vandalism detection based on topic categorization does not work.
  - Some vandalism edits are very tricky to detect

Previous Work I

Most work outside NLP

- Rule-based Robots:
  - e.g., Cluebot (Carter 2007)
- Machine-learning based:
  - features based on hand-picked rules, meta-data, and lexical cues
  - capitalization, misspellings, repetition, compressibility, vulgarism, sentiment, revision size etc

➔ works for easier/obvious vandalism edits, but…

Previous Work II

Some recent work started exploring NLP, but most based on shallow lexico-syntactic patterns

- Wang and McKeown (2010), Chin et al. (2010), Adler et al. (2011)

Vandalism Detection

- Our Hypothesis: textual vandalism constitutes a unique genre where ***a group of people share a similar linguistic behavior***

Wikipedia Manual of Style

Extremely detailed prescription of style:

- **Formatting / Grammar Standards**
  - layout, lists, possessives, acronyms, plurals, punctuations, etc
- **Content Standards**
  - *Neutral point of view*, *No original research* (always include citation), *Verifiability*
  - "What Wikipedia is Not": propaganda, opinion, scandal, promotion, advertising, hoaxes

Example of Textual Vandalism

<**Edit Title**: *Harry Potter*>

- Harry Potter
  crack with h
  smuggling b
  dumbledore. He is dumb!

Long distance dependencies:

- **The theory is that […] is the one […]**
- **Therefore, […] will […]**

<**Edit Title**: *Global Warming*>

- Another popular theory involving global warming is the concept that global warming is not caused by greenhouse gases. The theory is that Carlos Boozer is the one preventing the infrared heat from escaping the atmosphere. Therefore, the Golden State Warriors will win next season.

Language Model Classifier

- Wikipedia Language Model **(P_w)**
  - trained on normal Wikipedia edits
- Vandalism Language Model **(P_v)**
  - trained on vandalism edits
- Given a new edit (x)
  - compute **$P_w(x)$** and **$P_v(x)$**
  - if **$P_w(x) < P_v(x)$,** then edit 'x' is vandalism

Language Model Classifier

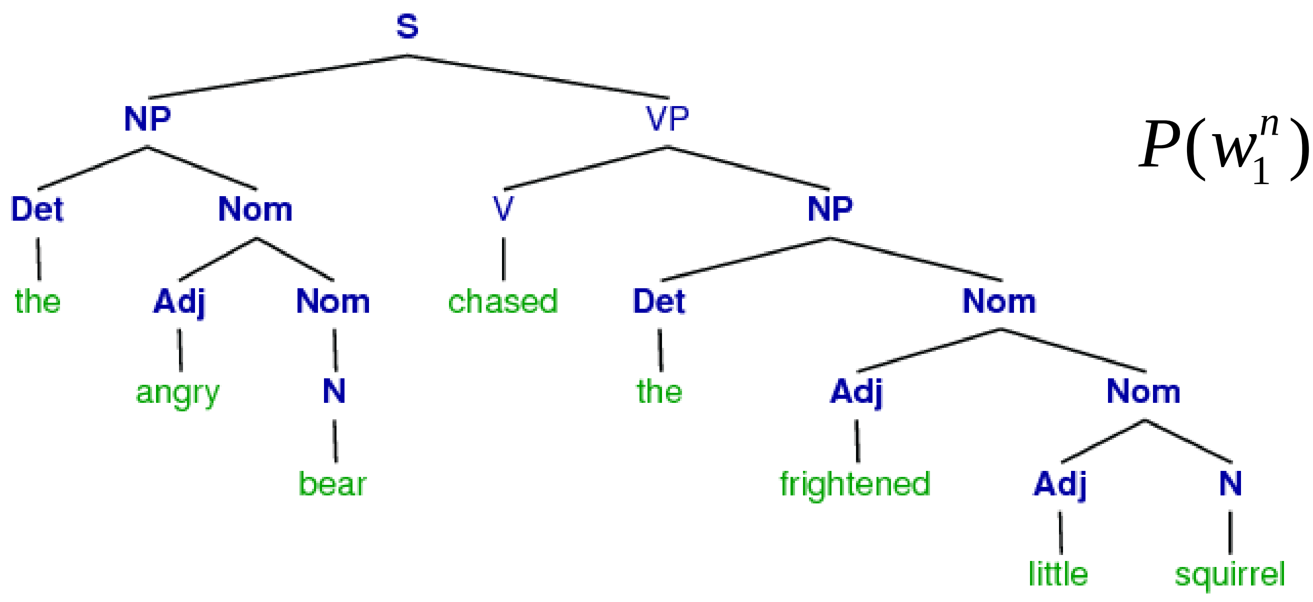
1. N-gram Language Models

   -- most popular choice

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

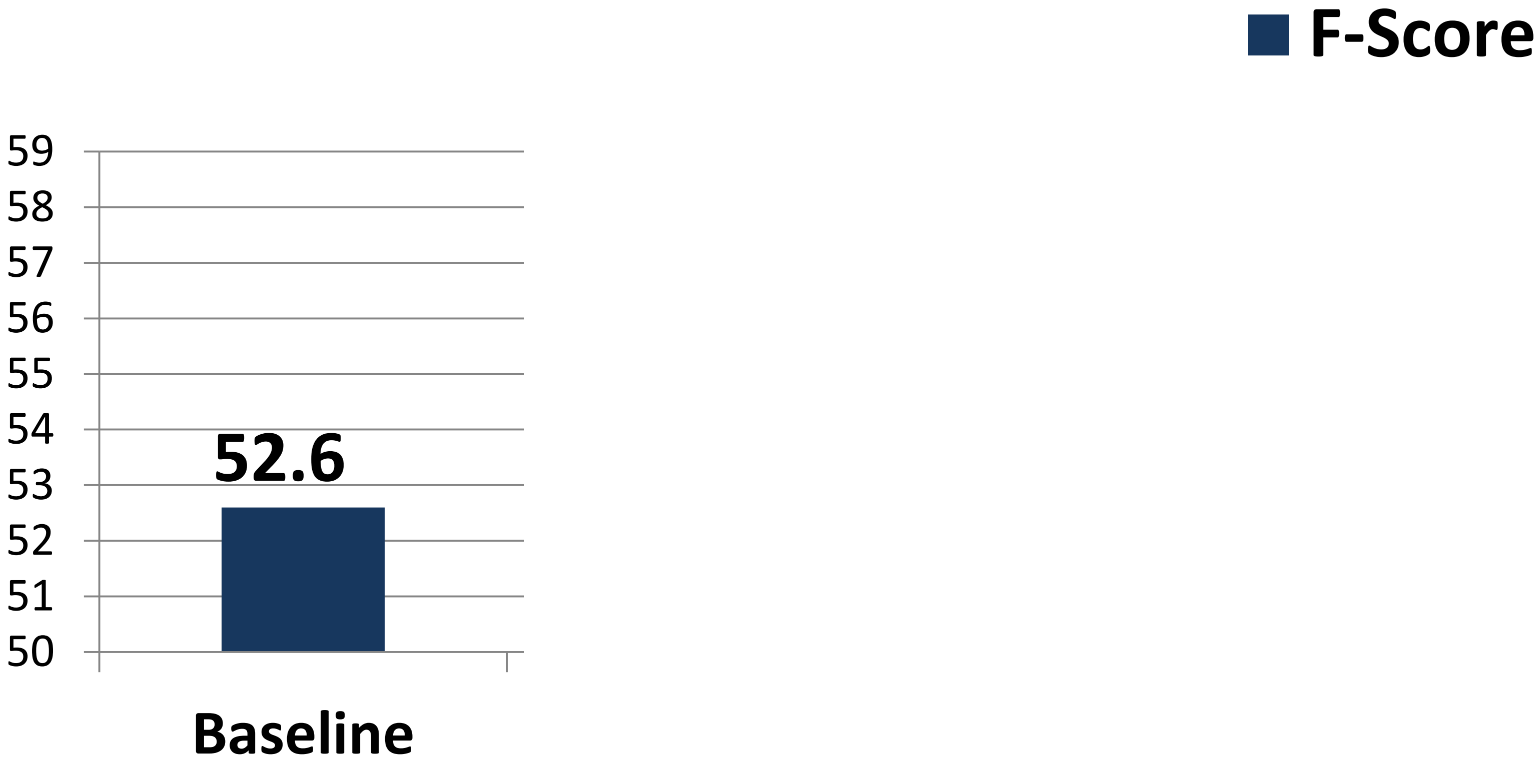
2. PCFG Language Models

   -- Chelba (1997), Raghavan et al. (2010),

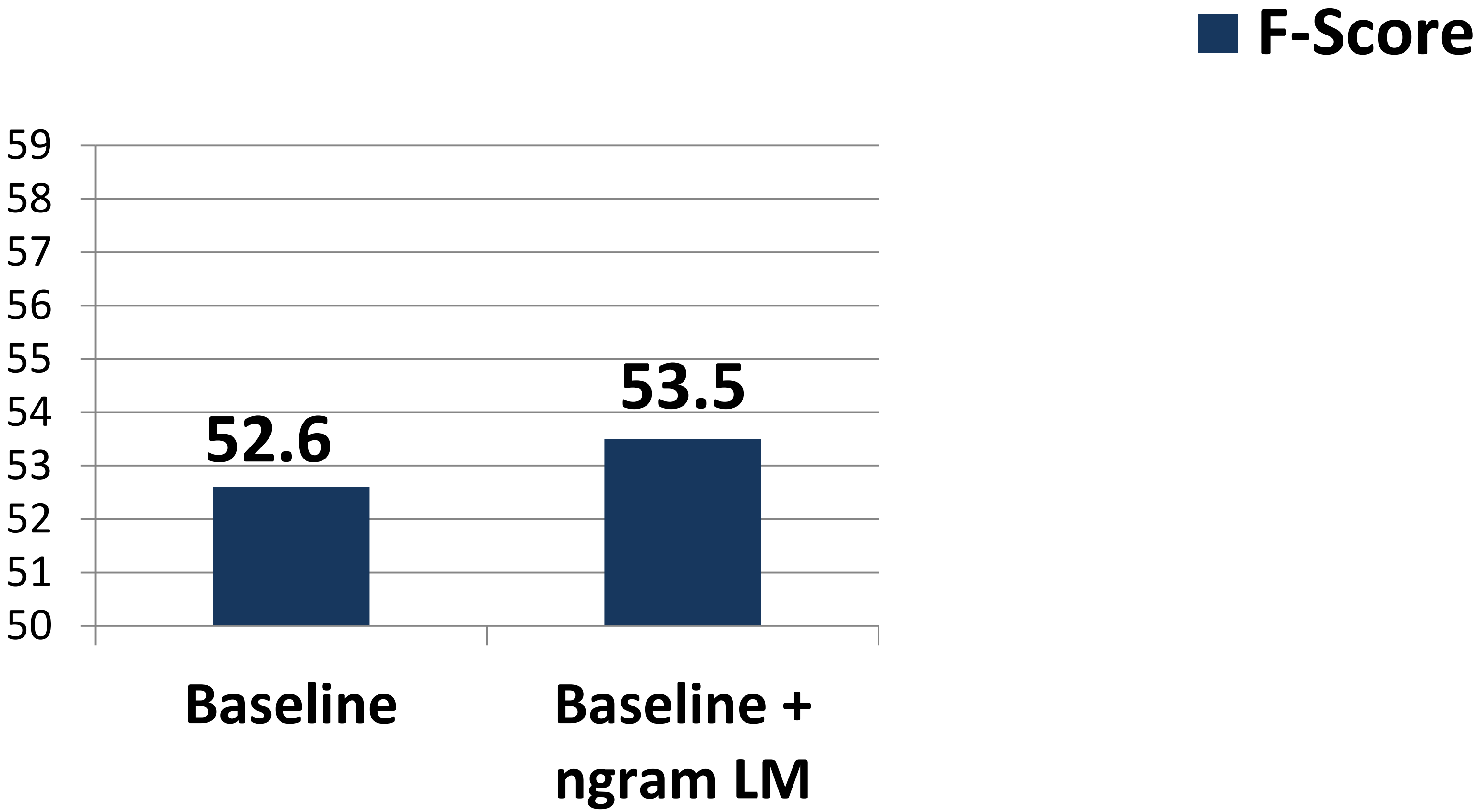


$$P(w_1^n) = \prod P(A \rightarrow \beta)$$

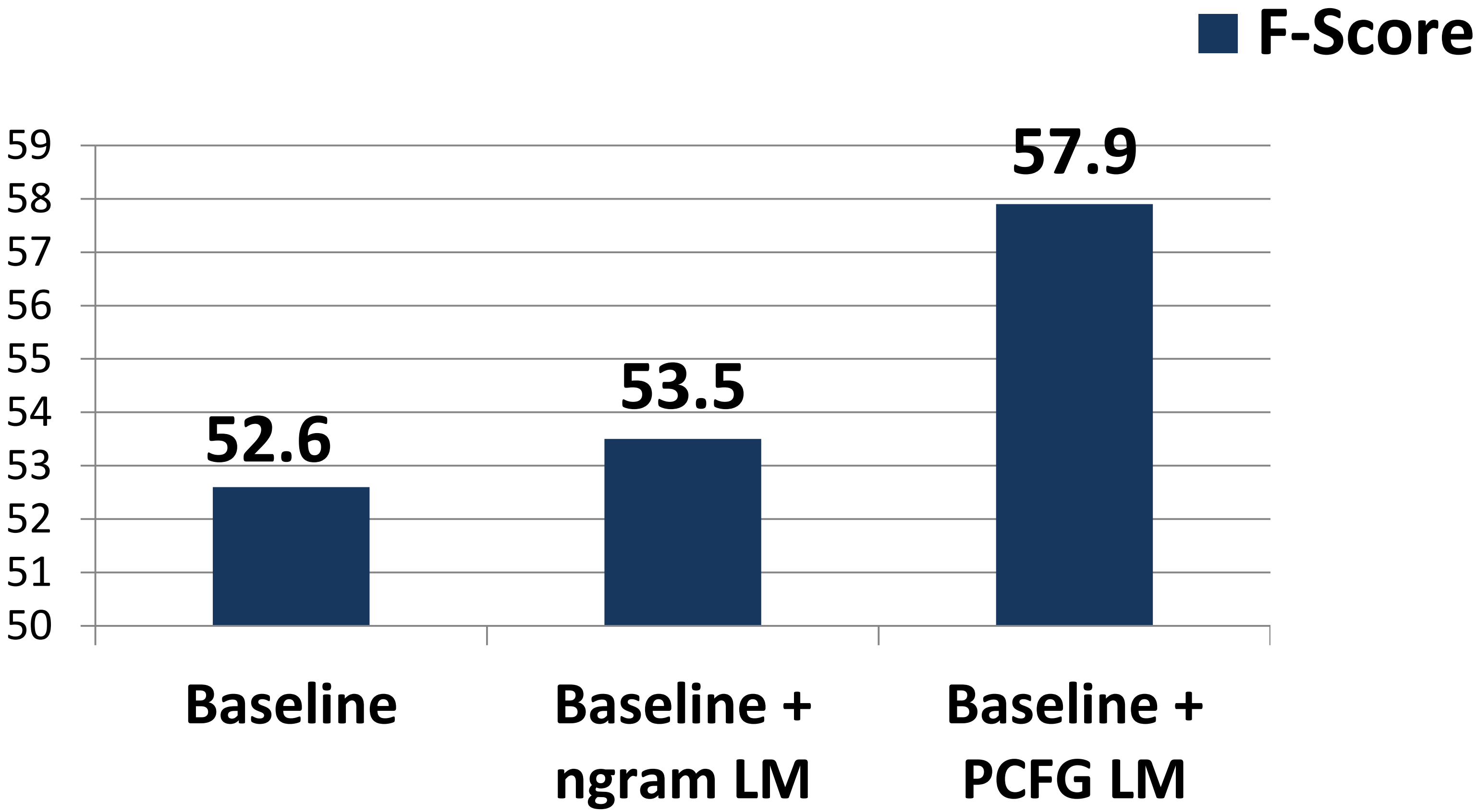
Classifier Performance



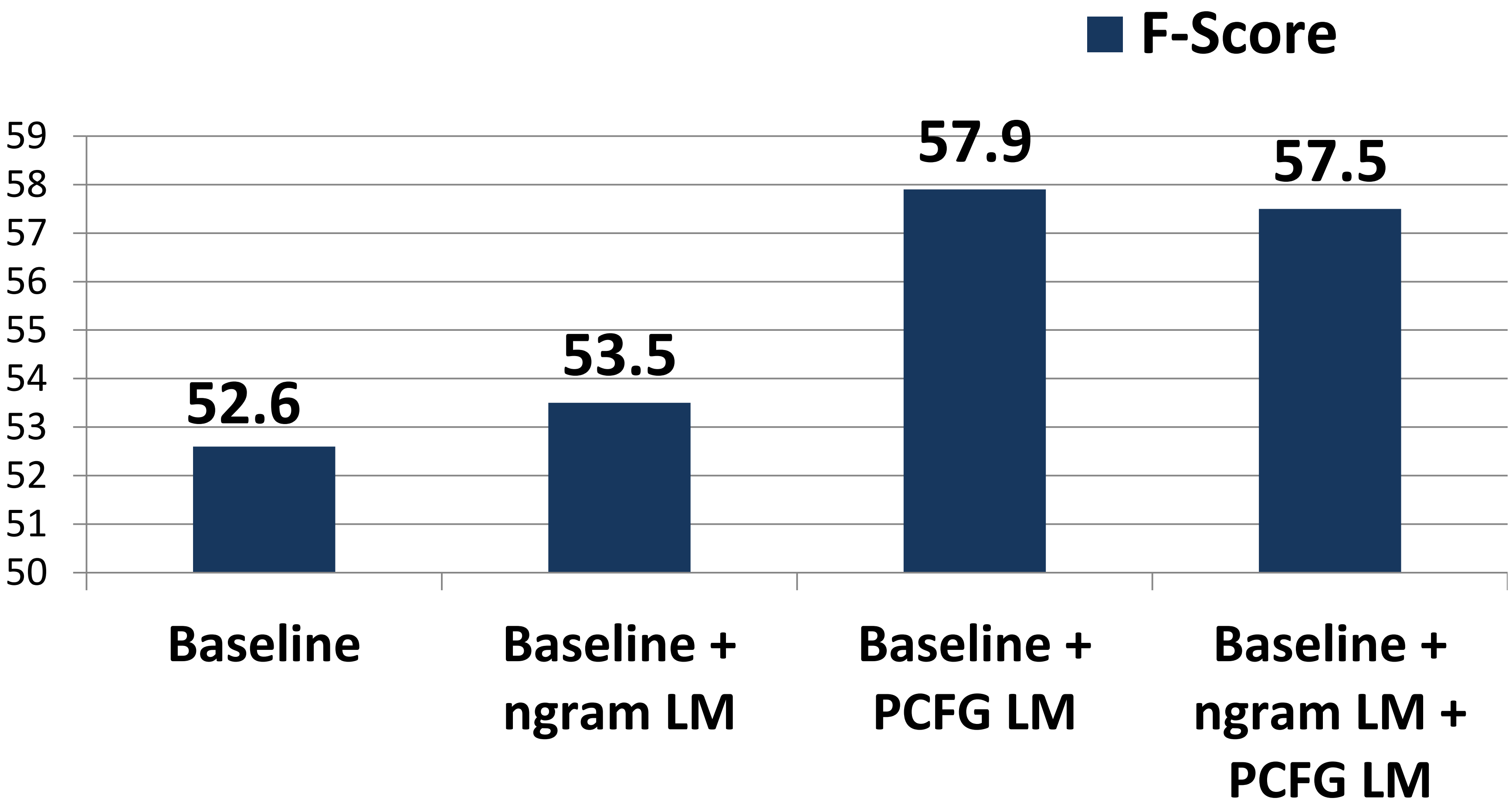
Classifier Performance



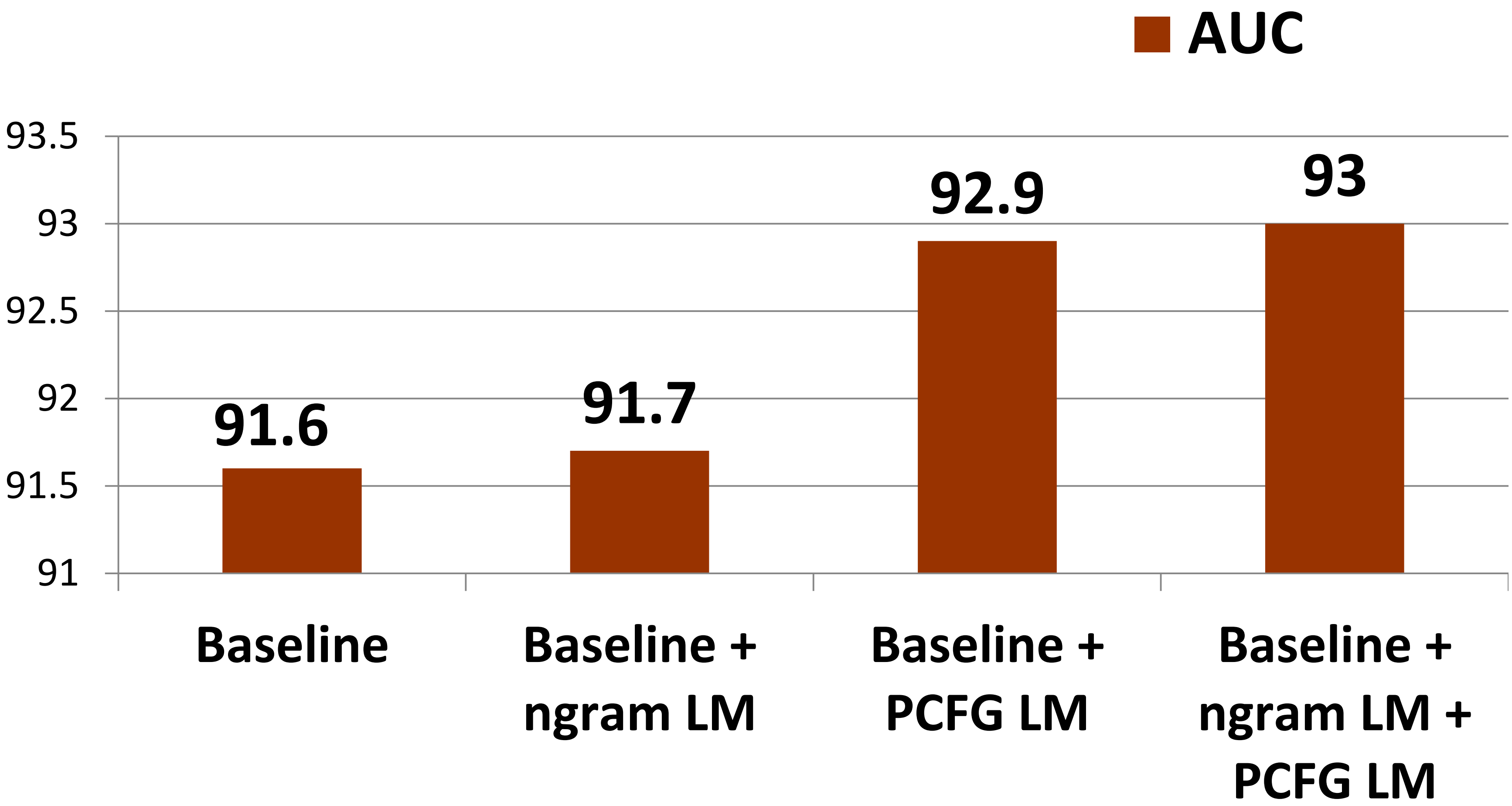
Classifier Performance



Classifier Performance



Classifier Performance



Vandalism Detected by PCFG LM

One day rodrigo was in the school **and** he saw a girl **and** she love her now **and** they are happy together.

Ranking of features

Feature	Score
Total number of author contributions	0.106
How long the author has been registered	0.098
How frequently the author contributed in the training set	0.097
If the author is registered	0.0885
Difference in the maximum PCFG scores	0.0437
Difference in the mean PCFG scores	0.0377
How many times the article has been reverted	0.0372
Total contributions of author to Wikipedia	0.0343
Previous vandalism count of the article	0.0325
Difference in the sum of PCFG scores	0.0320

Conclusion (Case Study II)

- There are unique language styles in vandalism, and stylometric analysis can improve automatic vandalism detection.
- Deep syntactic patterns based on PCFGs can identify vandalism more effectively than shallow lexico-syntactic patterns based on n-gram language models

In this talk: three case studies of **stylometric analysis**

- ✓ Deceptive Product Reviews
- ✓ Wikipedia Vandalism
- ✓ **THE GENDER OF AUTHORS**

“Against Nostalgia”

Excerpt from NY Times OP-ED, Oct 6, 2011

“STEVE JOBS was an enemy of nostalgia. (……) One of the keys to Apple’s success under his leadership was his ability to see technology with an unsentimental eye and keen scalpel, ready to cut loose whatever might not be essential. This editorial mien was Mr. Jobs’s greatest gift — he created a sense of style in computing because he could edit.”

“My Muse Was an Apple Computer”

Excerpt from NY Times OP-ED, Oct 7, 2011

“More important, you worked with that little blinking cursor before you. No one in the world particularly cared if you wrote and, of course, you knew the computer didn’t care, either. But it was waiting for you to type something. It was not inert and passive, like the page. It was listening. It was your ally. It was your audience.”

The New York Times

"My Muse Was an Apple Computer"

Excerpt from NY Times OP-ED, Oct 7, 2011

"More important, you worked with that little blinking cursor before you. No one in the world

f you wrote and, of course,

uter didn't care, either. But it

u to type something. It was

ve, like the page. It was

Gish Jen

a novelist

The New York Times

"Against Nostalgia"

Excerpt from NY Times OP-ED, Oct 6, 2011

"STEVE JOBS was an enemy of nostalgia. (……)
Apple's success under his
ability to see technology with
eye and keen scalpel, ready to
might not be essential. This
Mr. Jobs's greatest gift — he

Mike Daisey

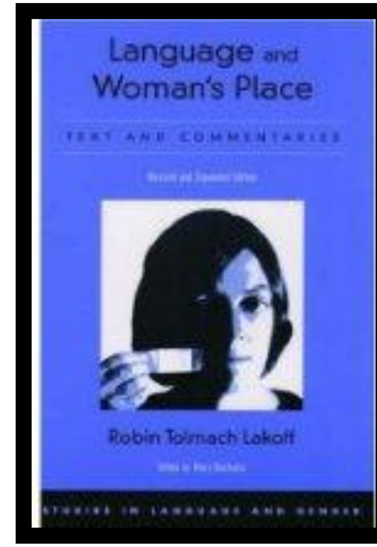
an author and performer

Motivations

Demographic characteristics of user-created web text

- New insight on social media analysis
- Tracking gender-specific styles in language over different domain and time
- Gender-specific opinion mining
- Gender-specific intelligence marketing

Women’s Language



Robin Lakoff(1973)

1. Hedges: “kind of”, “it seems to be”, etc.
2. Empty adjectives: “lovely”, “adorable”, “gorgeous”, etc.
3. Hyper-polite: “would you mind ...”, “I’d much appreciate if ...”
4. Apologetic: “I am very sorry, but I think...”
5. Tag questions: “you don’t mind, do you?”

…

Related Work

Sociolinguistic and Psychology

- Lakoff(1972, 1973, 1975)
- Crosby and Nyquist (1977)
- Tannen (1991)
- Coates, Jennifer (1993)
- Holmes (1998)
- Eckert and McConnell-Ginet (2003)
- Argamon et al. (2003, 2007)
- McHugh and Hambaugh (2010)

Related Work

Machine Learning

- Koppel et al. (2002)
- Mukherjee and Liu (2010)

Concerns: Gender Bias in Topics

"Considerable gender bias in topics and genres"

– Janssen and Murachver (2004)

– Herring and Paolillo (2006)

– Argamon et al. (2007)

We want to ask…

- Are there indeed gender-specific styles in language?

- If so, what kind of statistical patterns discriminate the gender of the author?
  - *morphological patterns*
  - *shallow-syntactic patterns*
  - *deep-syntactic patterns*

We want to ask…

- Can we trace gender-specific styles beyond topics and genres?
  - train in one domain and test in another

We want to ask…

- Can we trace gender-specific styles beyond topics and genres?
  – train in one domain and test in another
  – what about ***scientific papers***?

Gender specific language styles are not conspicuous in formal writing.
Janssen and Murachver (2004)

Dataset

Balanced topics to avoid gender bias in topics

- ❖ Blog Dataset

  -- informal language

- ❖ Scientific Dataset

  -- formal language

Dataset

Balanced topics to avoid gender bias in topics

- Blog Dataset
  - informal language
  - 7 topics – education, entertainment, history, politics, etc.
  - 20 documents per topic and per gender
  - first 450 (+/- 20) words from each blog

Dataset

Balanced topics to avoid gender bias in topics

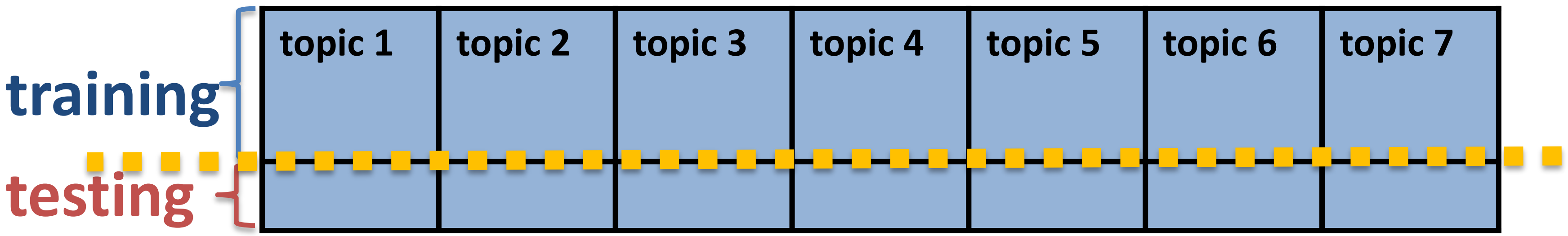
- Scientific Dataset
  - formal language
  - 5 female authors, 5 male authors
  - include multiple subtopics in NLP
  - 20 papers per author
  - first 450 (+/- 20) words from each paper

Plan for the Experiments

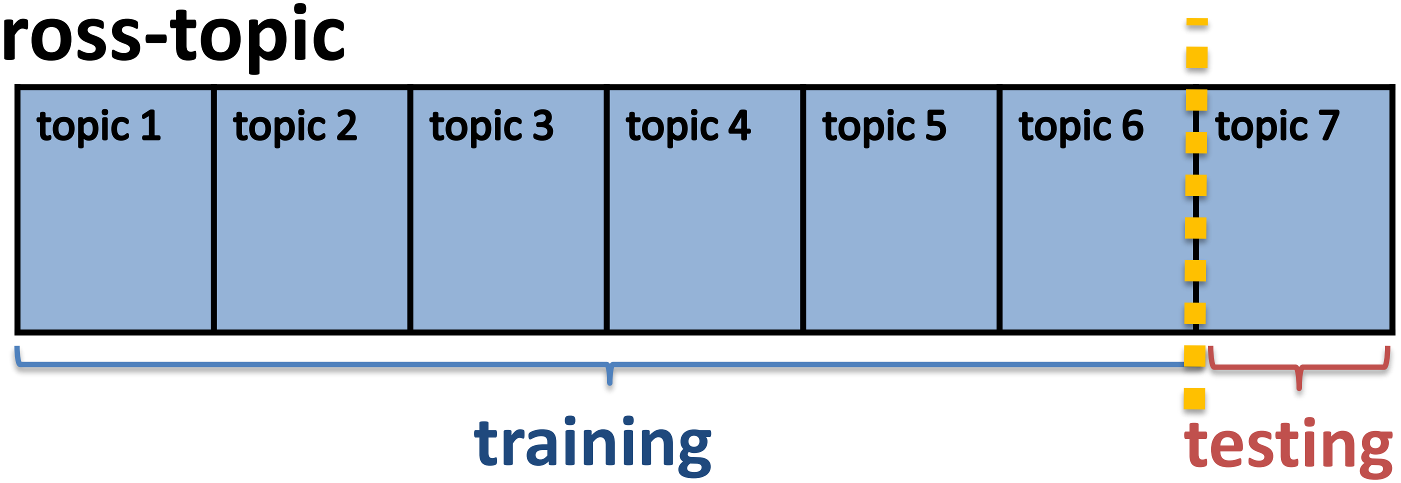
- Blog dataset
  1. balanced-topic
  2. cross-topic

Balanced-Topic / Cross-Topic

I. balanced-topic



II. cross-topic



Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic

Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Language Model Classifier

- Wikipedia Language Model **(P_w)**
  - trained on normal Wikipedia edits
- Vandalism Language Model **(P_v)**
  - trained on vandalism edits
- Given a new edit (x)
  - compute **$P_w(x)$** and **$P_v(x)$**
  - if **$P_w(x) < P_v(x)$,** then edit 'x' is vandalism

Language Model Classifier

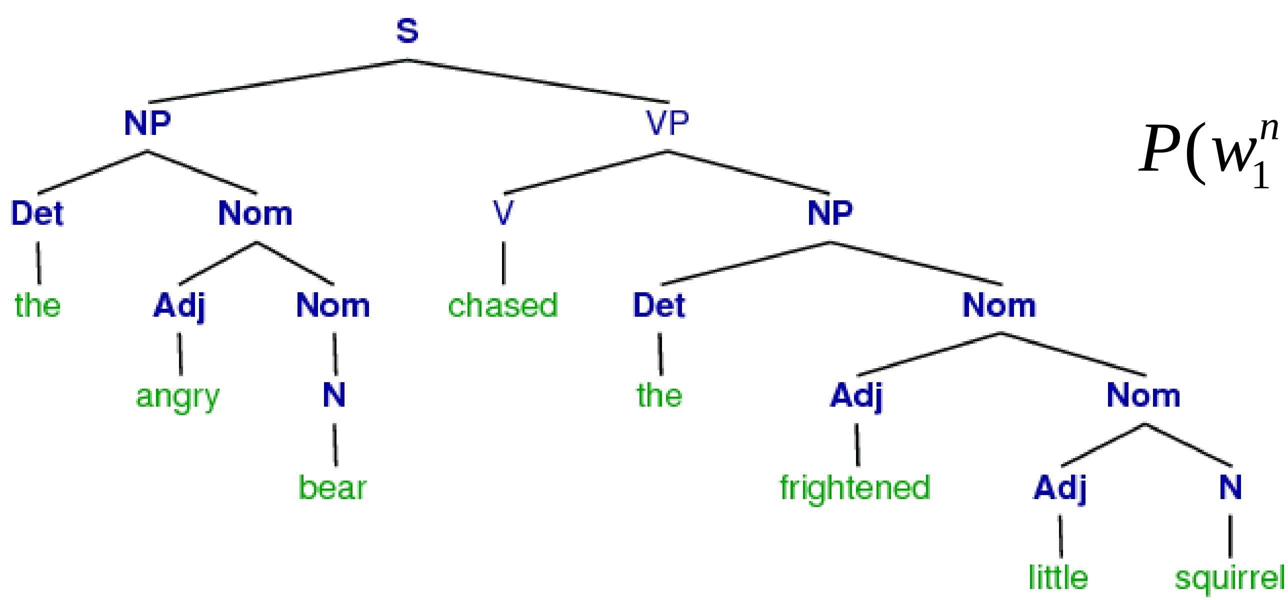
1. N-gram Language Models

   -- most popular choice

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

2. PCFG Language Models

   -- Chelba (1997), Raghavan et al. (2010),



$$P(w_1^n) = \prod P(A \rightarrow \beta)$$

Statistical Stylometric Analysis

1. Shallow Morphological Patterns
   ➔ Character-level Language Models (**Char-LM**)
2. Shallow Lexico-Syntactic Patterns
   ➔ Token-level Language Models (**Token-LM**)
3. Deep Syntactic Patterns
   ➔Probabilistic Context Free Grammar (**PCFG**)
   – Chelba (1997), Raghavan et al. (2010),

Baseline

1. Gender Genie:

   http://bookblog.net/gender/genie.php

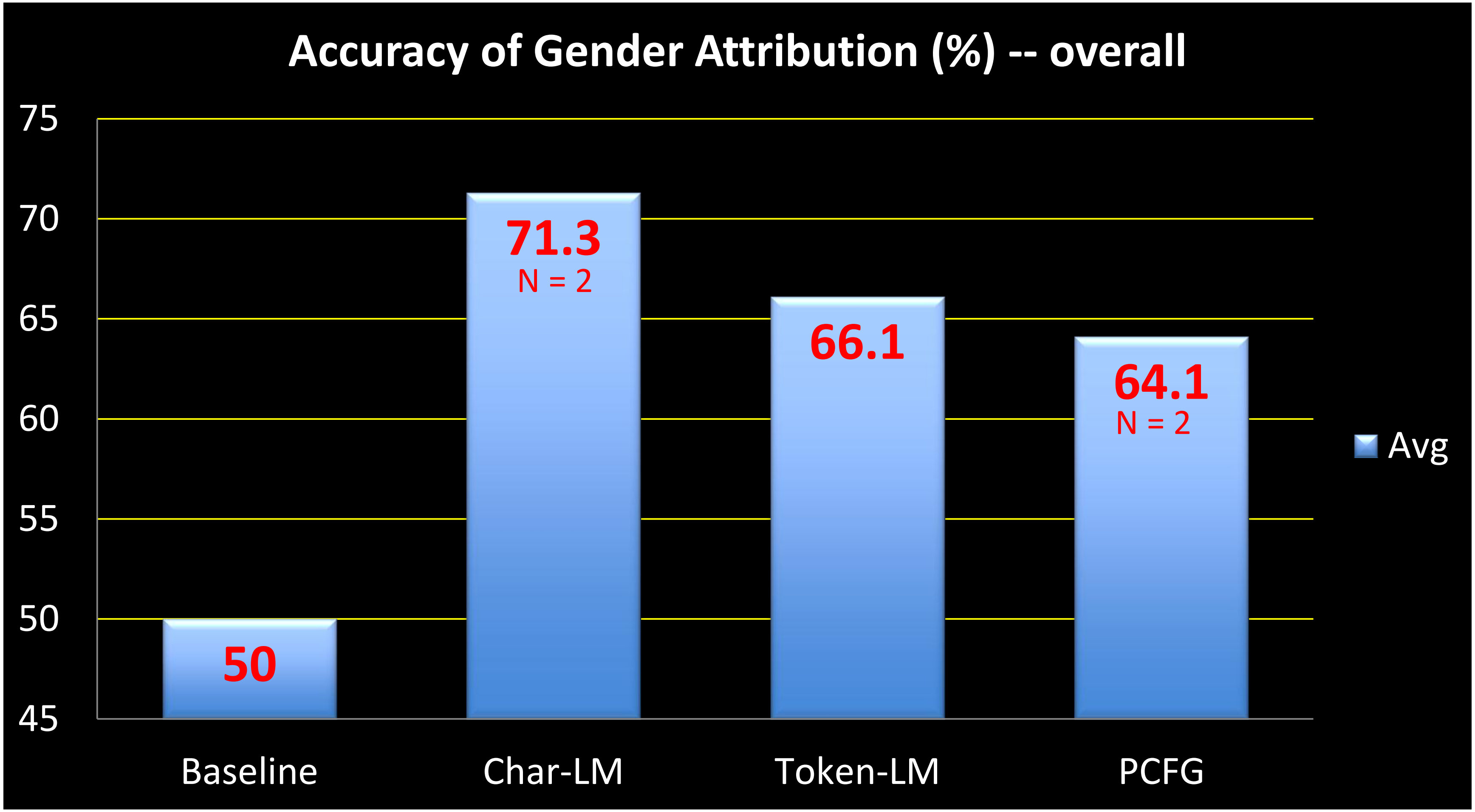
2. Gender Guesser

   http://www.genderguesser.com/

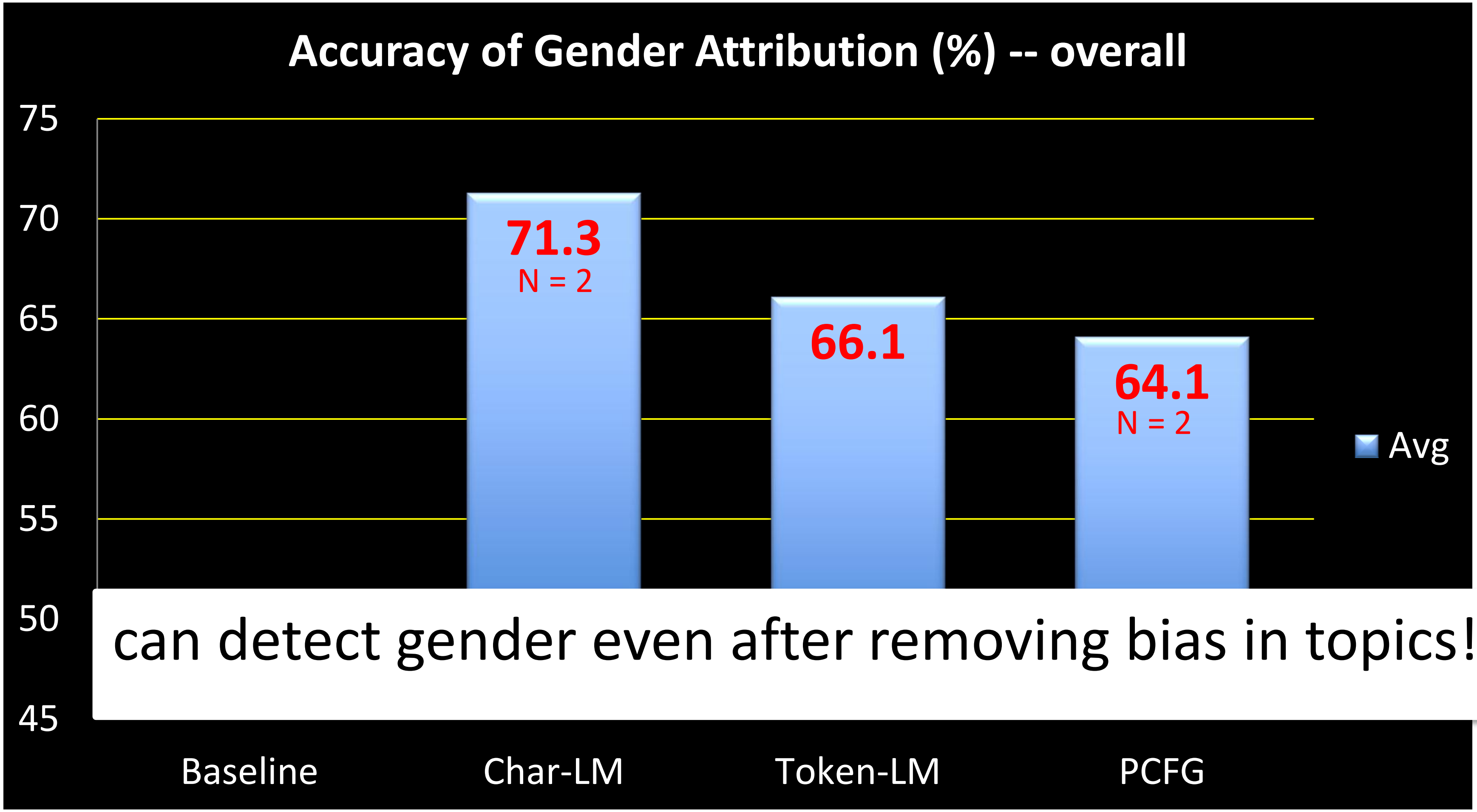
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment I: balanced-topic, blog



Experiment I:
balanced-topic, blog

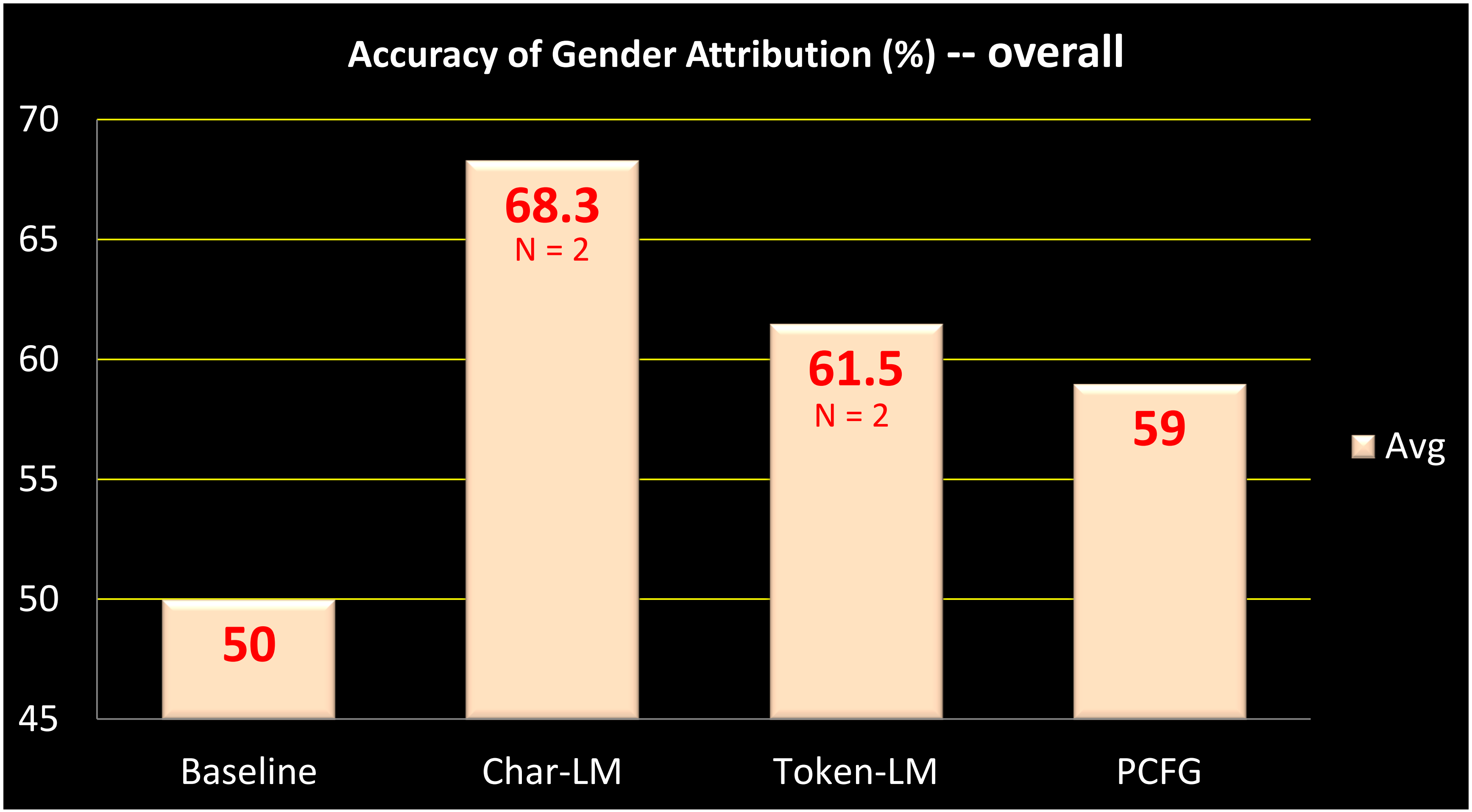


can detect gender even after removing bias in topics!

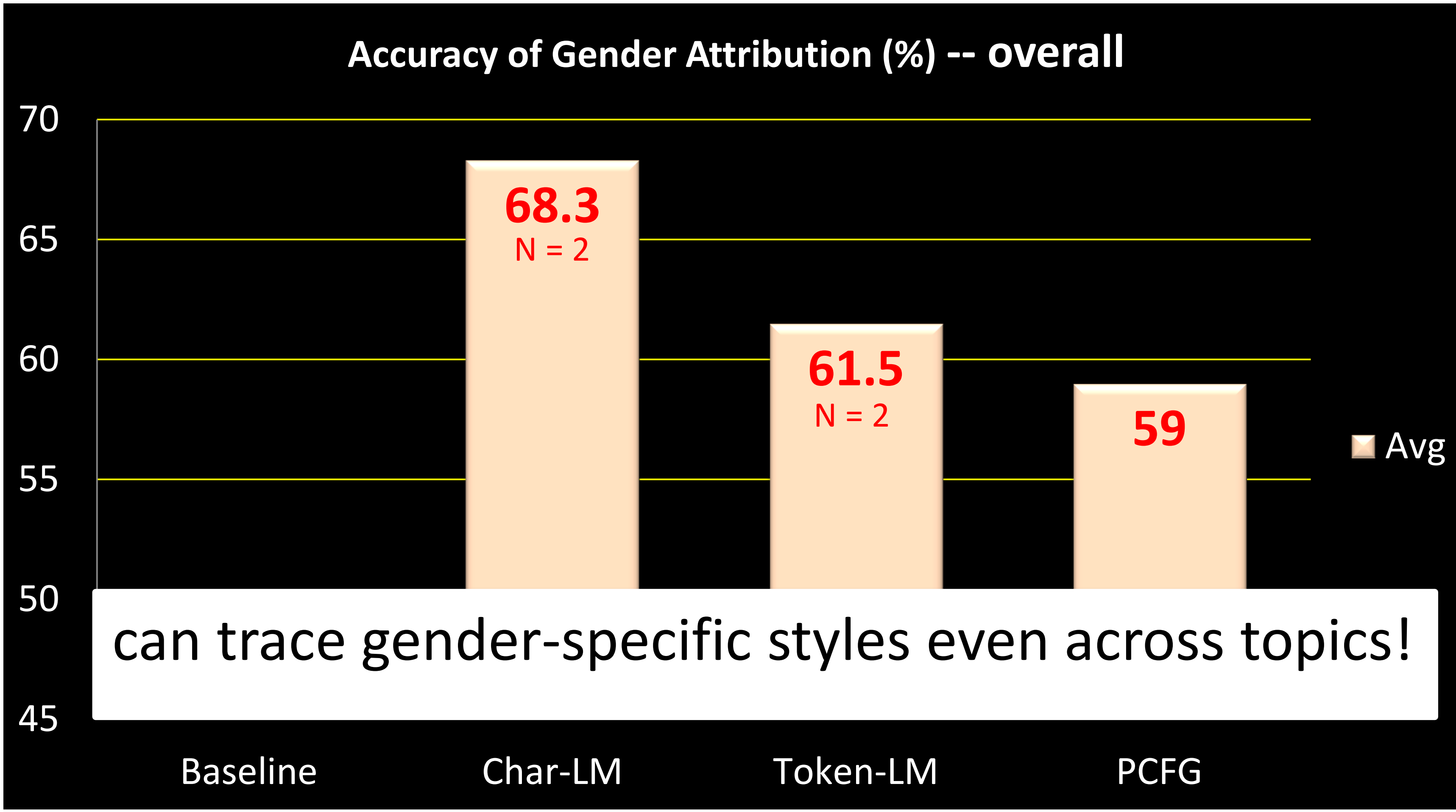
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment II: cross-topic, blog



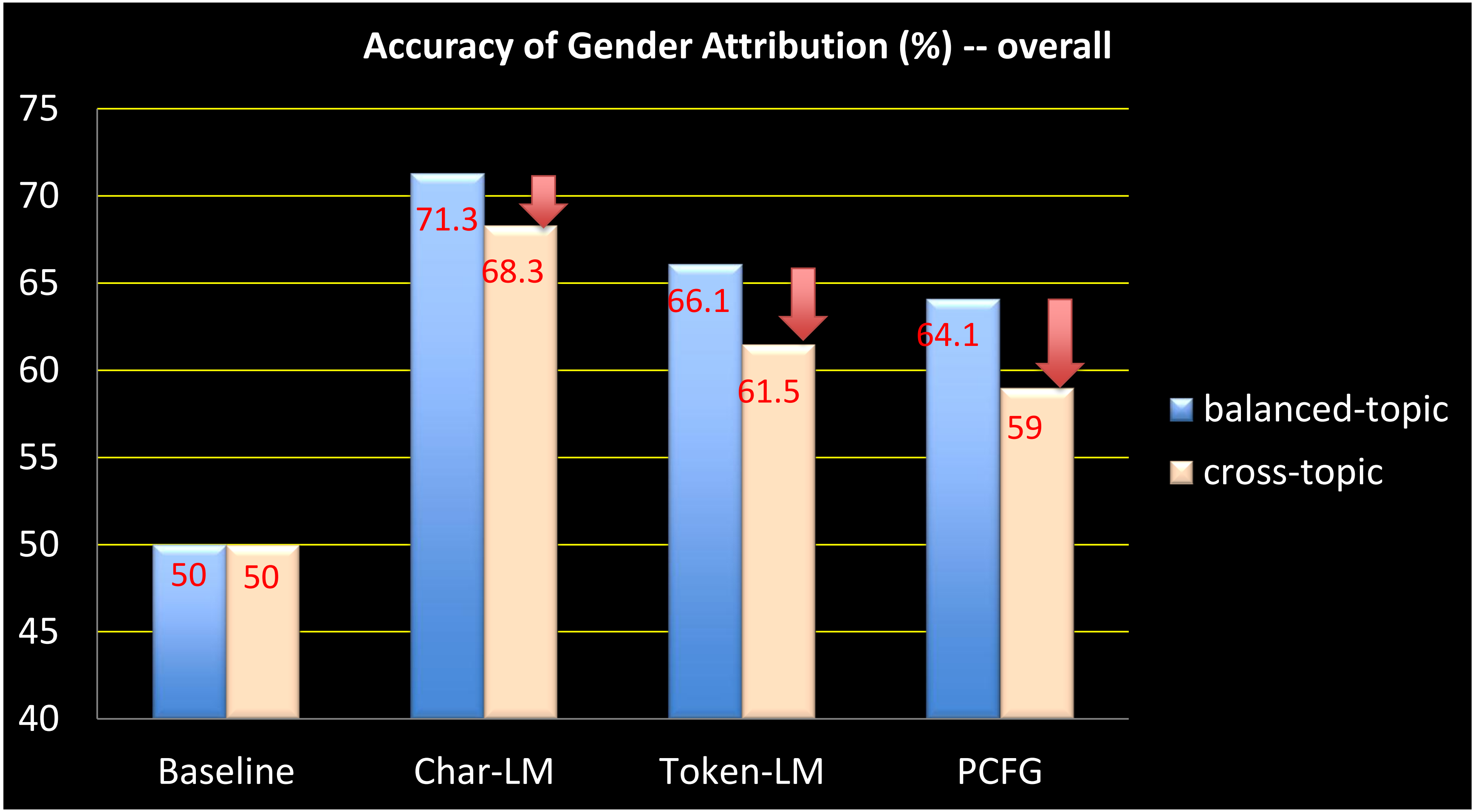
Experiment II:
cross-topic, blog



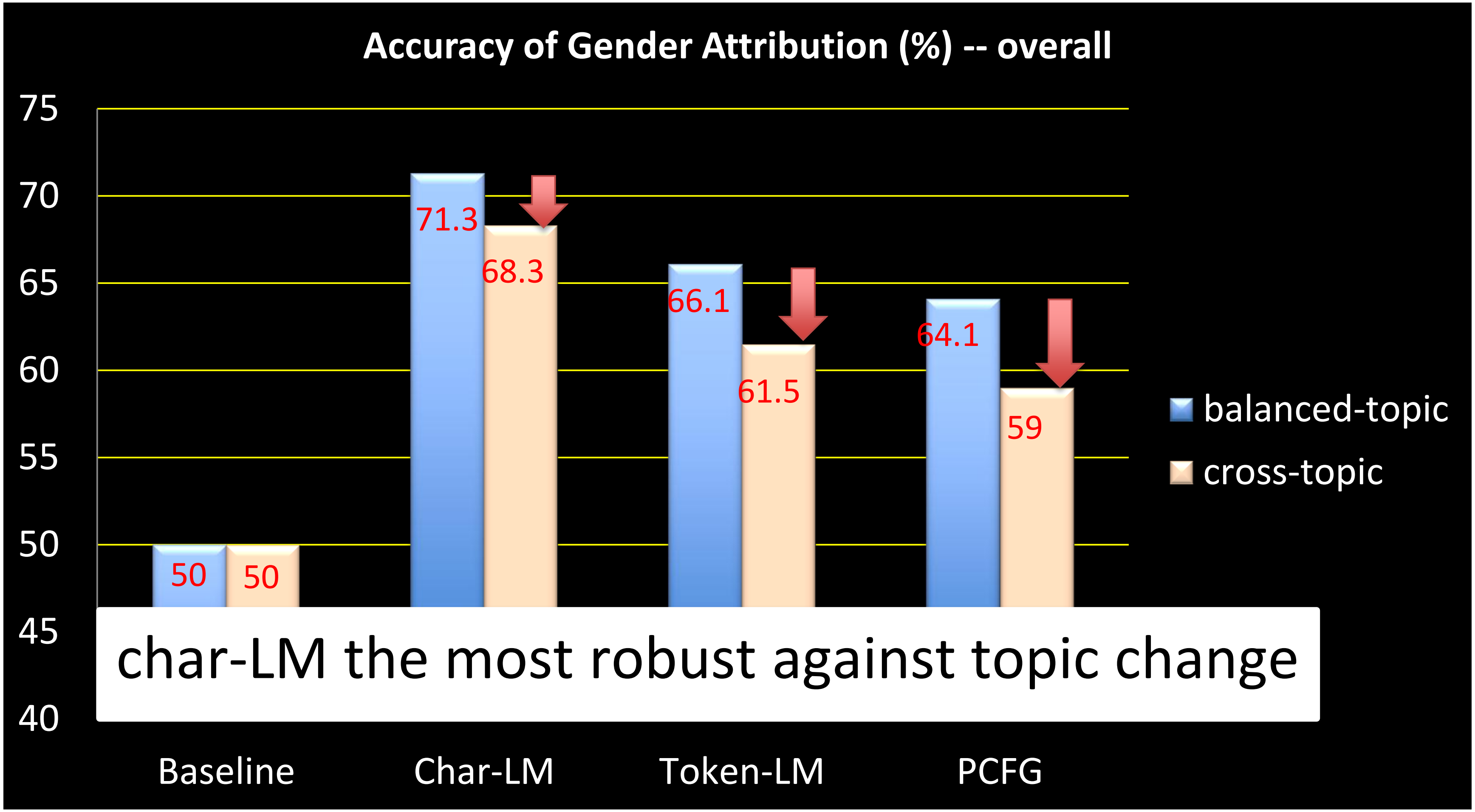
Plan for the Experiments

- Blog dataset (7 different topics)
  I. balanced-topic
  II. cross-topic
- Scientific paper dataset (10 different authors)
  III. balanced-topic (balanced-author)
  IV. cross-topic (cross-author)
- Both datasets
  V. cross-topic & cross-genre

Experiment I & II: balanced-topic v.s. crossed-topic



Experiment I & II: balanced-topic v.s. crossed-topic

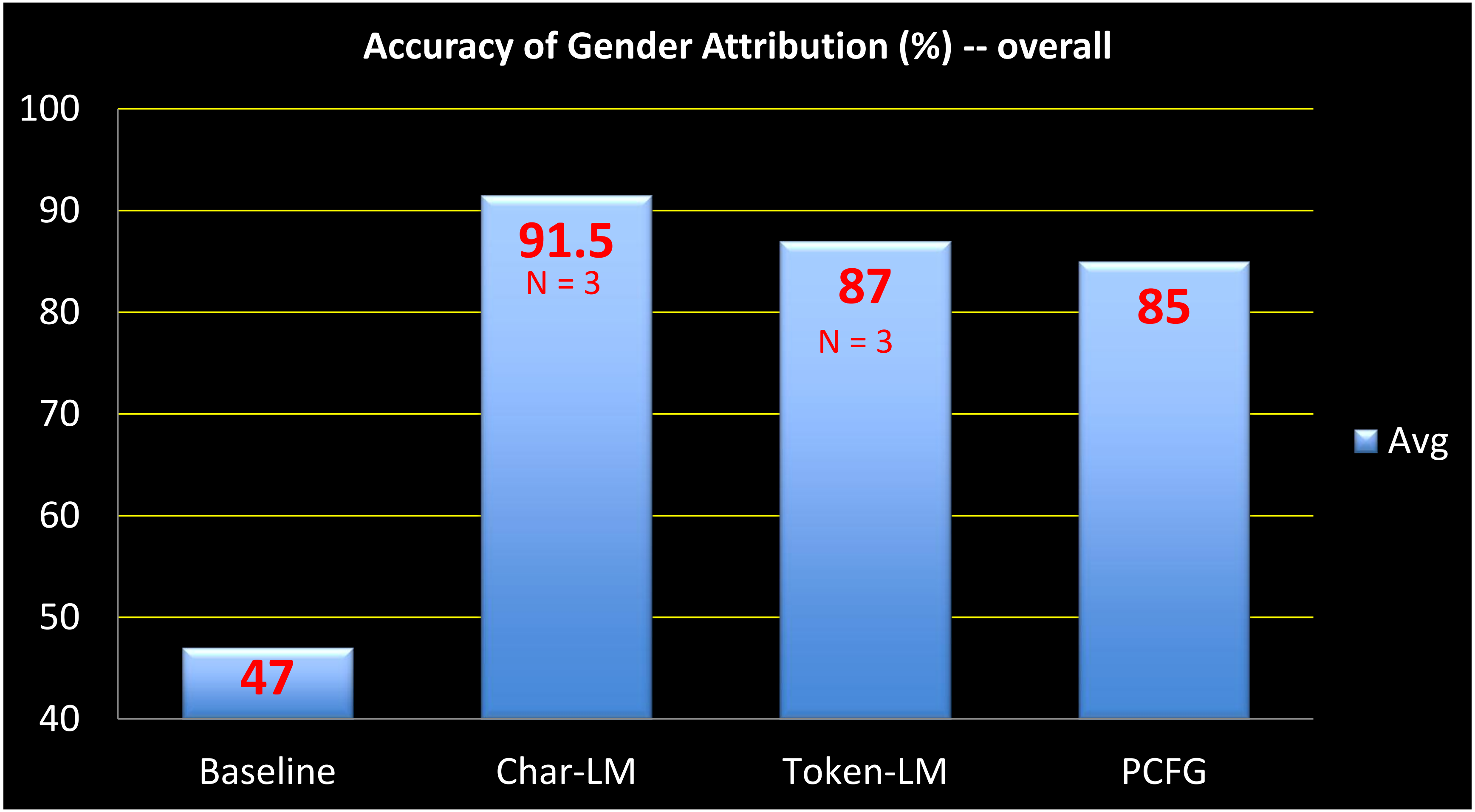


char-LM the most robust against topic change

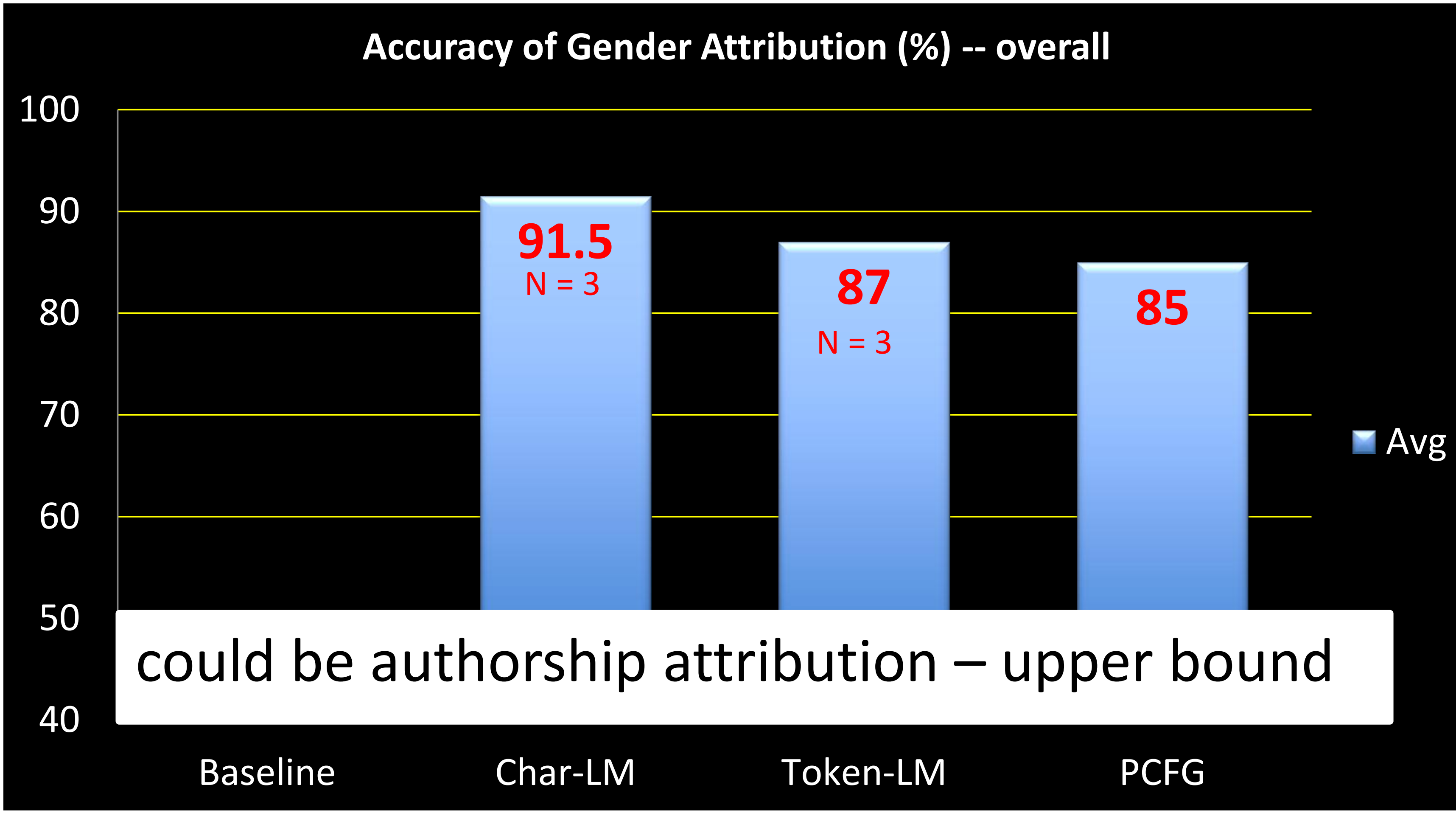
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment III: balanced-topic, scientific



Experiment III:
balanced-topic, scientific

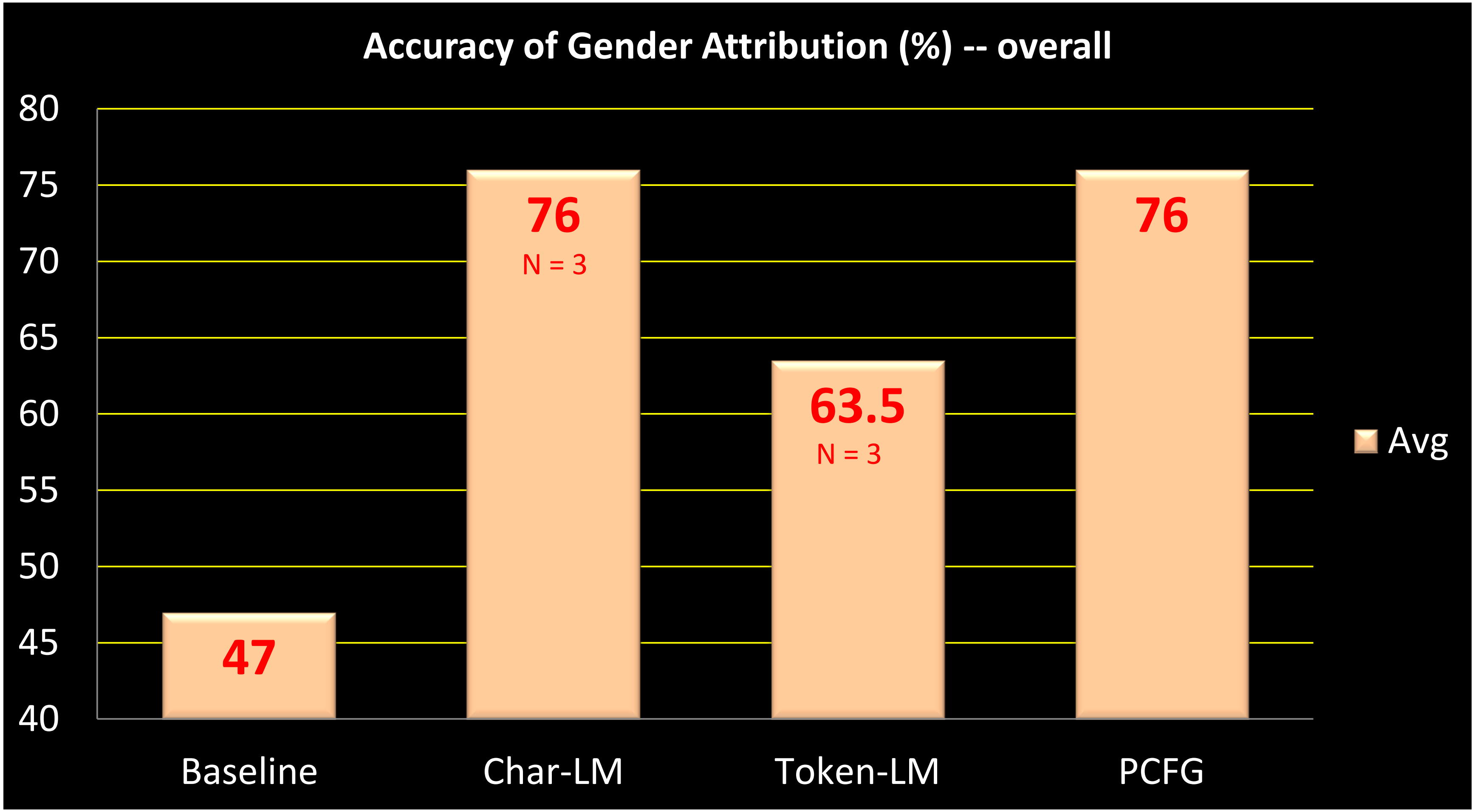


could be authorship attribution – upper bound

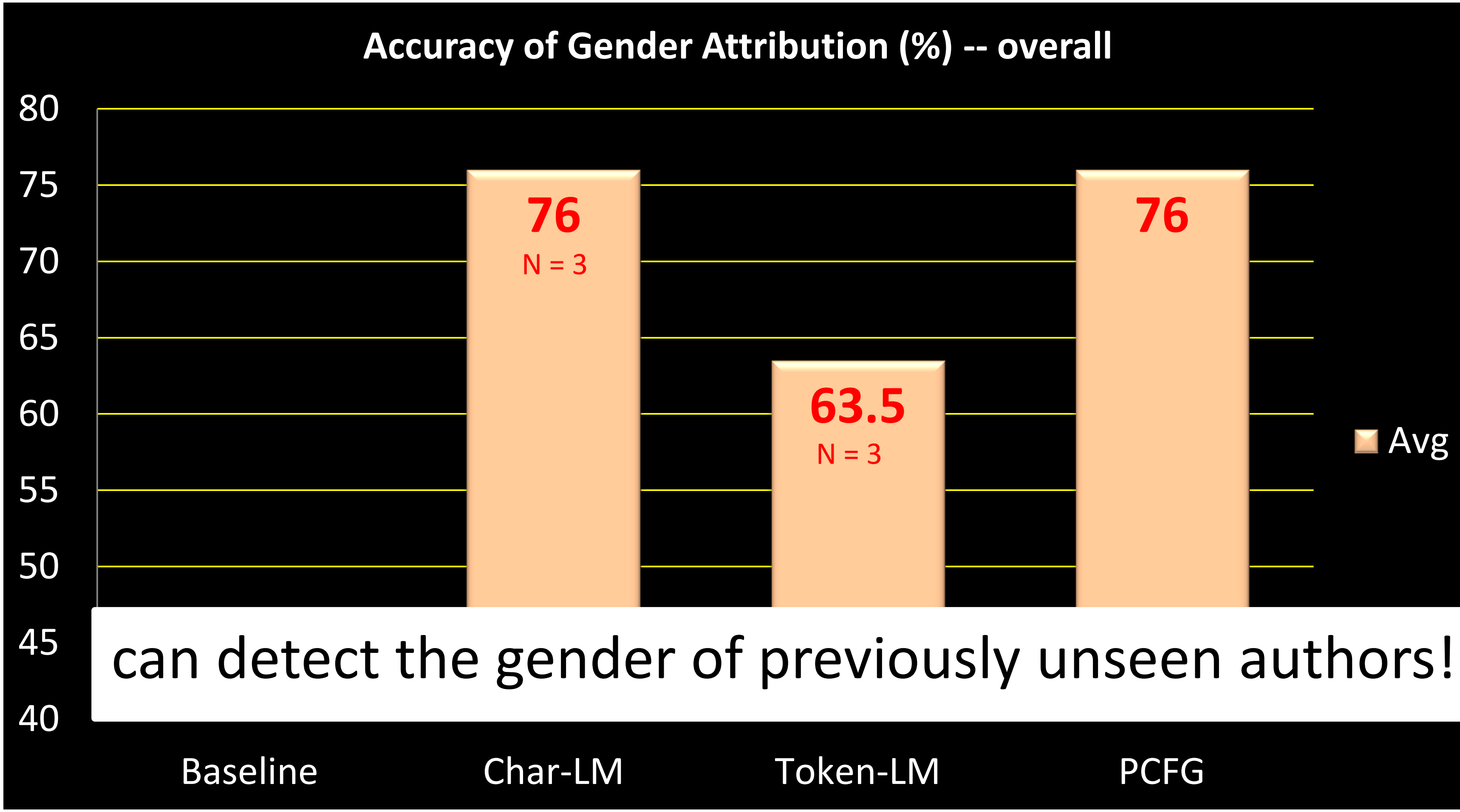
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment IV: cross-topic, scientific



Experiment IV: cross-topic, scientific

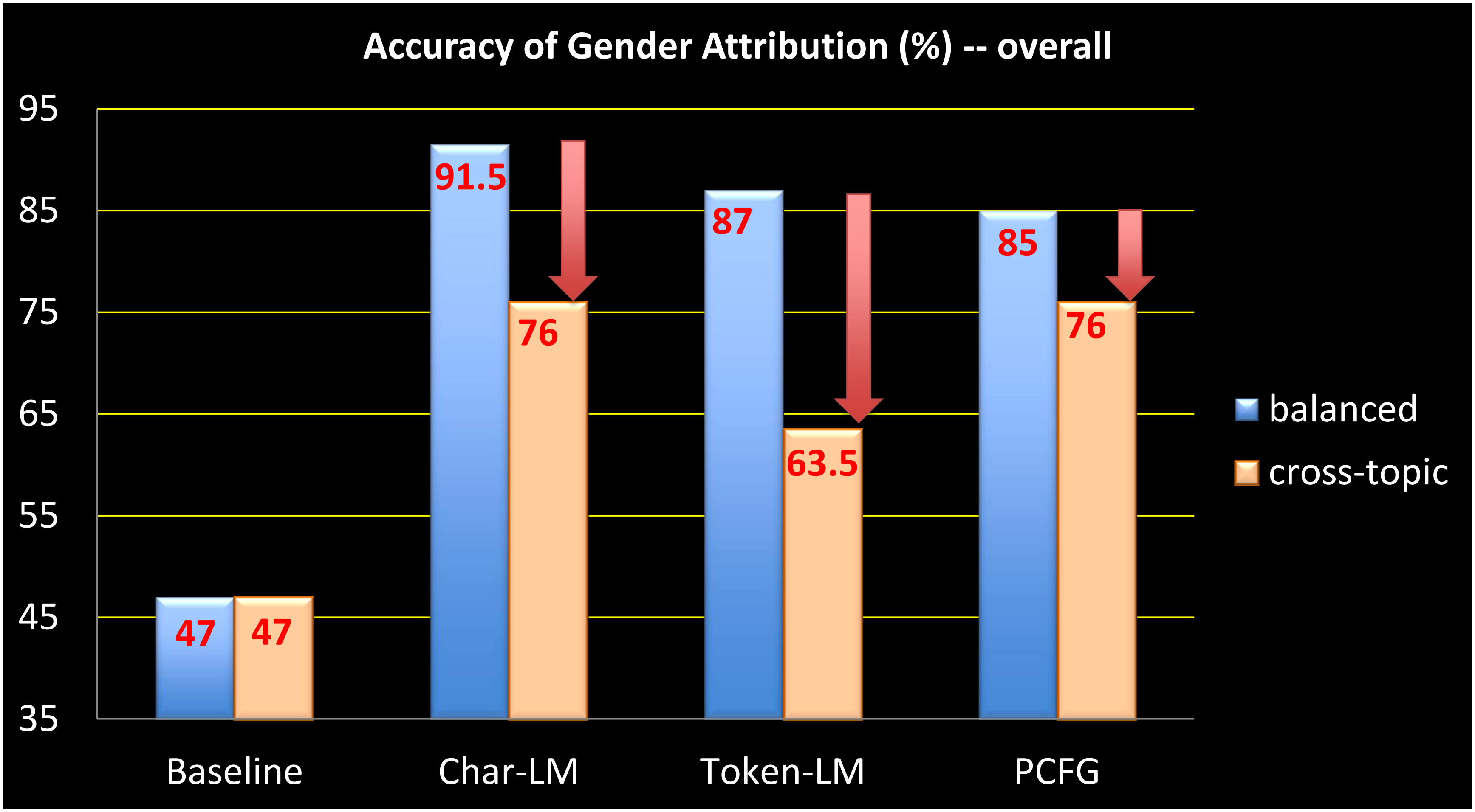


can detect the gender of previously unseen authors!

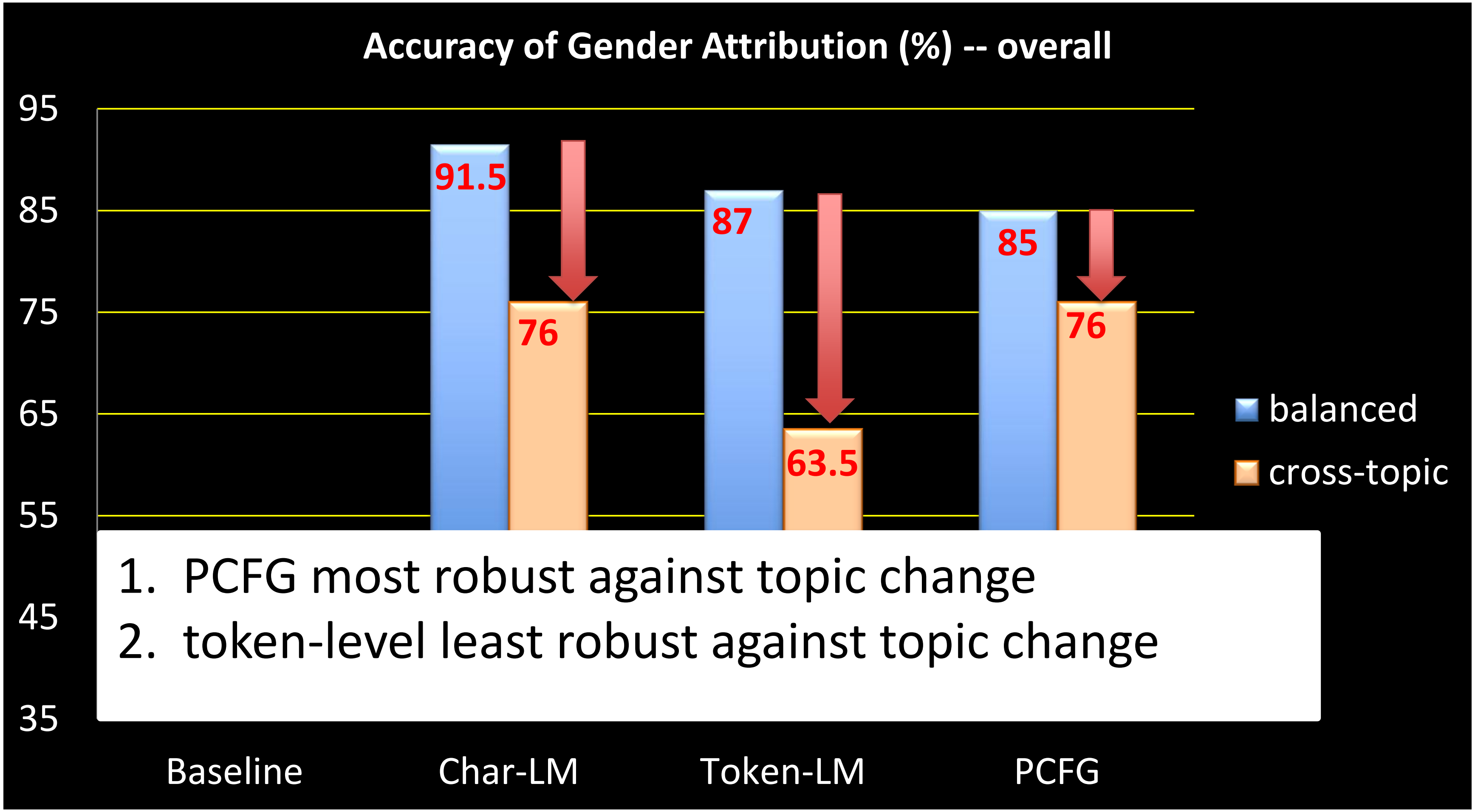
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment II & IV: cross-topic, scientific v.s. blog



Experiment II & IV: cross-topic, scientific v.s. blog

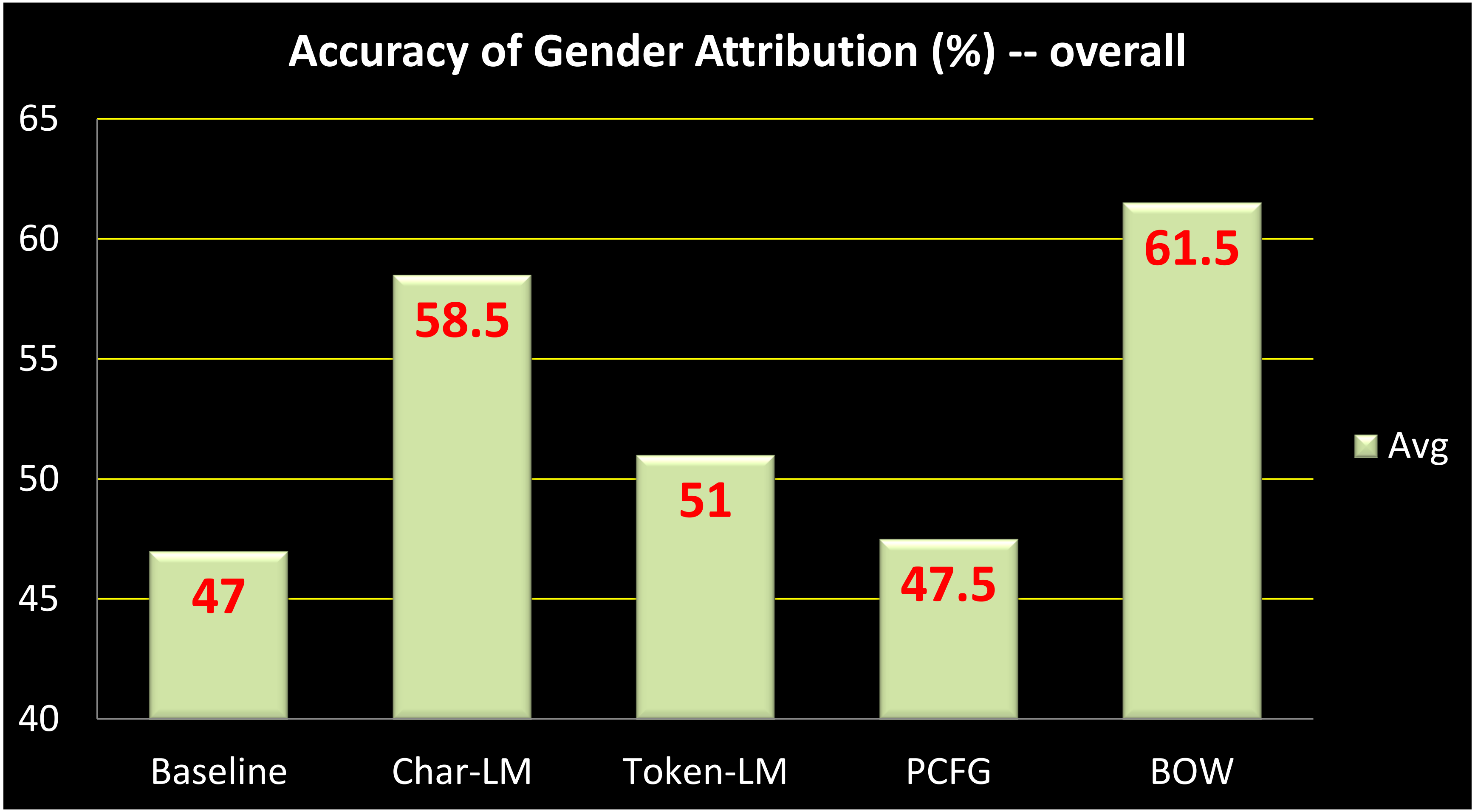


1. PCFG most robust against topic change
2. token-level least robust against topic change

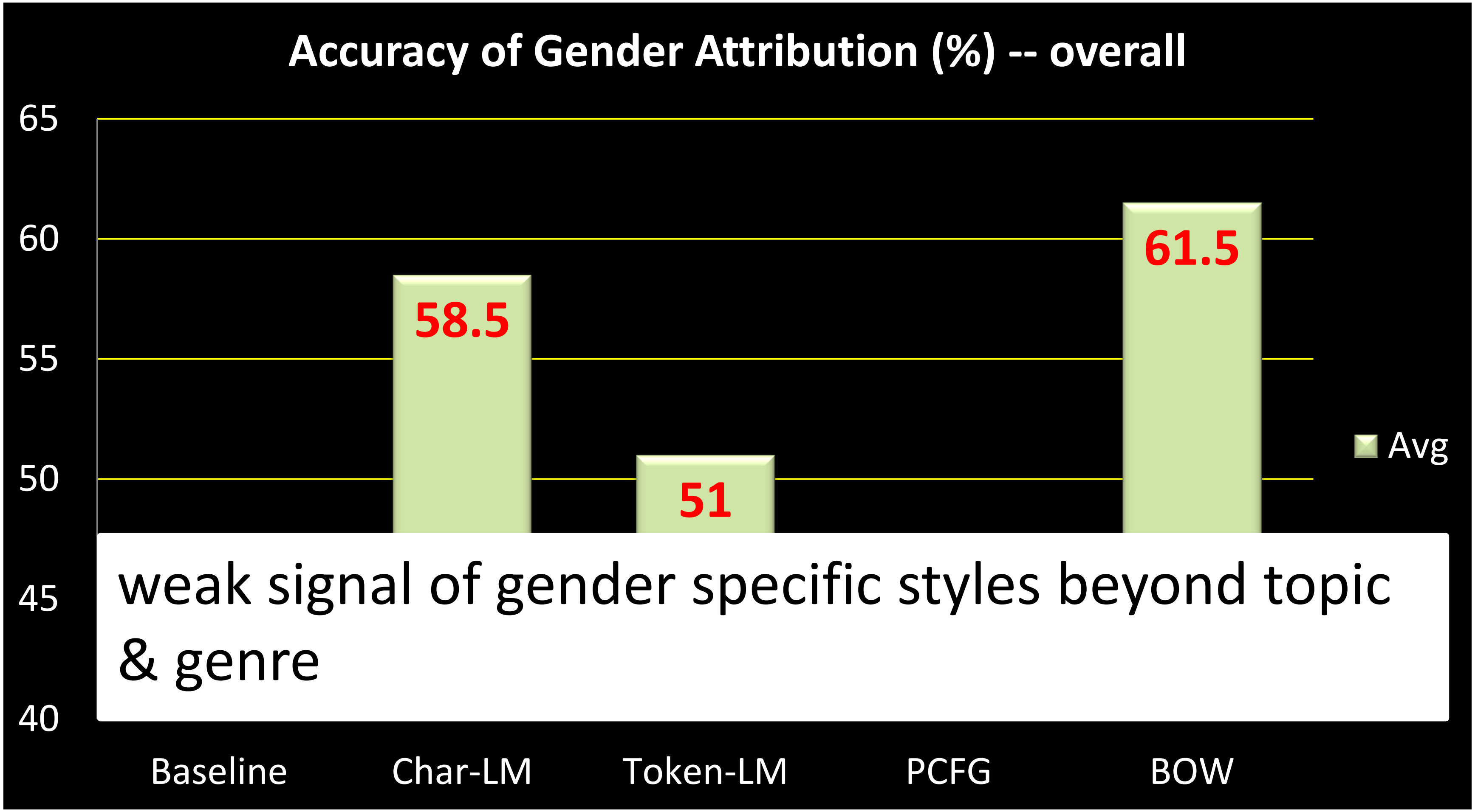
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre

Experiment V: cross-topic/genre, blog/scientific



Experiment V: cross-topic/genre, blog/scientific



weak signal of gender specific styles beyond topic & genre

Conclusions (Case Study III)

- comparative study of machine learning techniques for gender attribution consciously removing gender bias in topics.
- statistical evidence of gender-specific language styles beyond topics and genres.

Collaborators

- @ Stony Brook University:

  Kailash Gajulapalli, Manoj Harpalani, Rob Johnson, Michael Hart, Ruchita Sarawgi , Sandesh Singh

- @ Cornell University:

  Claire Cardie, Jeffrey Hancock, Myle Ott

- Based on
  - Ott et al., 2011 (ACL)
  - Harpalani et al., 2011 (ACL)
  - Sarawgi et al., 2011 (CoNLL)

THANK YOU!