User Attribute Prediction on Social Media

Daniel Preoţiuc-Pietro

Positive Psychology Center
University of Pennsylvania

21 November 2016

Motivation

Traditional psychological (social science) research study traits and behaviors reported through surveys:

- Personality
- Health habits
- Preferences

Motivation

These have some caveats:

- reach a limited sample size
- people might not be accurate in how they present themselves
- sample is usually biased
  - most psychological research is based on **W**estern **E**ducated from **I**ndustrialized **R**ich **D**emocratic countries samples

Motivation

Surveys: 500–1000

Facebook: 1.79 Billion

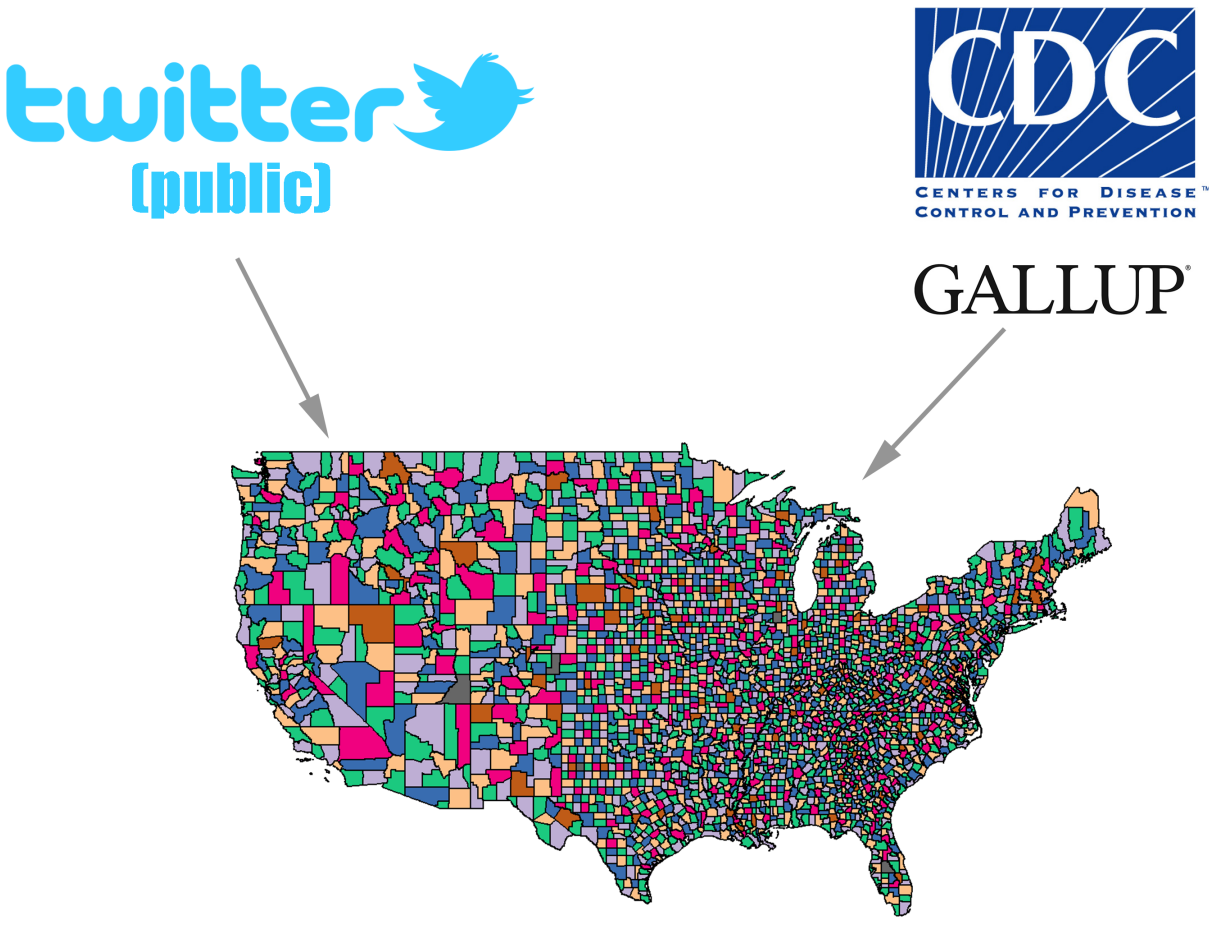
Motivation

Social media allows to take measurements:

- real-time
- unobtrusive
- fine temporal granularity
- fine spatial granularity

Motivation

Real-time community level research



Ethics

Twitter – profiles are public by default

Facebook/Instagram – users provide informed consent to share data

Data sets containing sensitive trait-level information provided through surveys are anonymized.

All studies were approved by the institutional UPenn Internal Review Board (IRB)

Motivation

User attribute prediction from text is successful:

- Age (Rao et al. 2010, ACL)
- Gender (Burger et al. 2011, EMNLP)
- Location (Eisenstein et al. 2010, EMNLP)
- Personality (Schwartz et al. 2013, PLoS One)
- Popularity (Lampos et al. 2014, EACL)
- Political Orientation (Volkova et al. 2014, ACL)
- Mental illness (Coppersmith et al. 2014, ACL)
- Occupation (Preoţiuc-Pietro et al. 2015, ACL)
- Income (Preoţiuc-Pietro et al. 2015, PLoS ONE)
- 'Dark Triad' Personality (Preoţiuc-Pietro et al. 2016, CIKM)

Hypothesis

Online footprints reveal individual differences in both demographic and psychological traits.

Task

Two sides of the problem:

1. **Measurement**
   - **Goal:** build models to predict traits of unknown users
   - predictive setup (regression/classification)
   - using large scale Machine Learning
2. **Insight**
   - **Goal:** gain a better understanding of group differences
   - interpretable features
   - use domain experts in analysis

Socio-economic factors

Socio-economic factors (occupation, social class, education, income) play a vital role in language use (Bernstein 1960, Labov 1972/2006)

Our contributions:

- Predicting new user attributes: occupation and income
- New dataset: user $\longleftrightarrow$ occupation $\longleftrightarrow$ income
- Gaussian Process classification for NLP tasks
- Feature ranking and analysis using non-linear methods

Standard Occupational Classification

Standardised job classification taxonomy

Developed and used by the UK Office for National Statistics (ONS), but applicable to any country

Hierarchical:

- 1-digit (major) groups: 9
- 2-digit (sub-major) groups: 25
- 3-digit (minor) groups: 90
- 4-digit (unit) groups: 369

Jobs grouped by **skill requirements**

Standard Occupational Classification

C1 Managers, Directors and Senior Officials

- 11 Corporate Managers and Directors
  - 111 Chief Executives and Senior Officials (£111,413)
    - 1115 Chief Executives and Senior Officials
      Job: chief executive, bank manager
    - 1116 Elected Officers and Representatives
  - 112 Production Managers and Directors (£50,952)
  - 113 Functional Managers and Directors (£70,457)
  - 115 Financial Institution Managers and Directors (£73,911)
  - 116 Managers and Directors in Transport and Logistics (£35,589)
  - 117 Senior Officers in Protective Services (£111,413)
  - 118 Health and Social Services Managers and Directors (£46,629)
  - 119 Managers and Directors in Retail and Wholesale (£29,009)
- 12 Other Managers and Proprietors

Standard Occupational Classification

C2 Professional Occupations
Job: mechanical engineer, pediatrist, postdoctoral researcher

C3 Associate Professional and Technical Occupations
Job: system administrator, dispensing optician

C4 Administrative and Secretarial Occupations
Job: legal clerk, company secretary

C5 Skilled Trades Occupations
Job: electrical fitter, tailor

C6 Caring, Leisure, Other Service Occupations
Job: school assistant, hairdresser

C7 Sales and Customer Service Occupations
Job: sales assistant, telephonist

C8 Process, Plant and Machine Operatives
Job: factory worker, van driver

C9 Elementary Occupations
Job: shelf stacker, bartender

Data

5191 users $\longleftrightarrow$ 3-digit job group $\longleftrightarrow$ mean income

Users collected by self-disclosure of job title in profile

Manually filtered by the authors

10M tweets, average 94.4 users per 3-digit group

Here we classify only at the 1-digit top level group (9 classes)

Feature representation and labels available online

Features

User Level features (**18**), such as:

- number of:
  - followers
  - friends
  - listings
  - tweets
- proportion of:
  - retweets
  - hashtags
  - @-replies
  - links
- average:
  - tweets/day
  - retweets/tweet

Features

Focus on **interpretable** features for analysis

Compute over reference corpus of 400M tweets:

- SVD embeddings and clusters
- Word2Vec(W2V) embeddings and cluster

SVD Features

Compute word × word similarity matrix

Similarity metric is Normalized PMI (Bouma 2009) using the entire tweet as context

SVD with different number of dimensions (30, 50, 100, 200)

User is represented by summing its word representations

The low-dimensional features offer no interpretability

SVD Features

Spectral clustering to get hard clusters of words (30, 50, 100, 200 clusters)

Each cluster consists of distributionally similar words $\longleftrightarrow$ topic

User is represented by the number of times he uses a word from each cluster

Word2Vec Features

Trained Word2Vec (layer size 50) on our Twitter reference corpus

Spectral clustering on the word $\times$ word similarity matrix (30, 50, 100, 200 clusters)

Similarity is cosine similarity of words in the embedding space

Gaussian Processes

Brings together several key ideas in one framework:

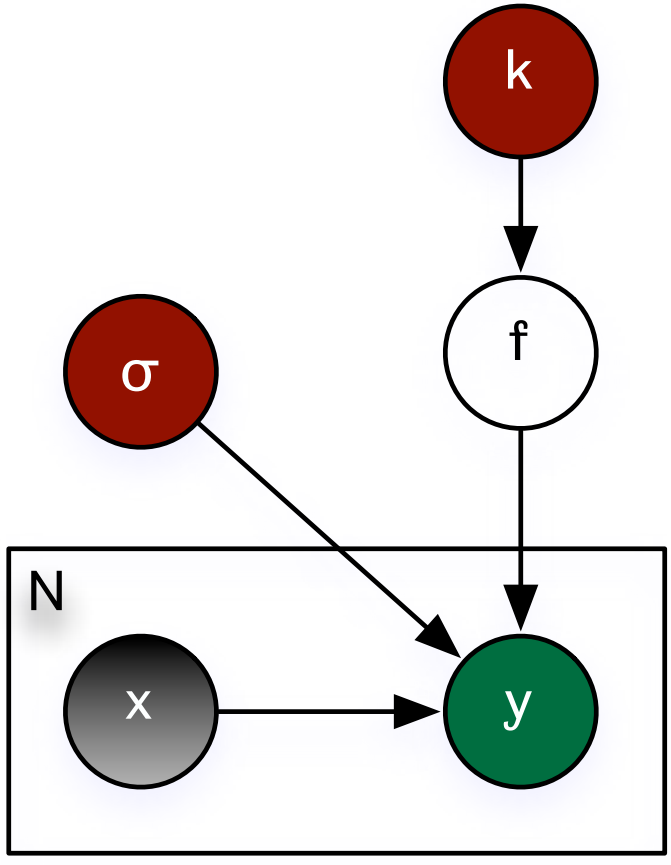
- Bayesian
- kernelised
- non-parametric
- non-linear
- modelling uncertainty

Elegant and powerful framework, with growing popularity in machine learning and application domains

Gaussian Process Graphical Model View

$$\mathbf{f} \sim \mathcal{GP}(m, k)$$
$$y \sim \mathcal{N}(f(\mathbf{x}), \sigma^2)$$

- $f : \mathcal{R}^D \to \mathbb{R}$ is a latent function
- y is a noisy realisation of $f(x)$
- k is the covariance function of kernel
- m and σ^2 are learnt from data



Gaussian Process Classification

Pass latent function through logistic function to squash the input from $(-\infty, -\infty)$ to obtain probability, $\pi(x) = p(y_i = 1 || f_i$ (similar to logistic regression)

The likelihood is non-Gaussian and solution is not analytical

Inference using Expectation Propagation (EP)

FITC approximation for large data

Gaussian Process Classification

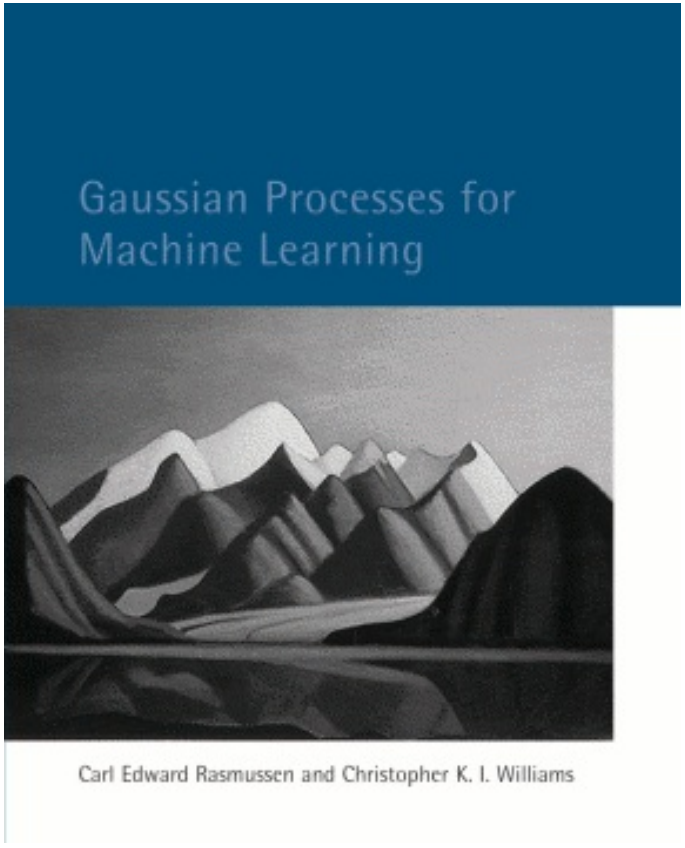
ARD kernel learns feature importance → features most **discriminative** between classes

We learn 9 one-vs-all binary classifiers

This way, we find the most predictive features consistent for all classes

Gaussian Process Resources

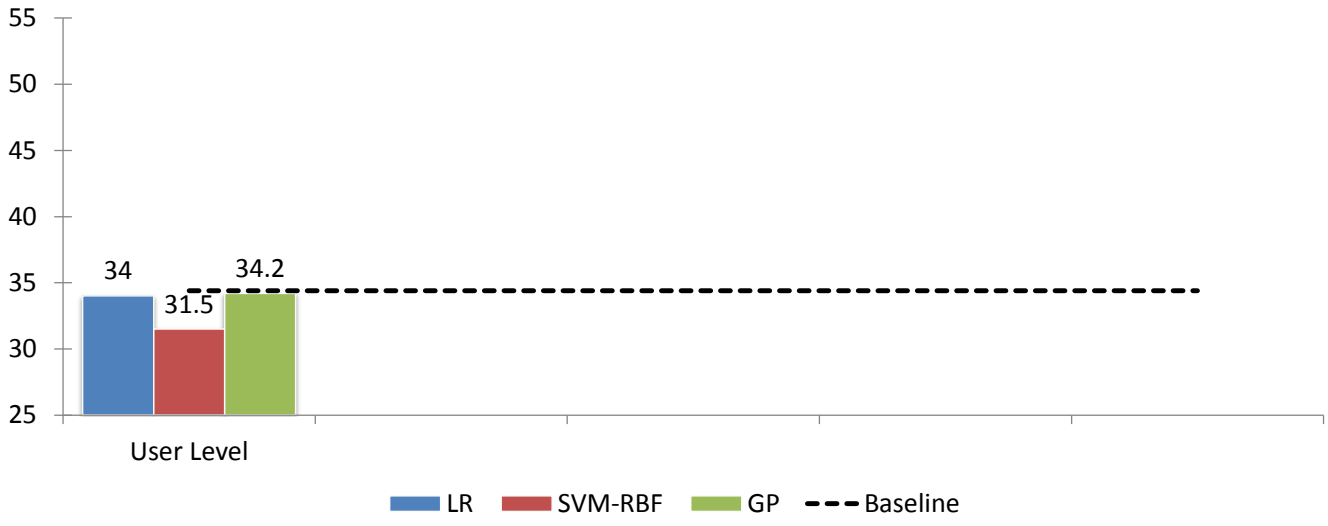
Free book:
`http://www.gaussianprocess.org/gpml/chapter`



Gaussian Process Resources

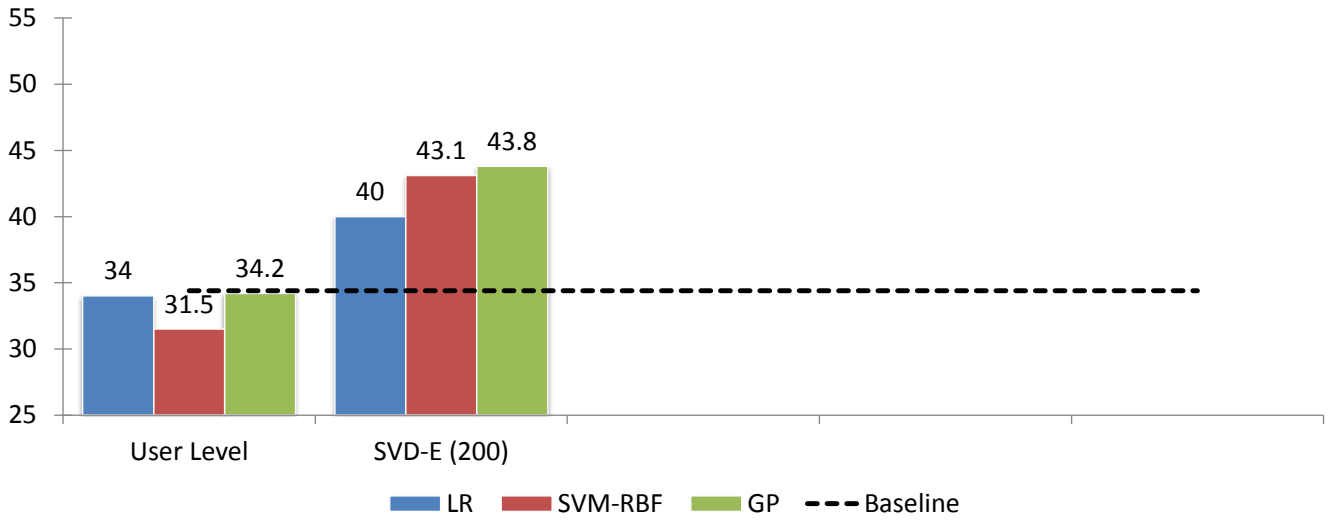
- GPs for Natural Language Processing tutorial (ACL 2014) `http://www.preotiuc.ro`
- GP School in Sheffield and roadshows in Kampala, Pereira, Nyeri, Melbourne `http://ml.dcs.shef.ac.uk/gpss/`
- Annotated bibliography and other materials `http://www.gaussianprocess.org`
- GPy Toolkit (Python) `https://github.com/SheffieldML/GPy`

Prediction



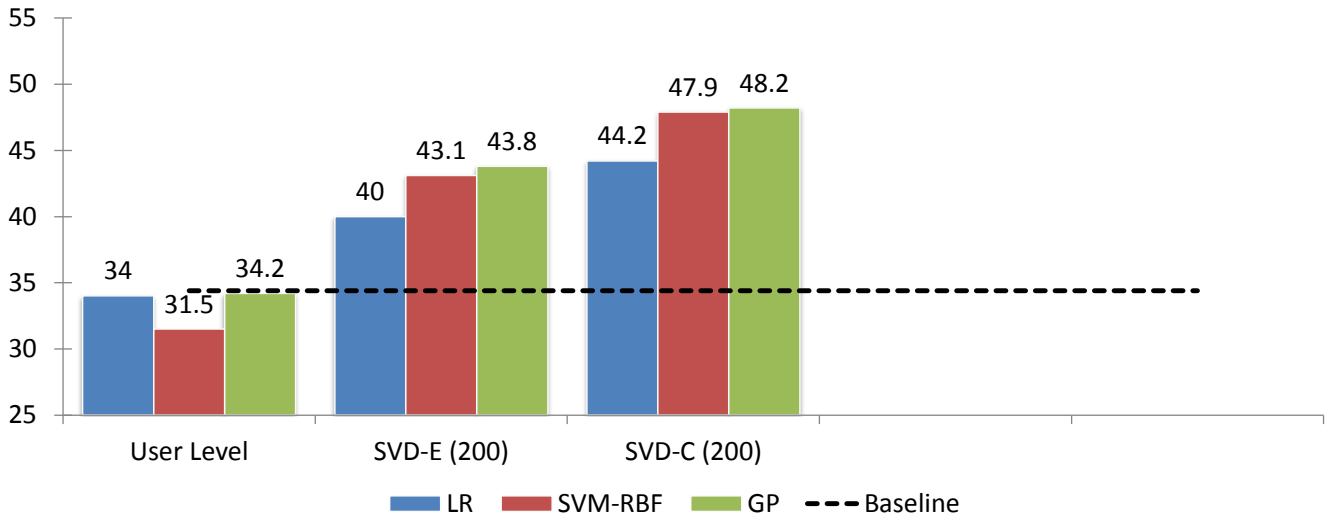
Stratified 10 fold cross-validation

Prediction



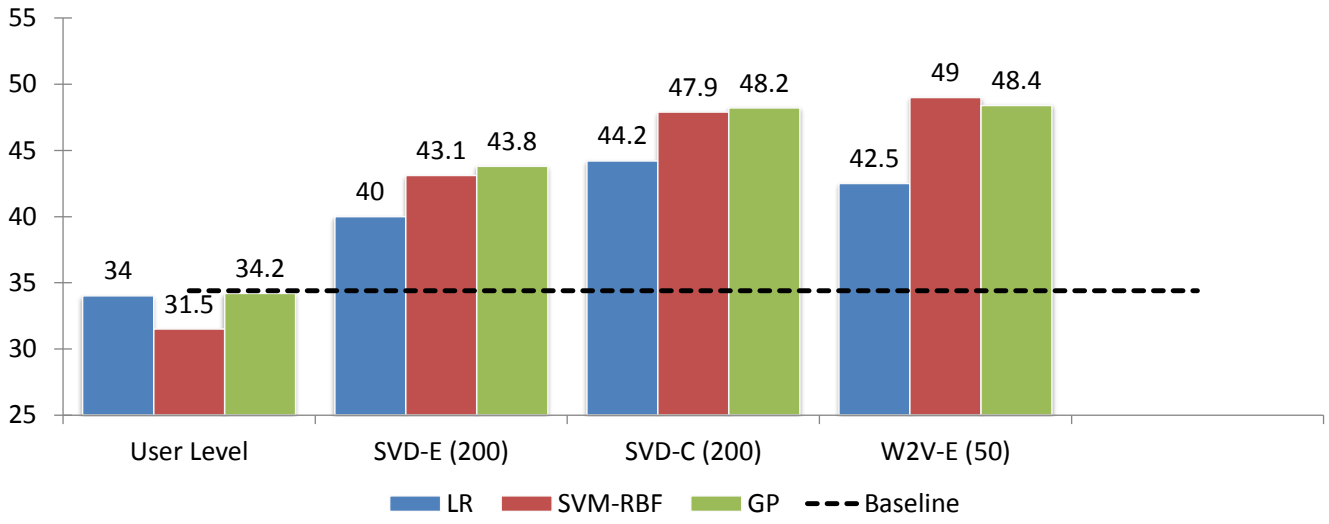
Stratified 10 fold cross-validation

Prediction



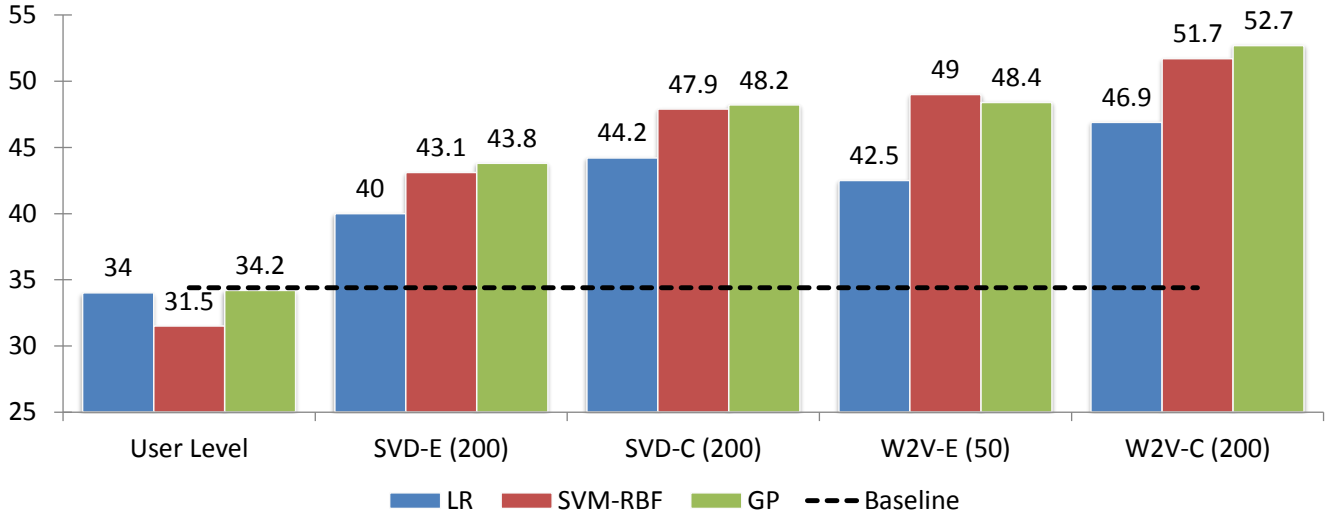
Stratified 10 fold cross-validation

Prediction



Stratified 10 fold cross-validation

Prediction



Stratified 10 fold cross-validation

Prediction Analysis

User level features have no predictive value

Clusters outperform embeddings

Word2Vec features are better than SVD/NPMI for prediction

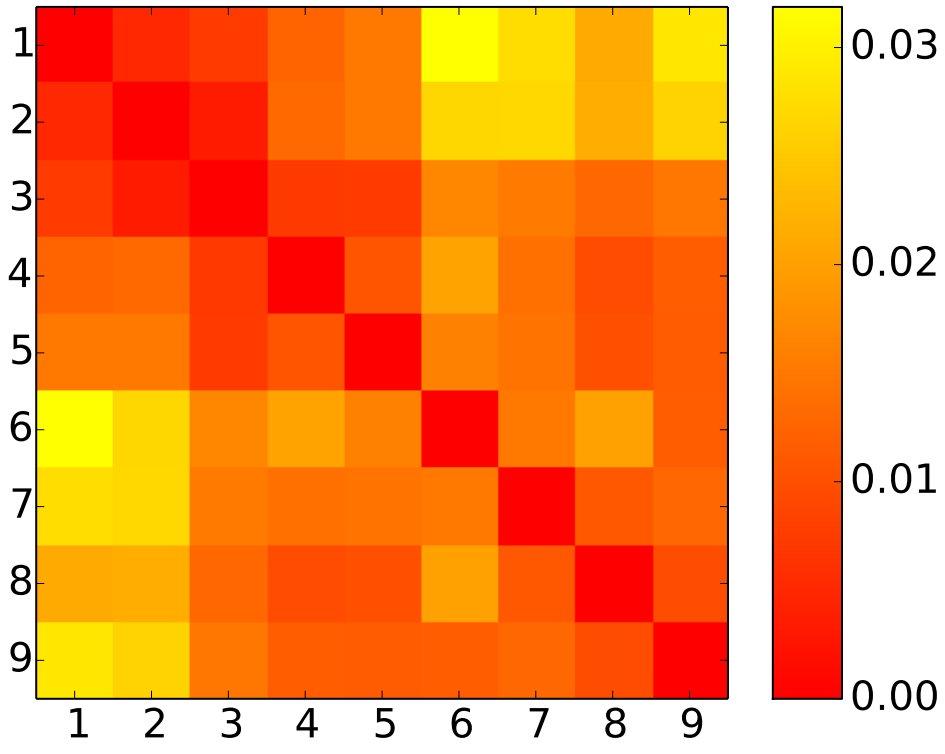
Non-linear methods (SVM-RBF and GP) significantly outperform linear methods

52.7% accuracy for 9-class classification is decent

Class Comparison

Jensen-Shannon Divergence between topic distributions across occupational classes

Some clusters of occupations are observable



Feature Analysis

Rank	**Manual Label**	**Topic** (most frequent words)
1	Arts	art, design, print, collection, poster, painting, custom, logo, printing, drawing
2	Health	risk, cancer, mental, stress, patients, treatment, surgery, disease, drugs, doctor
3	Beauty Care	beauty, natural, dry, skin, massage, plastic, spray, facial, treatments, soap
4	Higher Education	students, research, board, student, college, education, library, schools, teaching, teachers
5	Software Engineering	service, data, system, services, access, security, development, software, testing, standard

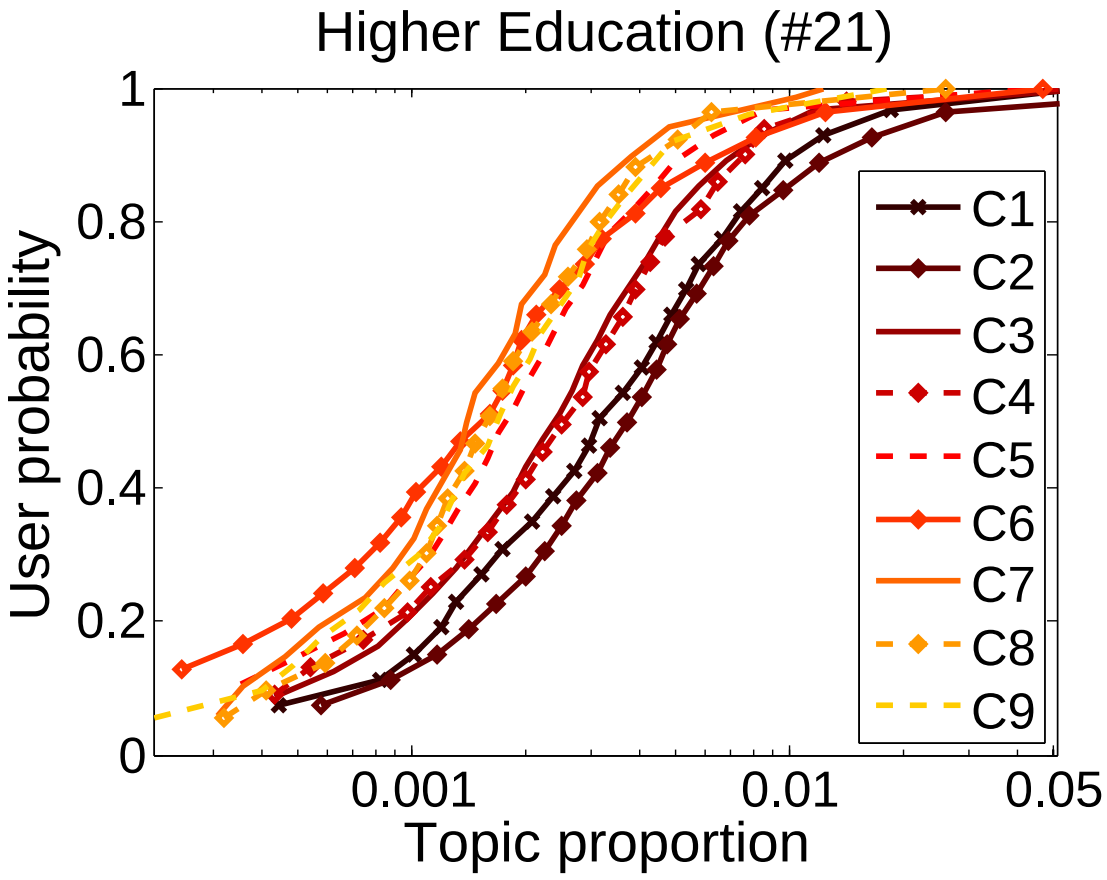
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

Feature Analysis

Rank	**Manual Label**	**Topic** (most frequent words)
7	Football	van, foster, cole, winger, terry, reckons, youngster, rooney, fielding, kenny
8	Corporate	patent, industry, reports, global, survey, leading, firm, 2015, innovation, financial
9	Cooking	recipe, meat, salad, egg, soup, sauce, beef, served, pork, rice
12	Elongated Words	wait, till, til, yay, ahhh, hoo, woo, woot, whoop, woohoo
16	Politics	human, culture, justice, religion, democracy, religious, humanity, tradition, ancient, racism

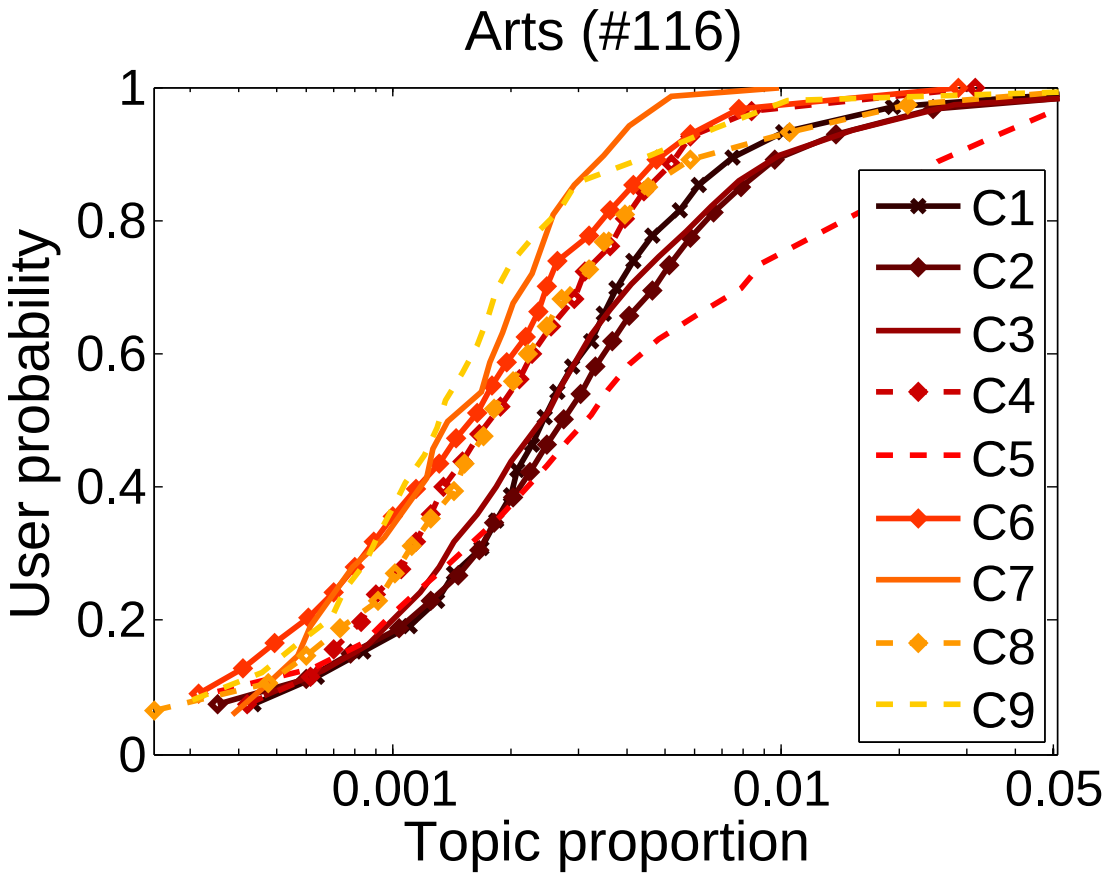
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

Feature Analysis - Cumulative Density Functions



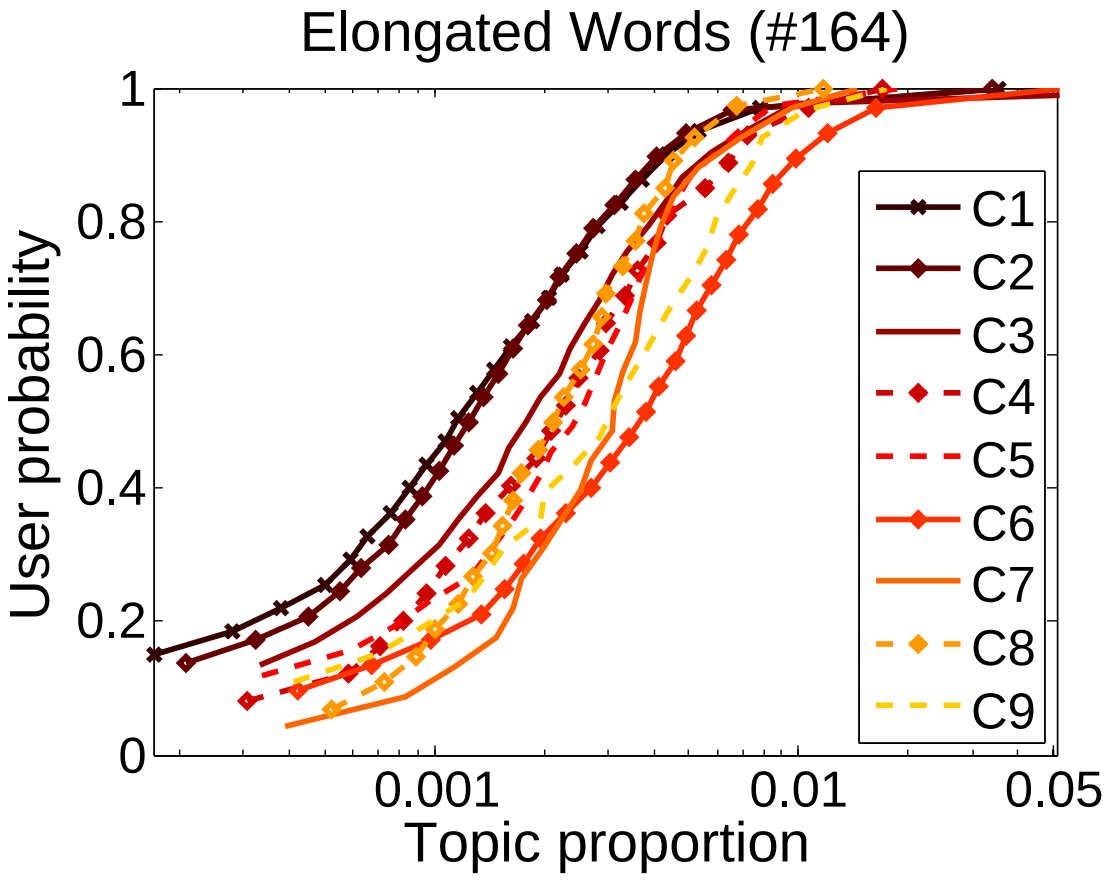
Topic more prevalent → CDF line closer to bottom-right corner

Feature Analysis - Cumulative Density Functions



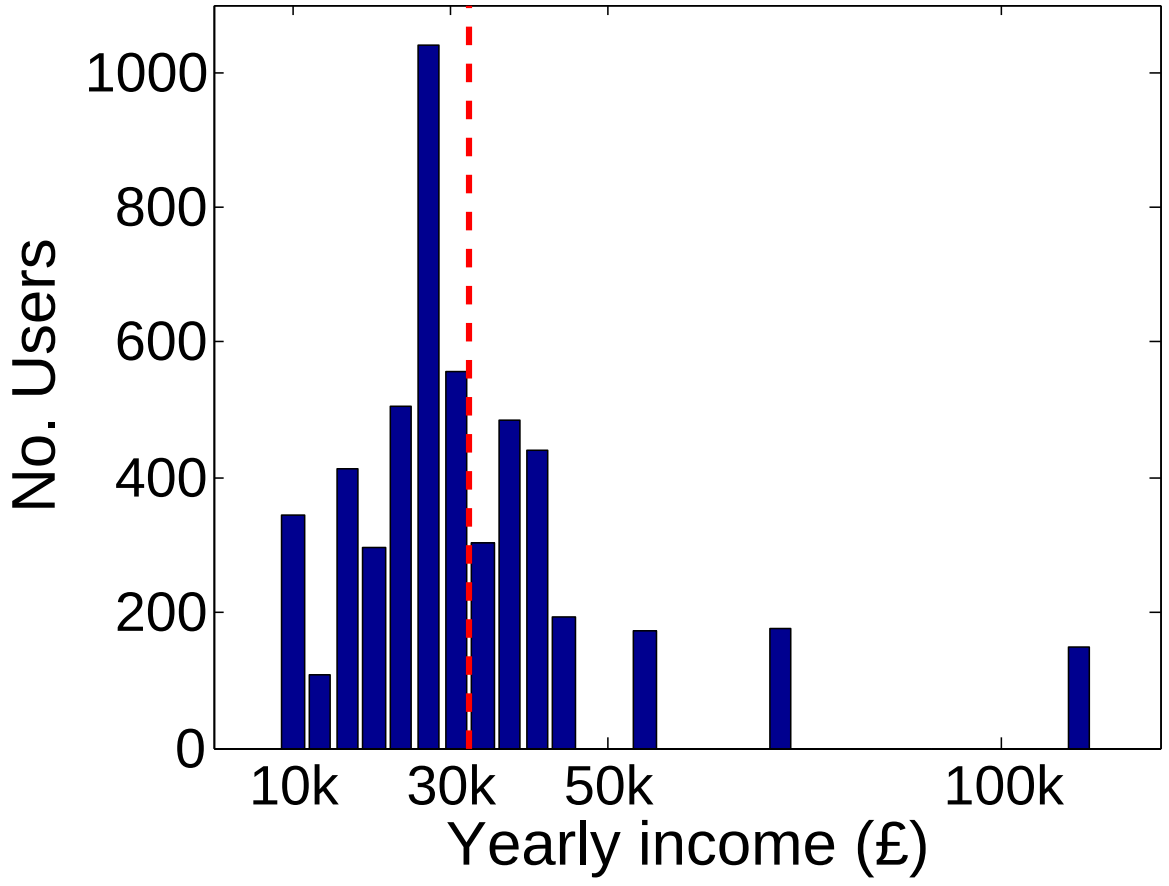
Topic more prevalent → CDF line closer to bottom-right corner

Feature Analysis - Cumulative Density Functions



Topic more prevalent → CDF line closer to bottom-right corner

Income Prediction (PLoS ONE 2015)

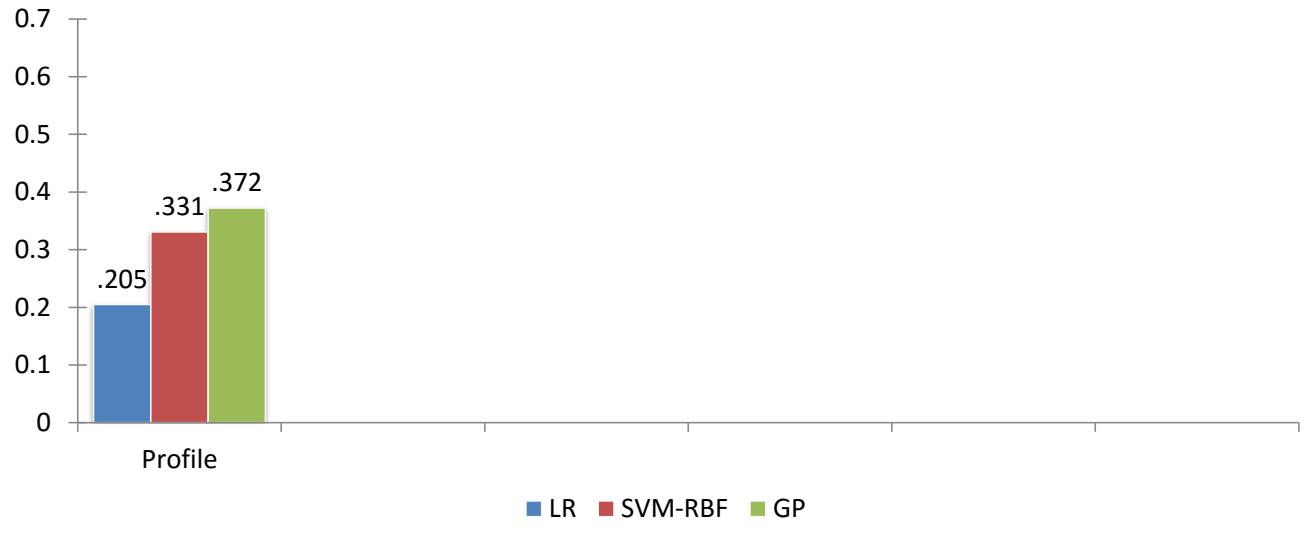


We approach the task as regression.

Features

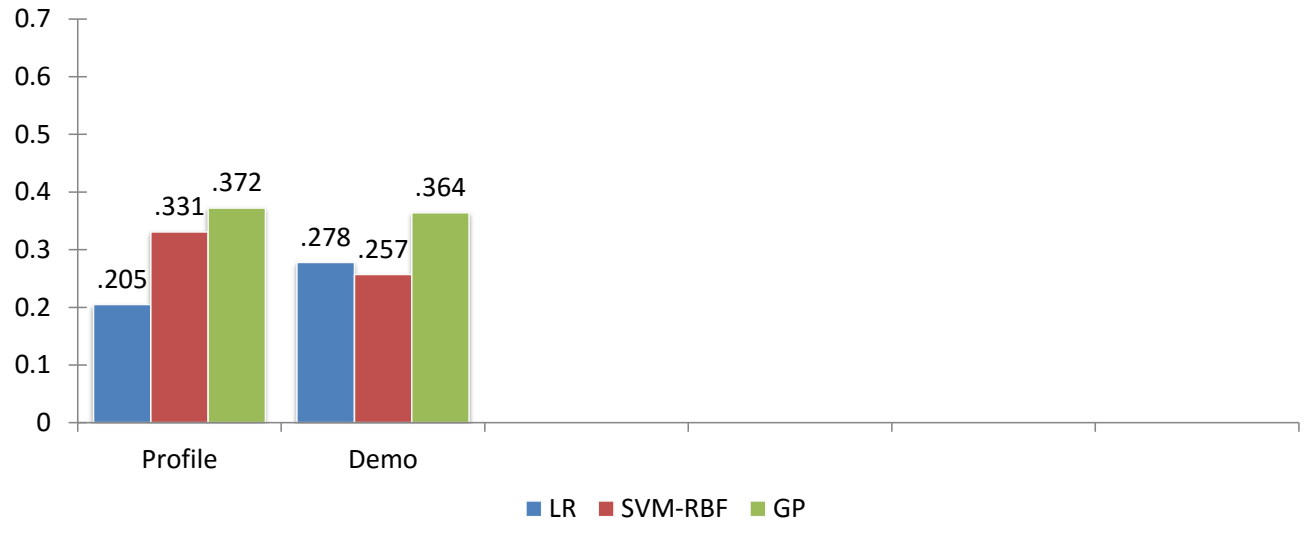
- Profile
  - Properties of the account (e.g. # followers)
- Predicted Psycho-Demographic Traits
  - e.g. Age, gender, life satisfaction
- Emotions
  - Sentiment and Ekman's six emotions
- Shallow Textual
  - Summary statistics from text (e.g. % of @-replies)

Prediction



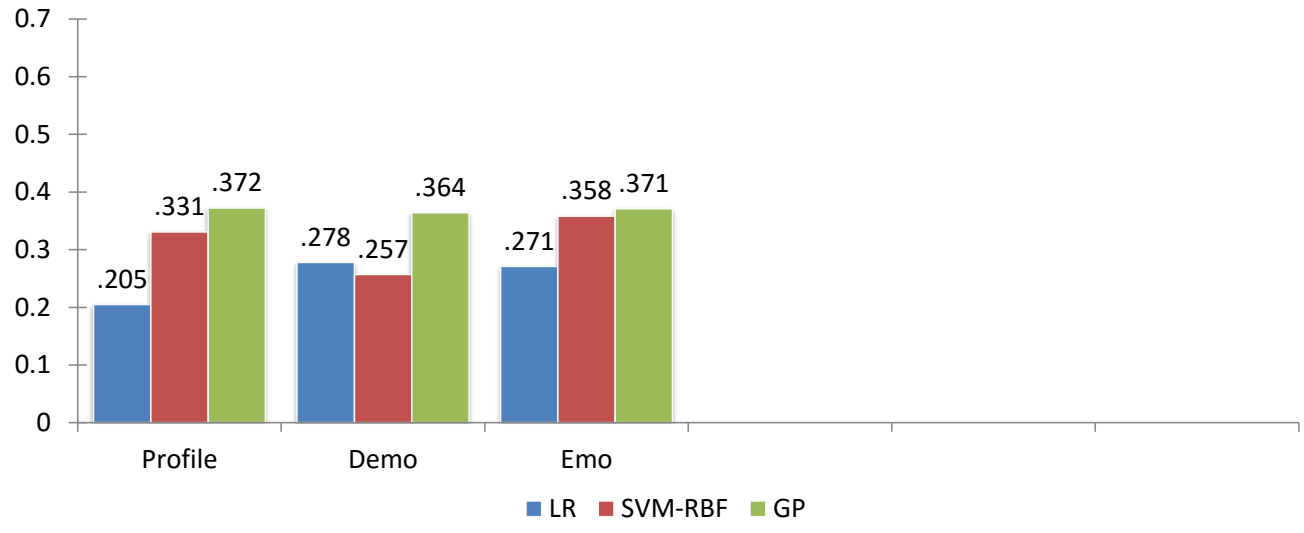
10 fold cross-validation

Prediction



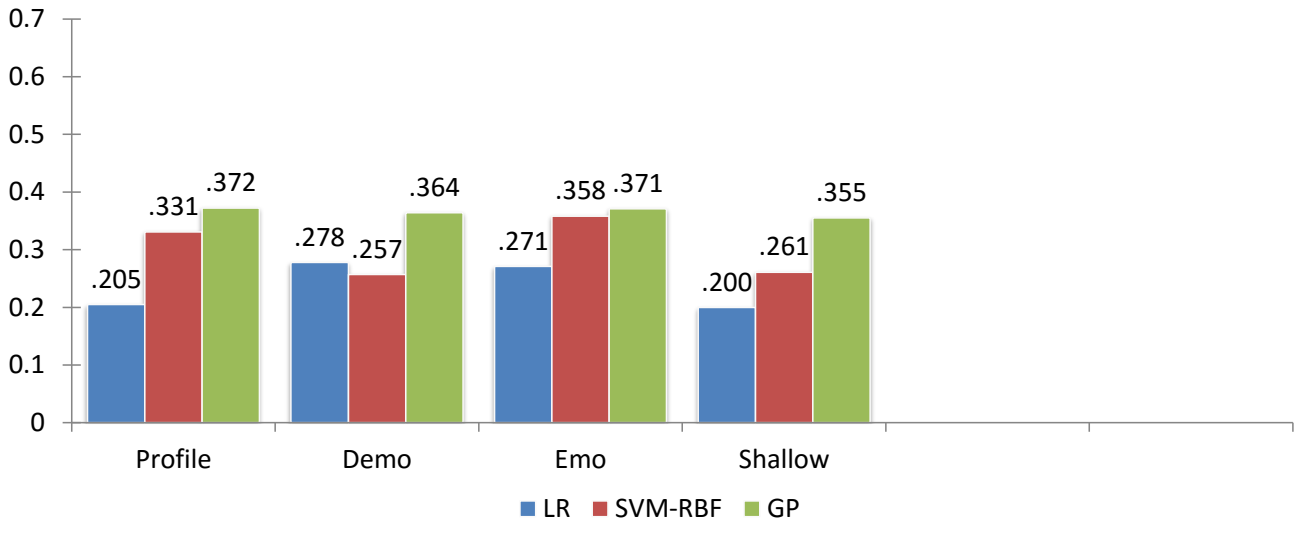
10 fold cross-validation

Prediction



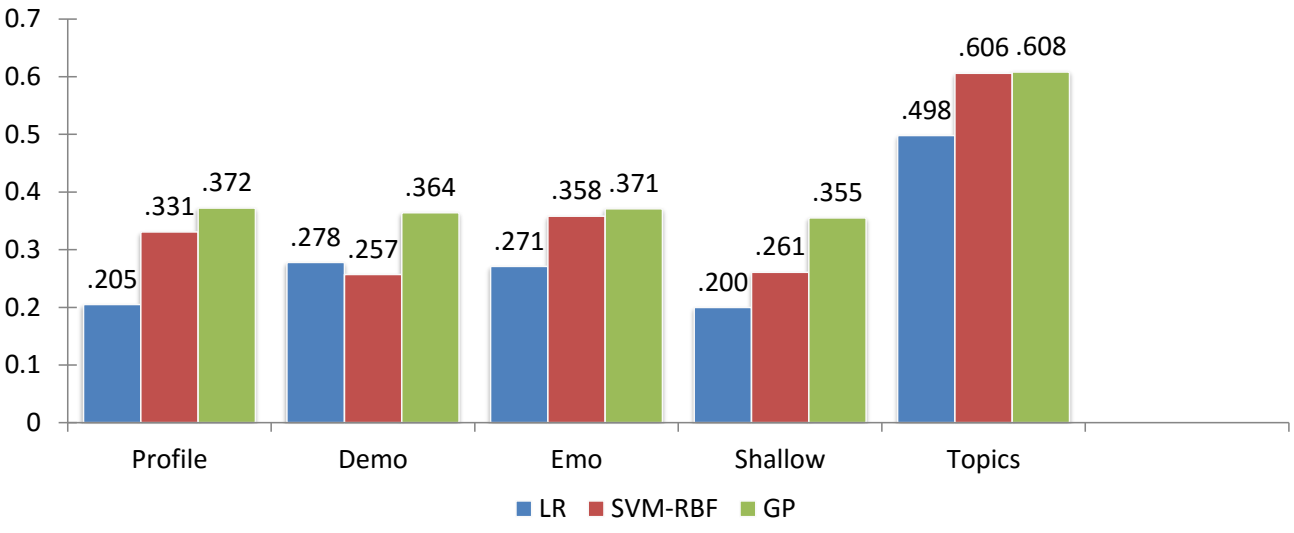
10 fold cross-validation

Prediction



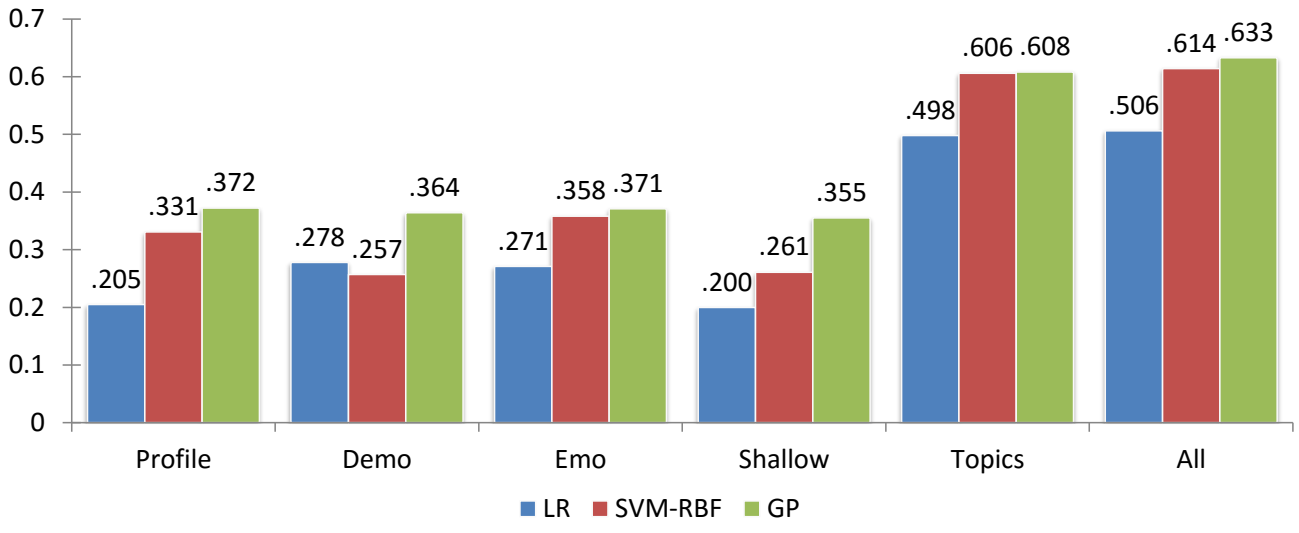
10 fold cross-validation

Prediction



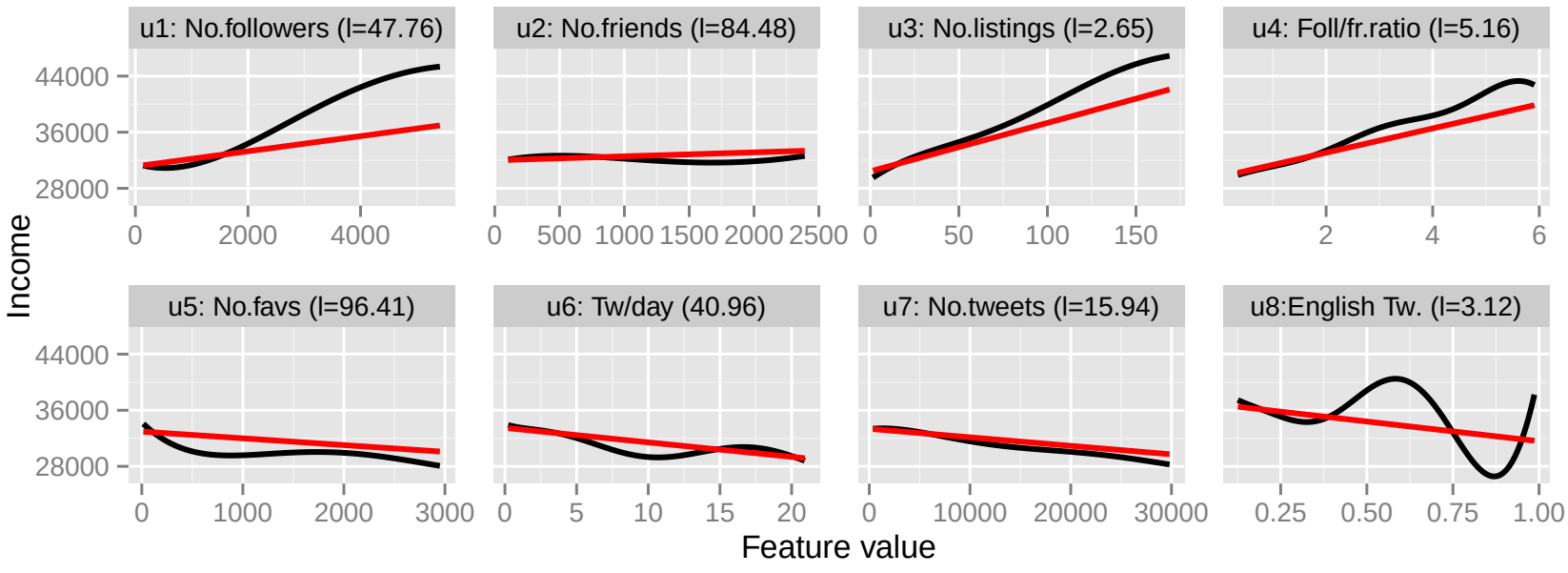
10 fold cross-validation

Prediction



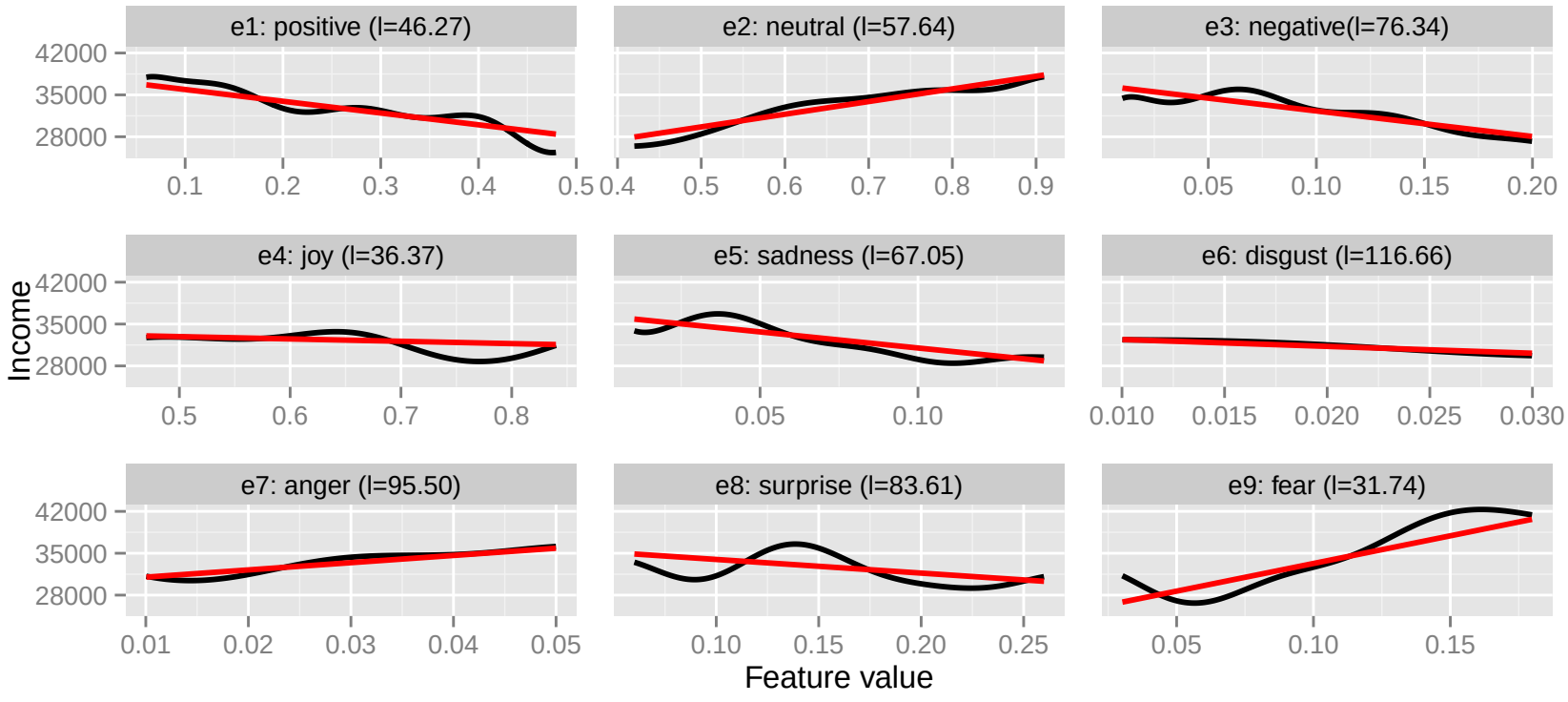
10 fold cross-validation

Profile Features



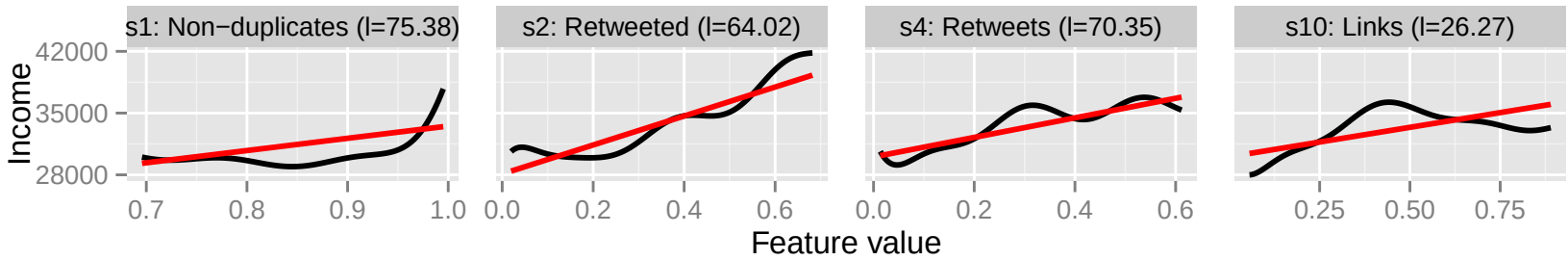
Linear fit, GP fit

Emotions



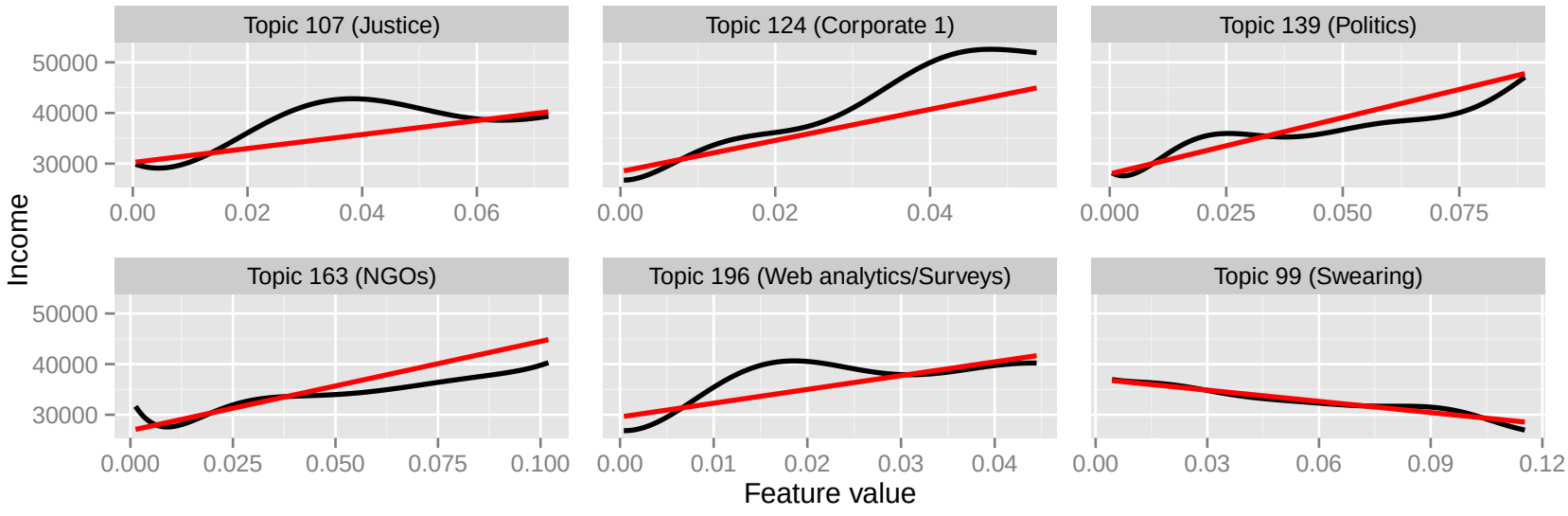
Linear fit, GP fit

Shallow Textual Features



Linear fit, GP fit

Topics



Linear fit, GP fit

Take Aways

User socio-economic status influences topic use in social media

Non-linear methods (Gaussian Processes) obtain significant gains over linear methods

Topic (clusters) features are both predictive and interpretable

New dataset available for research

Differences in Style (ACL 2016)

Most text prediction methods uncover **topical** differences

Not useful for many practical applications that adapt to traits:

- machine translation
- tutoring systems
- conversational agents
- control unwanted socio-demographic biases in NLP tools

We need to exploit differences in **style**, rather than in topic choice

Data Set

Stylistic differences often show similar tendencies across different traits (*wheee*: gender, extraversion, age, income...)

We study two related traits jointly:

- **Age**
  4,279 Twitter users; 574,095 tweets (Volkova et. al, 2015)
- **Income**
  5,191 Twitter users; 10,796,836 tweets (Preoţiuc-Pietro et al., ACL 2015)
  Age of this users predicted by an in-house state-of-the-art model (Sap et al., EMNLP 2014)

Features

Our features can be characterized by the four groups below:

- Surface
- Readability
- Syntax
- Style

Surface Features

- Average length of tweets in words and characters, length of words
- Ratio of words longer than five letters (shorter words are considered more readable)
- Type-token ratio per user – lexical density, a readability predictor (Oakland and Lane, 2004)
- Average number of smileys
- Average number of URLs
- Punctuation

Readability Features

- Automatic Readability Index (Senter and Smith, 1967)
- Flesch-Kincaid Grade Level (Kincaid et al., 1975)
- Coleman-Liau Index (Coleman and Liau, 1975)
- Flesch Reading Ease (Flesch, 1948)
- LIX Index (Anderson, 1983)
- Gunning-Fog Index (Gunning, 1969)
- SMOG grade (McLaughlin, 1969)

$$SMOG = 1.0430 \sqrt{\text{no. polysyllables} \times \frac{30}{\text{no. sentences}}} + 3.1291$$

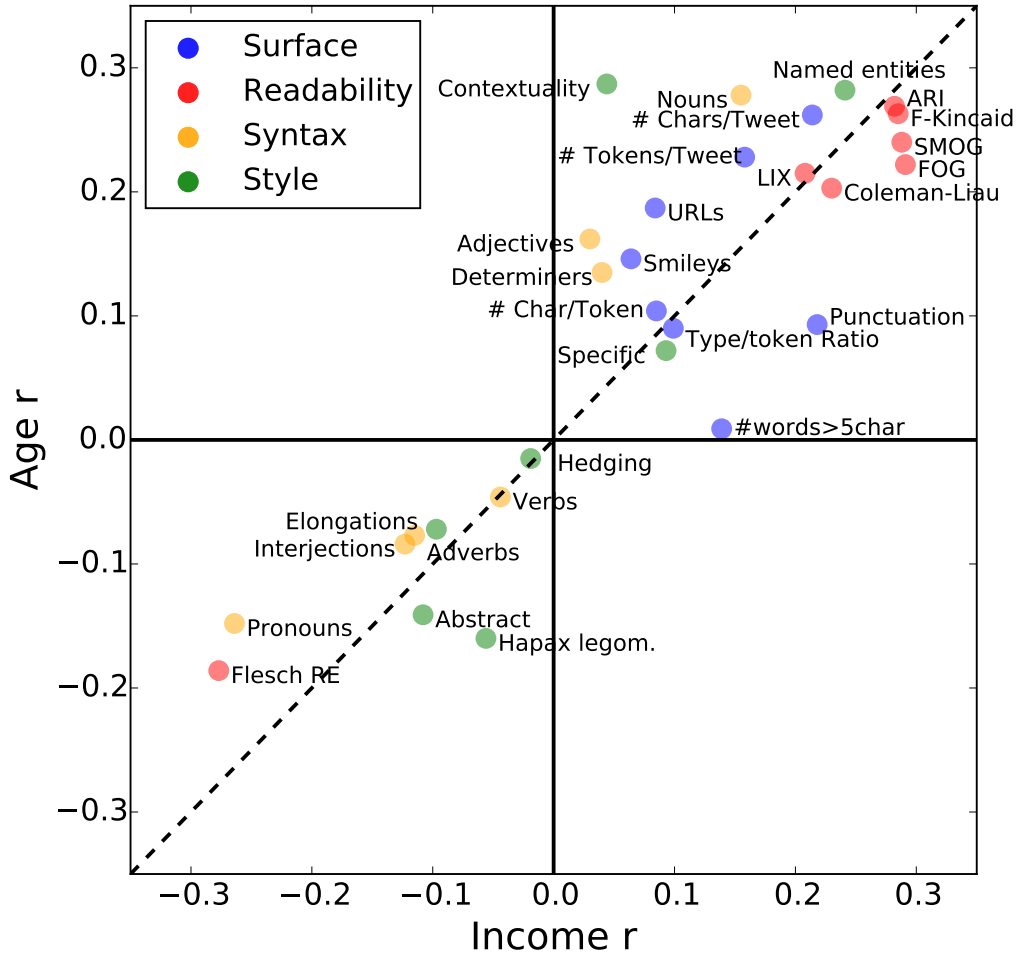
Syntactic Features

- proportions of individual Part-of-Speech tags (universal tagset)

Style Features

- F-score – *contextuality* (Heylighen and Dewaele, 2002)
- Ratio of Named Entities (3-classed) to words using Stanford NER
- Elongations and netspeak
- Words with numbers
- Hedge word list (Hyland, 2005)
- Abstract nouns
- Ratio of numbers to words
- Hapax legomena
- Ratio of superlatives
- Ratio of plural to singular pronouns and nouns

Feature Correlations



Predictive performance (Pearson correlation) for Income and Age. Individual points display univariate correlations (including sign) of the most predictive features, $p<0.001$

Take Aways

- User traits are related via features:
  + age $\rightarrow$ = **+** income
  ... and we expect this to hold for others
- Stylistic features offer linguistic insight
- Some features are more predictive of income than age: readability, punctuation, ...
- Other features are more associated to age than income: adjective use, contextuality, ...

Phrase Choice (AAAI 2016)

Another type of stylistic difference is phrase choice.

Paraphrases – alternative ways to convey the same information.

Paraphrase Database (PPDB) 2.0 (Pavlick et al. 2015 ACL) contains 100M automatically derived paraphrase pairs with type and confidence.

Data

We compute paraphrase preference over the following traits and corpora:

- Gender
  male vs. female
  67,337 Twitter users, 104M tweets
- Age
  $\geq$ 30 years old vs. $<$ 25 y.o.
  3,865 Twitter users, 690k tweets
- Occupational Class
  low skill (5–9) vs. high skill (1–2)
  3,909 Twitter users, 5M tweets

Method

Method for quantifying is straightforward:

$$\text{Gender}(w) = \log\left(\frac{\text{Female}(w)}{\text{Male}(w)}\right) \tag{1}$$

Within a paraphrase pair (w_1, w_2), the difference $\text{Gender}(w_1) - \text{Gender}(w_2)$ is the stylistic distance.

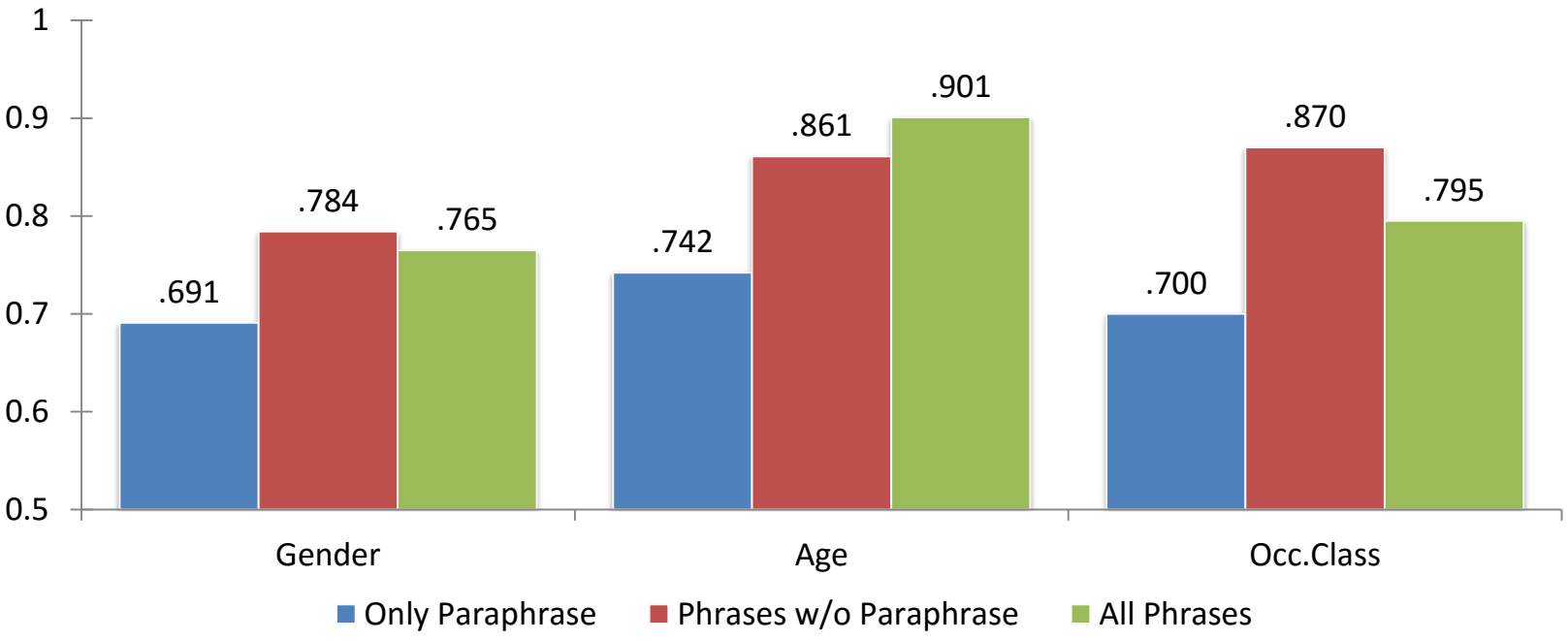
We use only equivalent paraphrases of 1–3 grams

Examples

Female	$\geqslant$ 30 y.o.	High Occ. Class
charming (-0.58)	delightful (-0.96)	gratifying (-0.33)
delightful (-0.49)	splendid (-0.55)	enjoyable (-0.30)
gratifying (0.03)	enjoyable (-0.46)	delightful (-0.17)
splendid (0.07)	gratifying (-0.04)	pleasant (-0.06)
good (0.12)	pleasant (0.01)	charming (-0.04)
pleasing (0.18)	charming (0.03)	splendid (-0.01)
nice (0.24)	pleasurable (0.13)	good (0.08)
pleasurable (0.24)	nice (0.53)	nice (0.17)
pleasant (0.33)	good (0.53)	pleasing (0.21)
enjoyable (0.38)	pleasing (0.80)	pleasurable (0.77)
Male	$<$ 25 y.o.	Low Occ. Class

$$\text{O} - \text{A}\ \tau = .51, \text{O} - \text{G}\ \tau = .2, \text{G} - \text{A}\ \tau = .06,$$

Prediction



Naive Bayes, 80-20 training-testing

Human Perception

Which word is more likely to be used by a **female** ?

Charming – Fascinating

Human Perception

Which word is more likely to be used by a **female** ?

Charming – **Fascinating**

Human Perception

Which word is more likely to be used by an **older** person?

Impressive – Amazing

Human Perception

Which word is more likely to be used by an **older** person?

Impressive – **Amazing**

Human Perception

Which word is more likely to be used by a person of **higher occupational class** ?

Suggestions – Proposals

Human Perception

Which word is more likely to be used by a person of **higher occupational class** ?

Suggestions – Proposals

Human Perception

Which word is more likely to be used by a **female** ?

Brutal – Fierce

Human Perception

Which word is more likely to be used by a **female** ?

Brutal – Fierce

Human Perception

Which word is more likely to be used by an **older** person?

Defensive – Protective

Human Perception

Which word is more likely to be used by an **older** person?

Defensive – Protective

Human Perception

Which word is more likely to be used by a person of **higher occupational class** ?

Humour – Wit

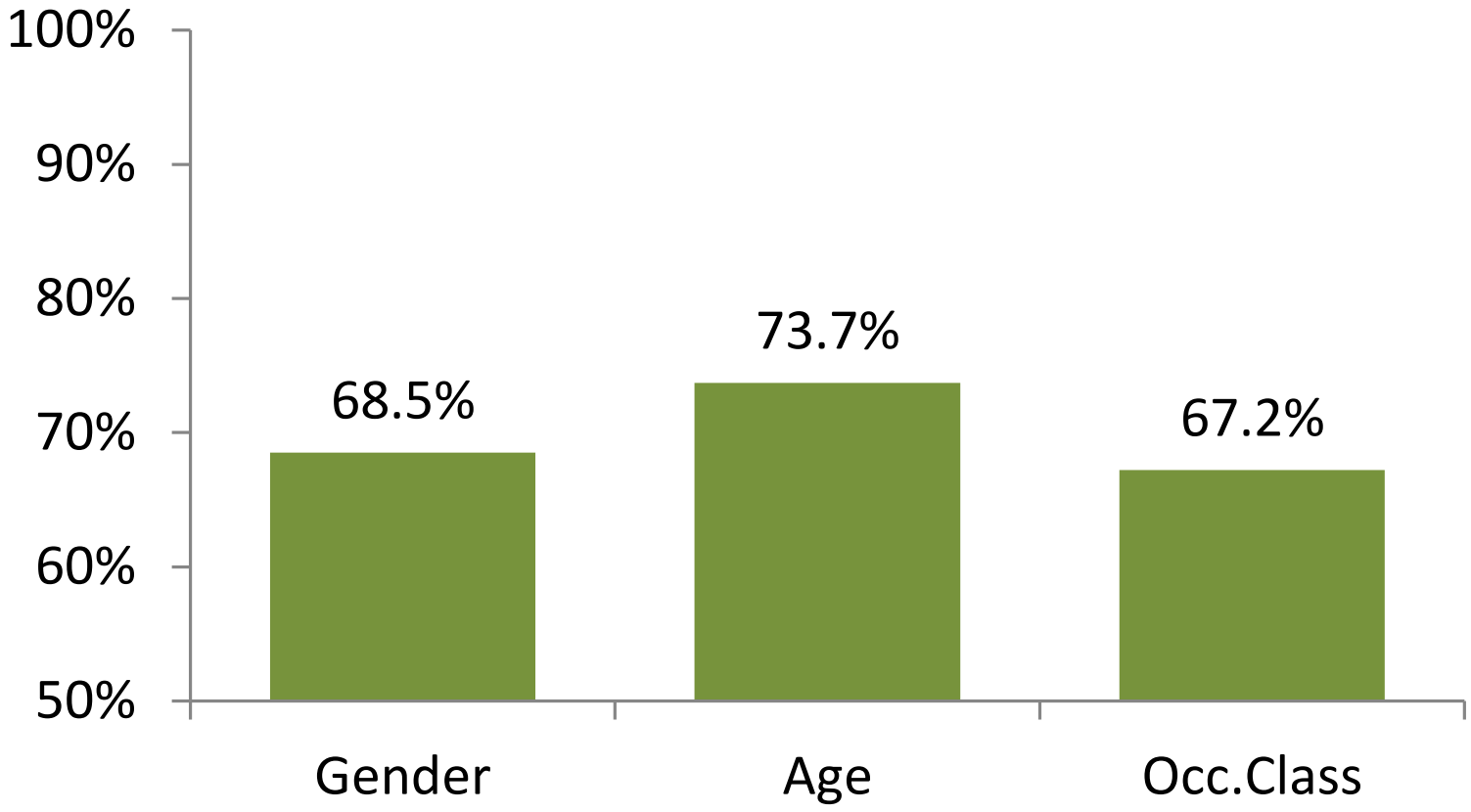
Human Perception

Which word is more likely to be used by a person of **higher occupational class** ?

Humour – Wit

Human Perception

amazon mechanical turk™
Artificial Artificial Intelligence



Take Aways

Stylistic features have important applicability

Paraphrase choice still contains valuable information (albeit less than topical)

Choices match human intuitions to some extent...

Real vs. Perceived (SPPS 2016)

...but for user attributes, not always:

Real vs. Perceived

Real traits: influence how people are going to behave

Perceived traits: influence how *other* people are going to behave

Two factors influence perception:

1. the actual behavior and cues;
2. the traits of the perceiver.

Experimental Setup

20 Tweets/user

9 ratings/user

Real traits known in advance through self-reports

Forced choice guess

Self-rated confidence (1-5)

Each rater took a demographic survey and a number of psychological questionnaires.

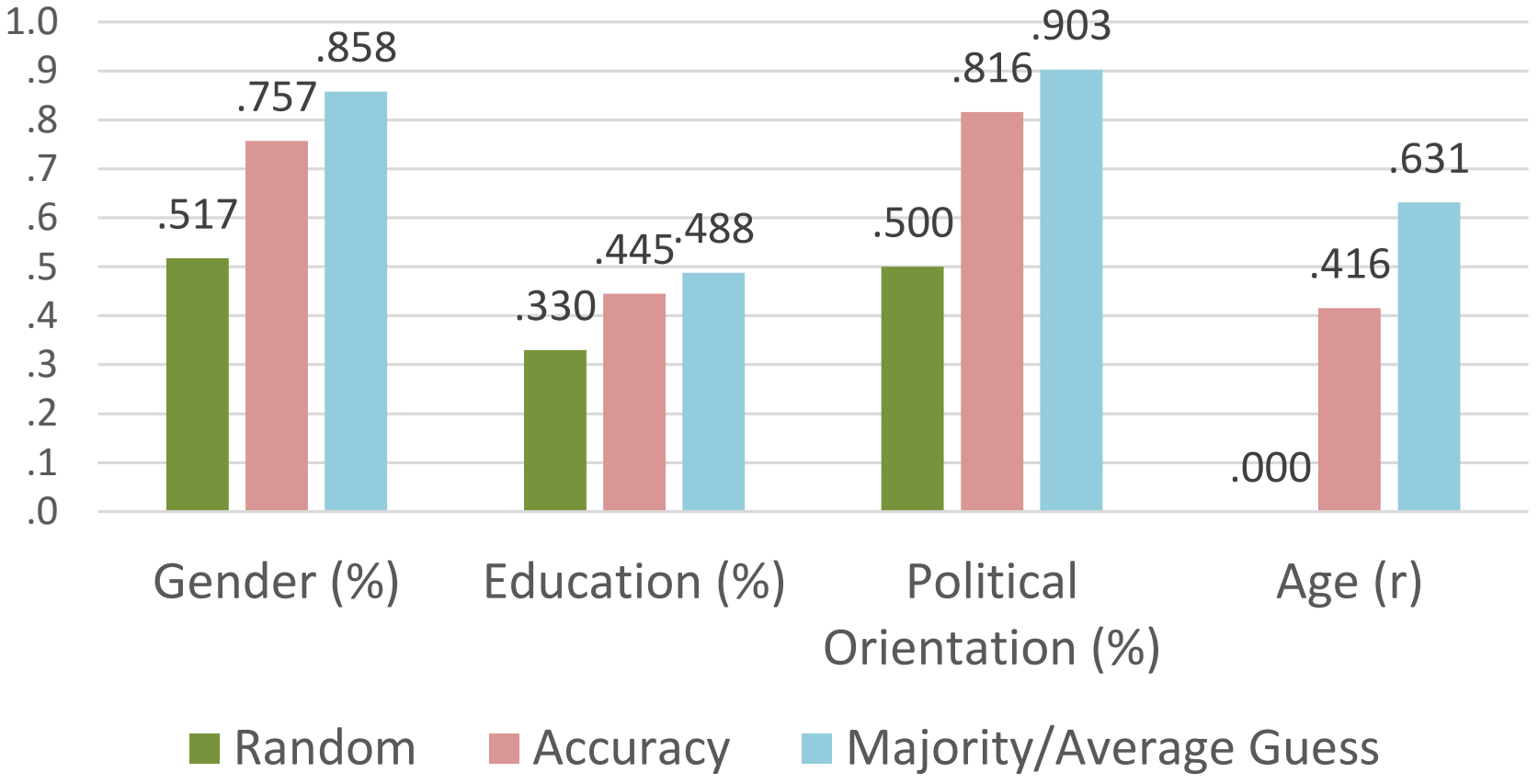
This way we isolate the textual cues from any other profile related cues (screen name, profile pic, etc)

Data Set

Trait	**Outcome**	**#Users**	**#Raters**
Gender	M/F	2607	1083
Age	Integer	1066	737
Education	Adv/BSc/HS	900	481
Political Orientation	Lib/Cons	2500	943

Data set statistics.

Guess Accuracies



People are usually correct.

Inaccurate Gender Stereotypes (SPPS 2016)

Words more likely to be associated with **females** among male authors

Words more likely to be associated with **males** among female authors

The size of the word is the strength to which they're inaccurate stereotypes i.e. 'love' is more likely to mislead people in guessing female compared to 'wonderful.'

Stereotypical Differences in Gender Perception

Rank	**Manual Label**	**Topic** (most freq.words)	**Perc**	**Real**	**Cont**
1	Politics I	nation, held, rally, defend, supporters	.372	.281	**.176**
2	Sports I	players, teams, crowds, athletes, clubs	.370	.284	**.171**
3	Sports II	training, team, field, coach, career	.323	.246	**.148**
8	Politics II	former, leader, chief, vice, minister	.316	.244	**.142**
15	Tech	title, tech, stats, division, technical	.314	.258	**.129**

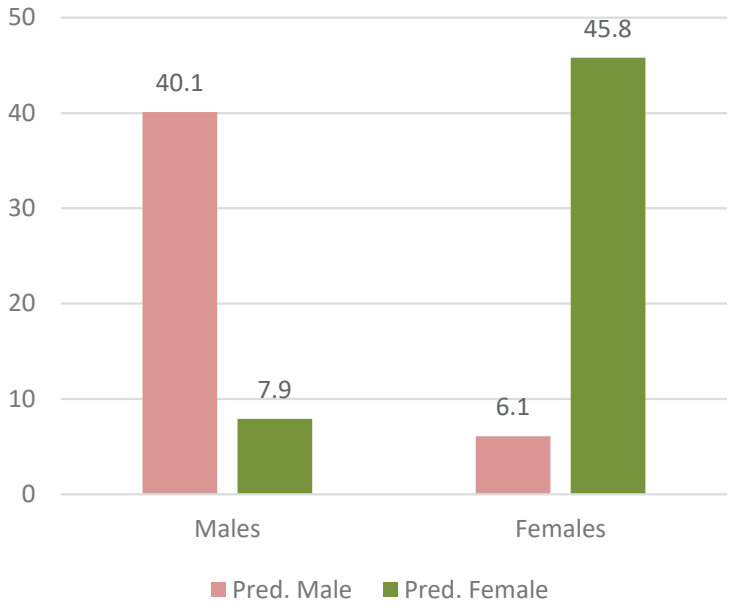
Word2Vec word clusters ('topics') correlated with the misperception of a female being male. Values show univariate correlations to **Perc** – percentage #male ratings, **Real** – self-reported male, **Cont** – Perc controlled for Real

Differences in Confidence

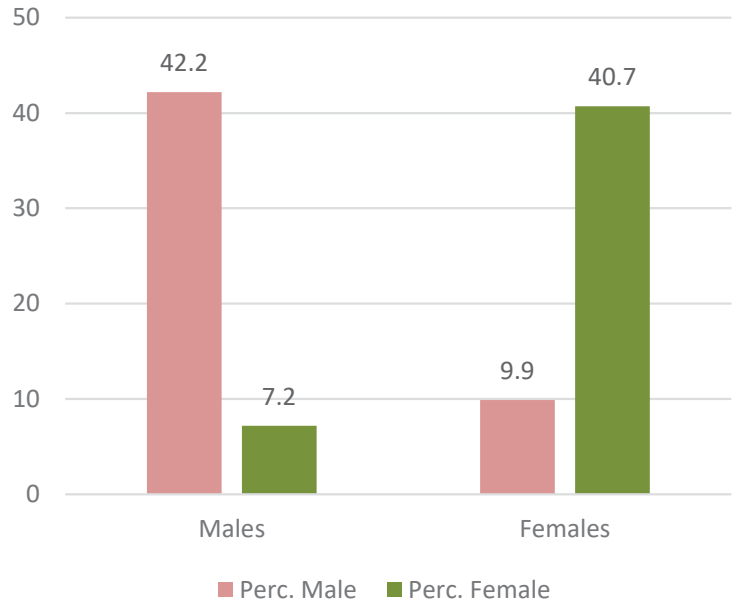
Rank	**Manual Label**	**Topic** (most freq.words)	**Conf**	**Real**	**Cont**
1	Family I	sibling, flirted, married, husband, wife	(.028)	(.071)	**.240**
3	Family II	brother, mom, daddy, daughter, sister	(.127)	.241	**.214**
6	Cosmetics	hair, blonde, shave, eyebrows, dye	.163	.182	**.199**
	. . .				
998	Astrology	horoscope, zodiac, gemini, taurus, virgo	-.269	(.087)	**-.288**
999	Planning	agenda, planning, activities, schedule	-.285	(.020)	**-.289**
1000	Adjectives	wiser, easier, shittier, happier, worse	-.277	(.081)	**-.295**

Word2Vec word clusters ('topics') correlated with the self-rated confidence for predicting gender. Values show univariate correlations to **Conf** – confidence, **Real** – ground truth, **Cont** – Conf controlled for Real

Inaccurate Gender Stereotypes (ACL 2016)



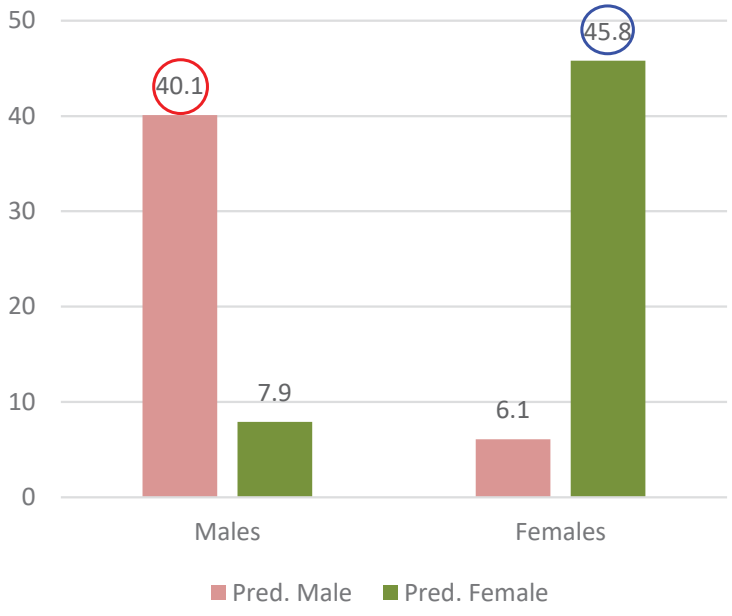
Model predictions.



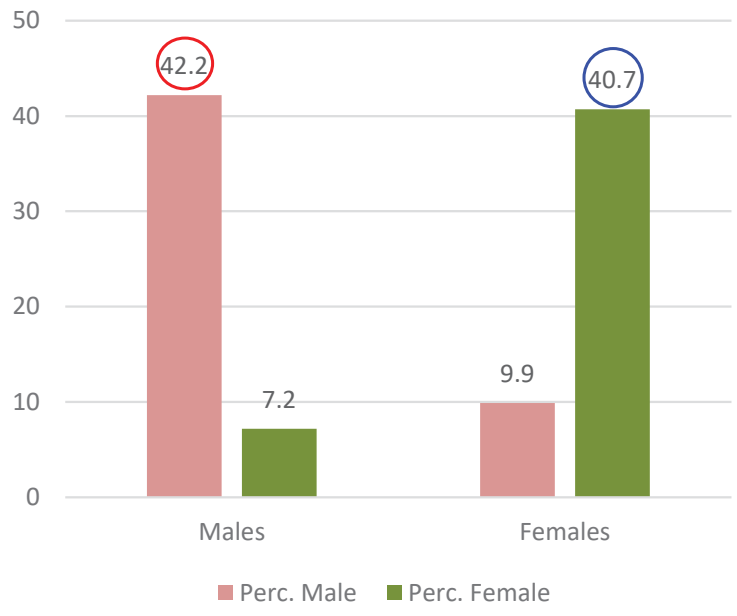
Human guesses.

Model trained on >10,000 users with self-reported gender.

Inaccurate Gender Stereotypes



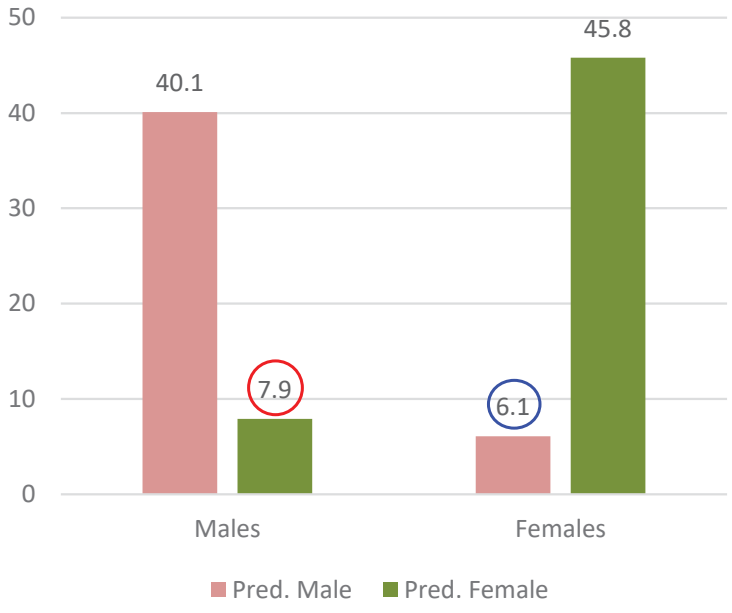
Model predictions.



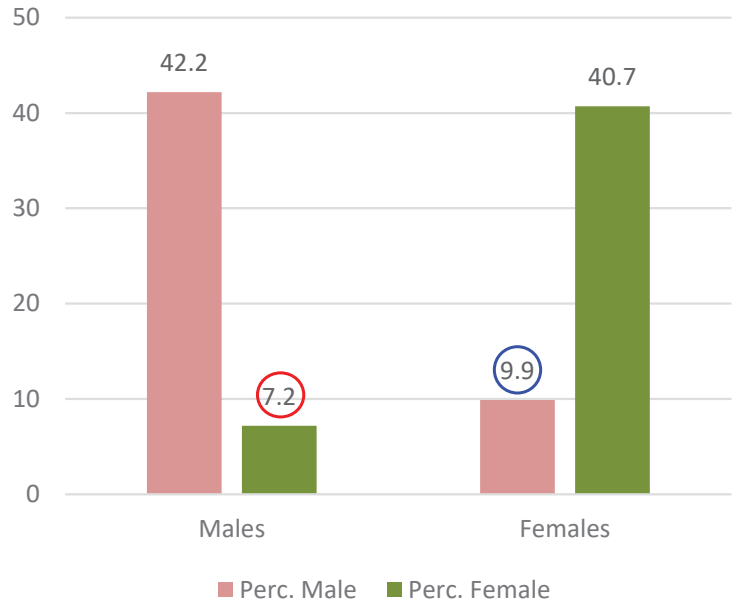
Human guesses.

The accuracies for correct predictions are reversed!

Inaccurate Gender Stereotypes



Model predictions.



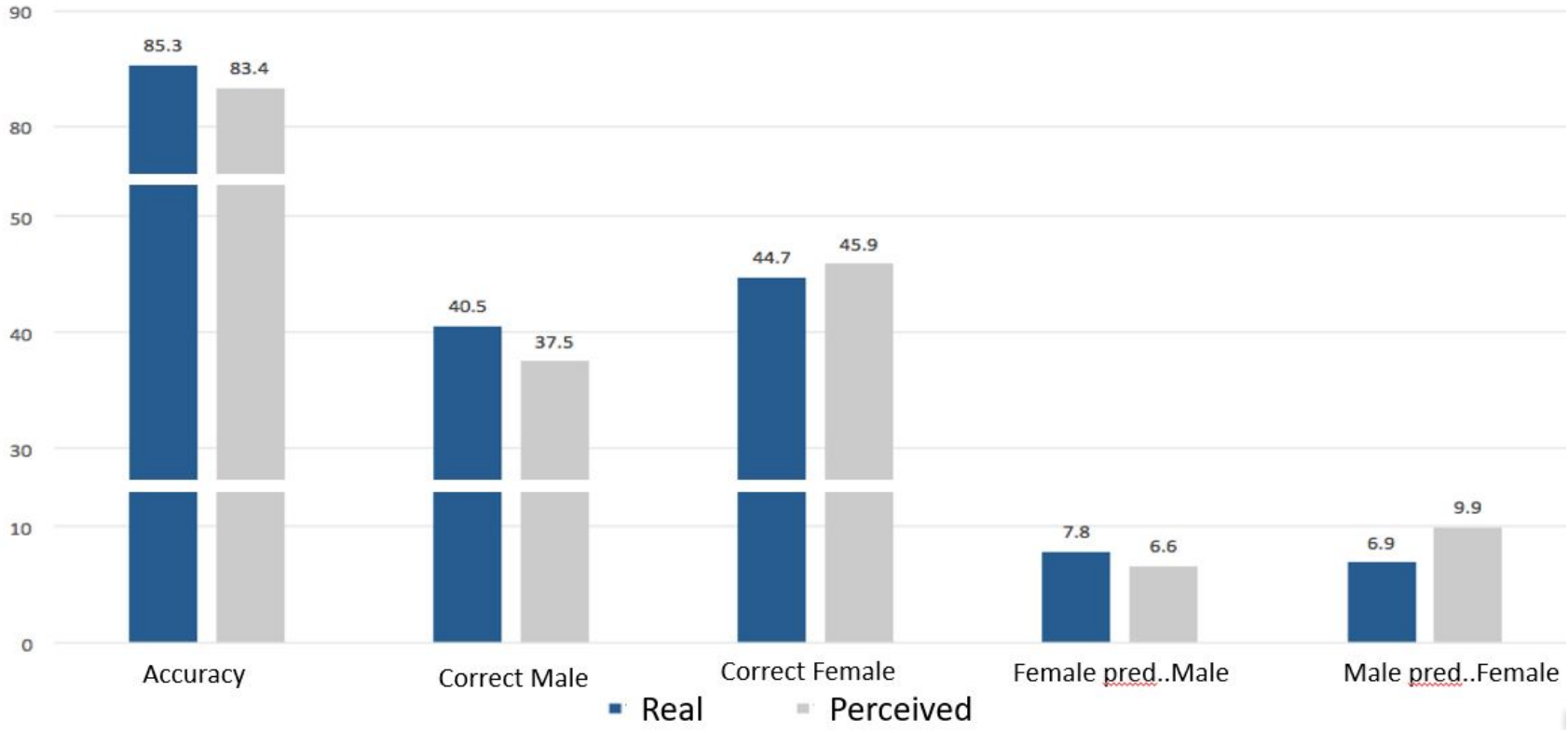
Human guesses.

The accuracies for incorrect predictions are also reversed!

Inaccurate Gender Stereotypes

Trained two models on the same data:

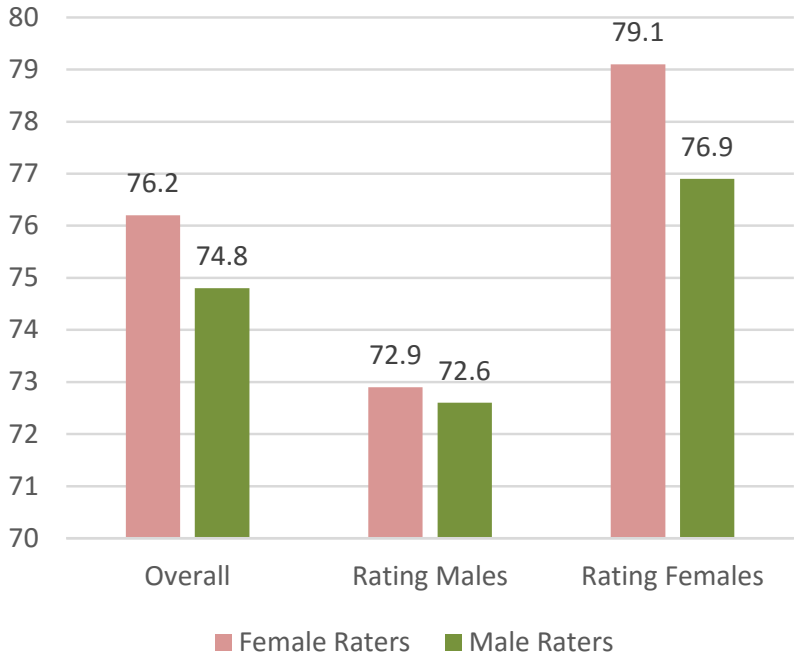
- perceived labels
- real labels



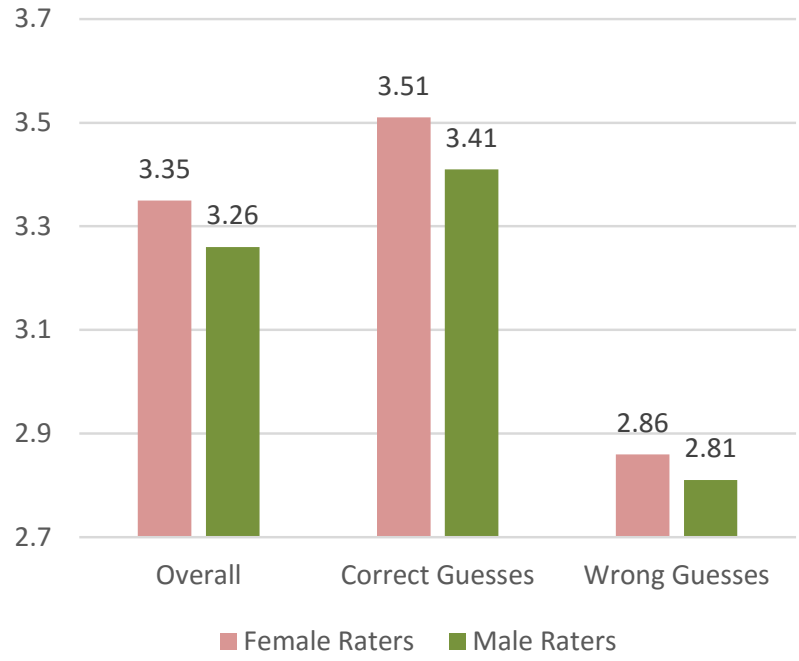
Training on perceived traits introduces a systematic bias!

User Attributes

The impact of the **rater** traits in gender prediction



Accuracy



Confidence

Take Aways

With user traits, **Real ≠ Perceived**.

We can measure stereotypes in the wild.

Annotators' own traits play an important role.

Future Directions

We need to better understand:

- stylistic differences
- difference between reality and perception of traits
- other psychological traits: e.g. empathy, trust
- longitudinal change within individuals
- how to combine multiple modalities i.e. text, images, network
- the goal of the personalization e.g. trait match or complement

Images and Personality (ICWSM 2016)

Task: Predict Big Five personality traits from a single profile picture

Data: 66.000 Twitter Profile Images $\longleftrightarrow$ tweet predicted personality

Features: 85 interpretable

- aesthetics
- composition
- no. faces
- facial presentation (emotions, smile, glasses)

Controlling for age and gender

Accuracy: r ~ .17 (single photo) vs. r > .30 text predictions (thousands of features from hundreds of posts)

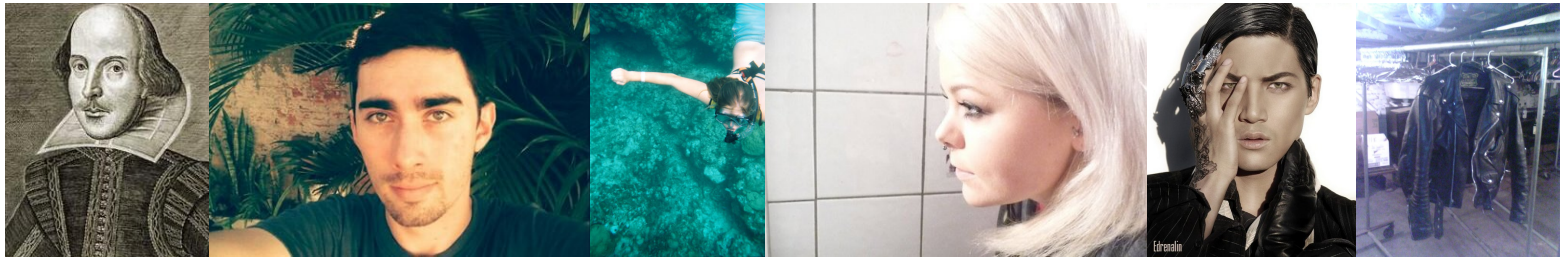
Images and Personality (ICWSM 2016)

Openness to experience

- curious, artistic, intellectual

Online:

- good aesthetic quality
- negative in colour and facial emotions
- non-conformist (grayscale, non-faces)



Images and Personality (ICWSM 2016)

Conscientiousness

- preference for planned behaviour, conformance to what is expected, hard working

Online:

- no relation to aesthetic quality
- conforming to what is expected: large single face
- happiest through facial emotions
- happiness not expressed through colors

Modelling complex spatio-temporal dynamics

User generated data changes with, and as a response to spatial and temporal proprieties

We can model these changes for practical tasks:

- discovering topic evolution and recurrence (**EMNLP 13**)
- uncover topic emergence (**LREC 16**)
- studying regional variation
- useful for text categorisation and retrieval

Predicting real-world outcomes

Nowcasting is a growing area for non-obtrusive social measurement:

- online content can potentially access a larger and more representative part of the population
- timely and regional information about events
- small cost when data access and expertise are in place

Nowcasting:

- political voting intention (ACL 13)
- economic indices (ACL 14 ws)

Location Based Social Networks and Urban Computing

Modelling future check-in behaviour (**WebScience 13**)

Characterising and clustering cities based on venues and activity patterns (**KDD 13 ws**)

Text Mining

Sentiment analysis

- modeling valence/arousal on Facebook posts (**NAACL 16 ws**)
- identifying sentiment modifiers (EMNLP 15 ws)

Political science applications:

- news biases (LREC 16)
- debate transcripts (Penn Current, wwbp.org/blog)

Summary of Research Abstracts

www.preotiuc.ro

Thank you!

Read more...

FORTUNE
QUARTZ
BBC SCIENCE • TECHNOLOGY • FUTURE
FOCUS
FAST COMPANY
Mashable
WIRED
The Telegraph
The Washington Post
Slate
SCIENCES
ET A
LE FIGARO
RT
New Scientist
SCIENTIFIC AMERICAN
the Atlantic
theguardian
BBC NEWS
vocativ
Inc.
VICE
Forbes
metro
marie claire
la Repubblica.it