

User attribute prediction from social media posts

Daniel Preoțiuc-Pietro

Positive Psychology Center
University of Pennsylvania

 Penn | World Well-Being Project

13 January 2016

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)
- ▶ Impact (Lamos et al. 2014 EACL)
- ▶ Political Orientation (Volkova et al. 2014 ACL)
- ▶ Mental illness (Coppersmith et al. 2014 ACL)

Motivation

User attribute prediction from text is useful:

- ▶ Sentiment analysis (Volkova et al. 2013 EMNLP)
- ▶ Text classification (Hovy 2015 ACL)
- ▶ Audience profiling (Yang et al. 2015 EMNLP)
- ▶ Marketing
- ▶ embedding to downstream tasks (e.g. controlling for demographic factors)
- ▶ ... and social science research

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, pediatricist, 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 \times 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

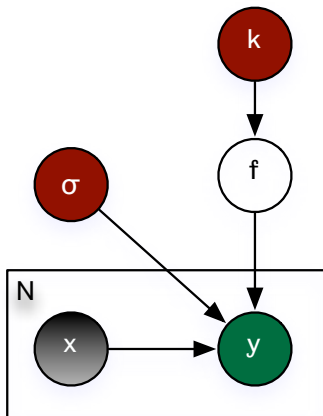
Elegant and powerful framework, with growing popularity in machine learning and application domains

Gaussian Process Graphical Model View

$$\mathbf{f} \sim \mathcal{GP}(m, k)$$

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

- ▶ $f : \mathcal{R}^D \rightarrow \mathbb{R}$ 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

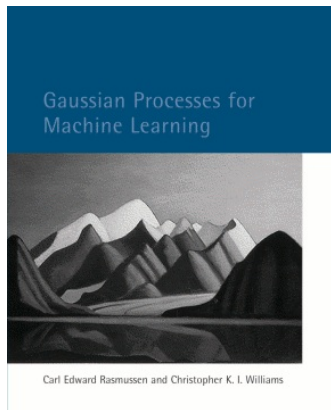
We learn 9 one-vs-all binary classifiers

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

Gaussian Process Resources

Free book:

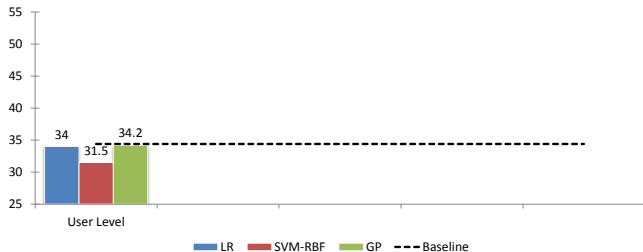
<http://www.gaussianprocess.org/gpml/chapter>



Gaussian Process Resources

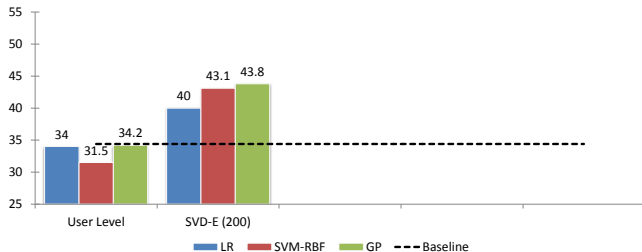
- ▶ GPs for Natural Language Processing tutorial (ACL 2014)
<http://www.preotiuc.ro>
- ▶ GP School in Sheffield and roadshows in Kampala, Pereira, Nyeri, Melbourne <http://ml.dcs.shef.ac.uk/gpss/>
- ▶ Annotated bibliography and other materials
<http://www.gaussianprocess.org>
- ▶ GPy Toolkit (Python)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

Prediction



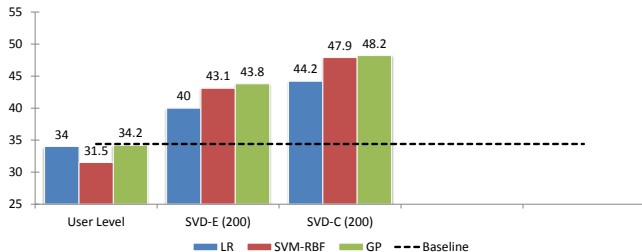
Stratified 10 fold cross-validation

Prediction



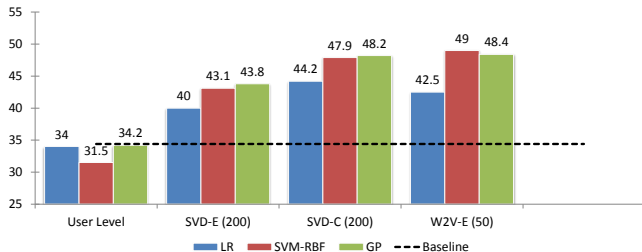
Stratified 10 fold cross-validation

Prediction



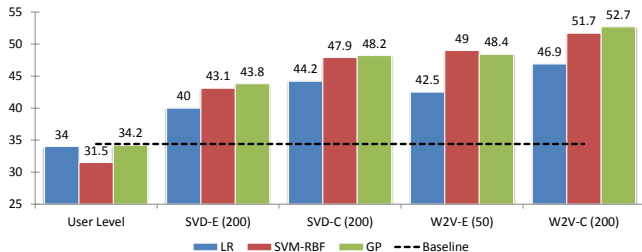
Stratified 10 fold cross-validation

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

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

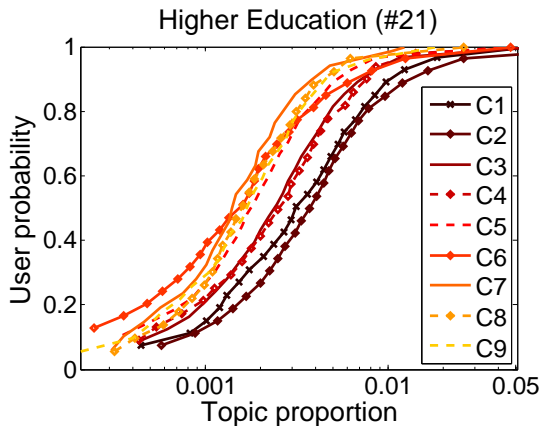
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

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

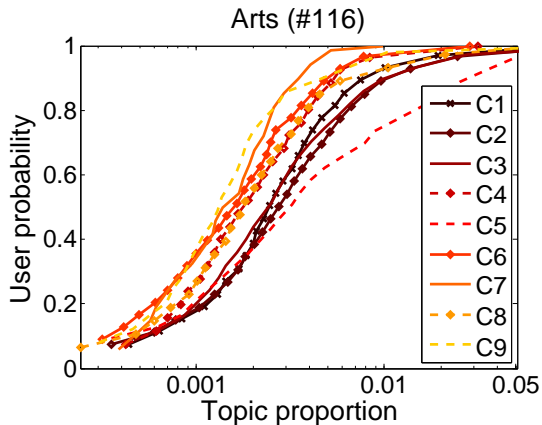
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

Feature Analysis - Cumulative Density Functions



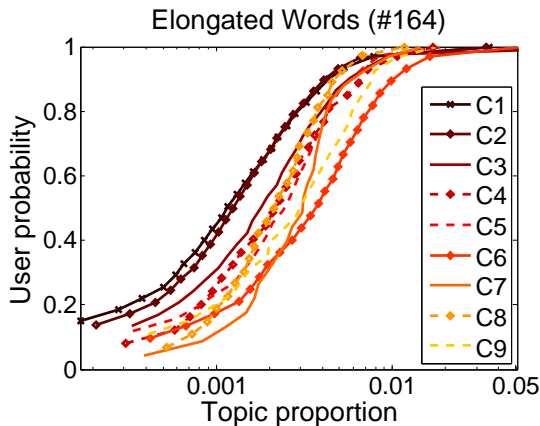
Topic more prevalent \rightarrow CDF line closer to bottom-right corner

Feature Analysis - Cumulative Density Functions



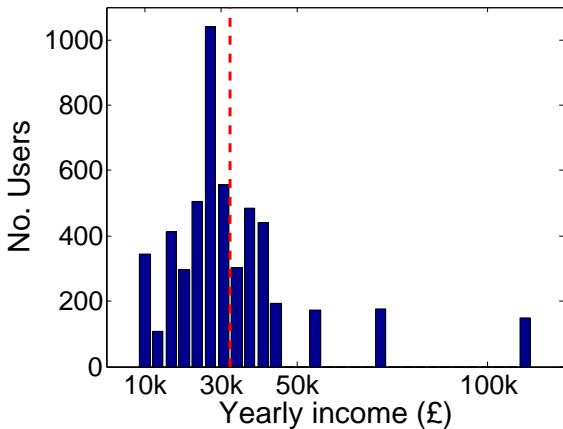
Topic more prevalent \rightarrow CDF line closer to bottom-right corner

Feature Analysis - Cumulative Density Functions



Topic more prevalent \rightarrow CDF line closer to bottom-right corner

Income prediction



We approach the task as regression.

Emotion Features

- ▶ Sentiment:
 - ▶ positive
 - ▶ neutral
 - ▶ negative
- ▶ Emotions:
 - ▶ anger
 - ▶ disgust
 - ▶ fear
 - ▶ joy
 - ▶ sadness
 - ▶ surprise

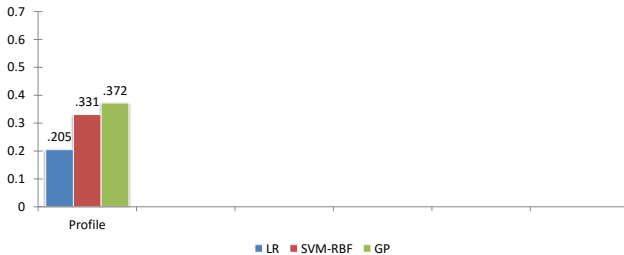
Predicted User Psycho-Demographic Features

- ▶ gender
male, female
- ▶ age
- ▶ political
independent, conservative, liberal, unaffiliated
- ▶ intelligence
> average, average, < average, >> average, << average
- ▶ relationship
married, in a relationship, single, other
- ▶ ethnicity
Asian, African American, Indian, Hispanic, Other, Caucasian
- ▶ education
bachelor, graduate, high school
- ▶ religion
Christian, Jewish, Muslim, Hindu, unaffiliated, other
- ▶ children
yes, no

Predicted User Psycho-Demographic Features

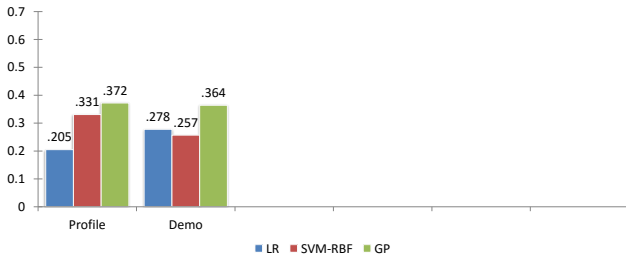
- ▶ income
below average, above average, very high
- ▶ life satisfaction
satisfied, dissatisfied, very satisfied, very dissatisfied, neither
- ▶ optimism
optimist, pessimist, extreme optimist, extreme pessimist, neither
- ▶ narcissism
agree strongly, agree, disagree, disagree strongly, neither
- ▶ excited
agree strongly, agree, disagree, disagree strongly, neither
- ▶ anxious
agree strongly, agree, disagree, disagree strongly, neither

Prediction



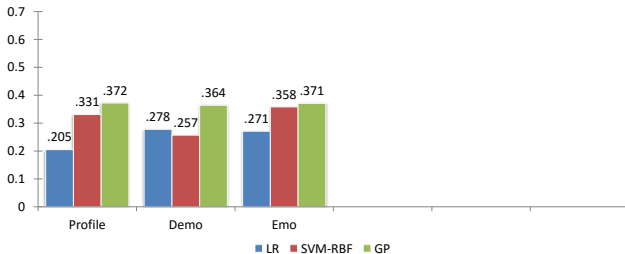
10 fold cross-validation

Prediction



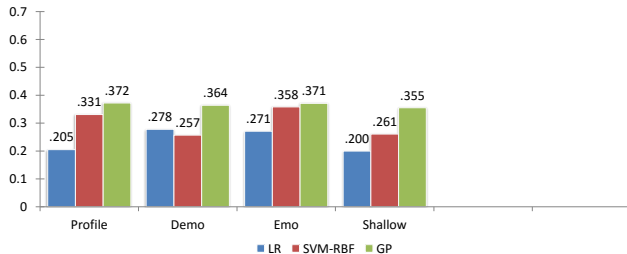
10 fold cross-validation

Prediction



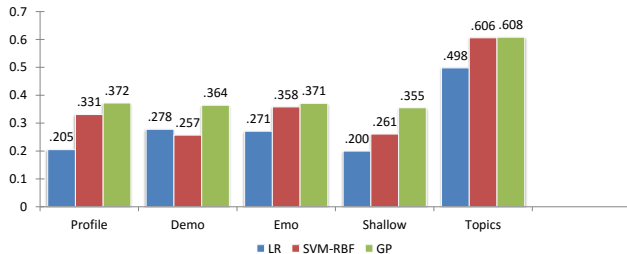
10 fold cross-validation

Prediction



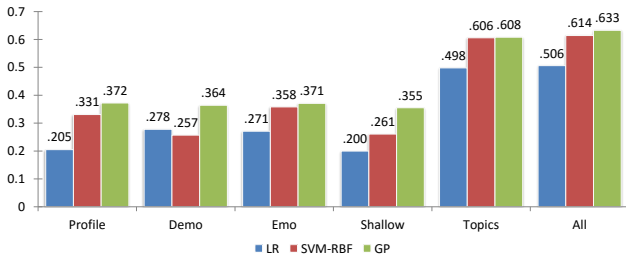
10 fold cross-validation

Prediction



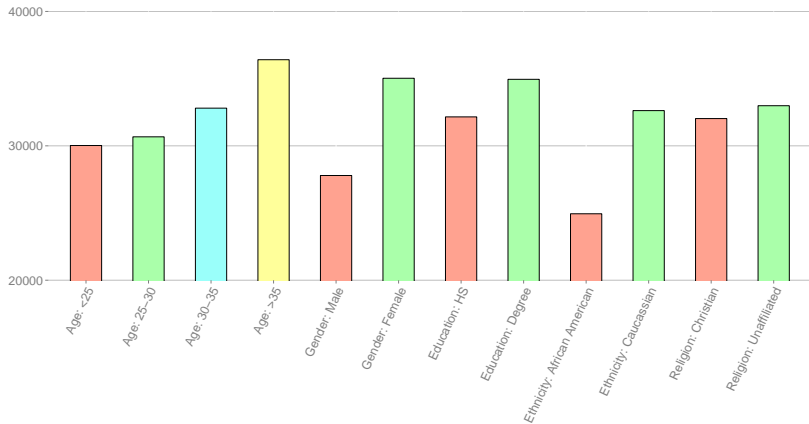
10 fold cross-validation

Prediction

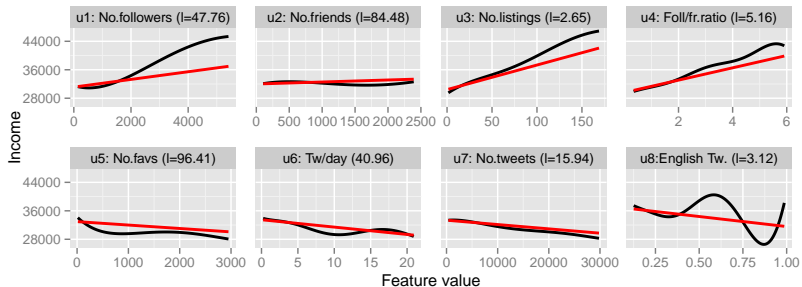


10 fold cross-validation

Psycho-Demographic Features

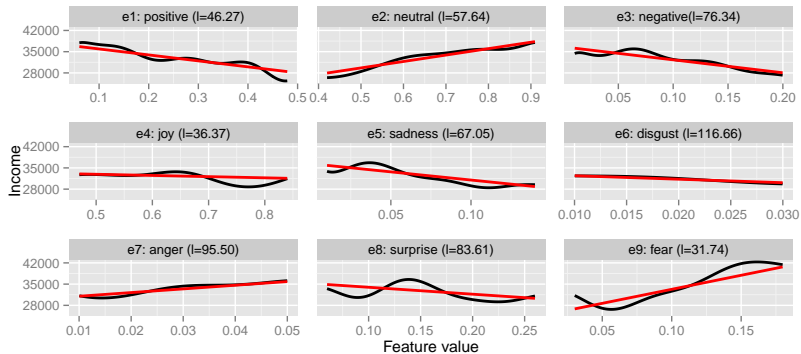


Profile Features



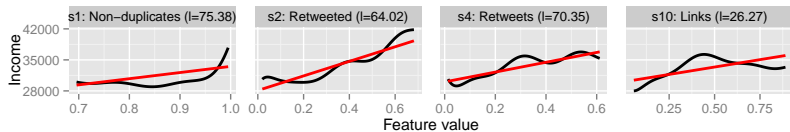
Linear fit, GP fit

Emotions



Linear fit, GP fit

Shallow Textual Features



Linear fit, GP fit

Topics



Linear fit, GP fit

Take Aways

User socio-economic status influences language 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

However...

Most text prediction methods uncover **topical** differences

Not useful for many practical applications that adapt to traits:

- ▶ machine translation (Mirkin et al. 2015 EMNLP)
- ▶ conversational agents
- ▶ tutoring systems

We need to be aware of **style** differences, rather than topical

One type of stylistic difference is phrase choice

Text-to-text Translation

Female or Male?

@USER A teeny tiny cute coffee shop (and a cheap fry-up style cafe). I'm intrigued...

@USER A small tiny clever coffee shop (and a cheap fry-up style cafe). I'm intrigued...

Text-to-text Translation

Female or Male?

@USER A teeny tiny cute coffee shop (and a cheap fry-up style cafe). I'm intrigued...

@USER A small tiny clever coffee shop (and a cheap fry-up style cafe). I'm intrigued...

Text-to-text Translation

Over 30 years old or under 25 years old?

RT @USER : how your body responds to exercise over time URL

RT @USER : how your body answers to workout over time URL

Text-to-text Translation

Over 30 years old or under 25 years old?

RT @USER : how your body responds to exercise over time URL

RT @USER : how your body answers to workout over time URL

We study three traits:

- ▶ Gender
male vs. female
67,337 Twitter users, 104M tweets
- ▶ Age
 ≥ 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

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

We use only equivalent paraphrases of 1–3 grams

Method is straightforward:

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

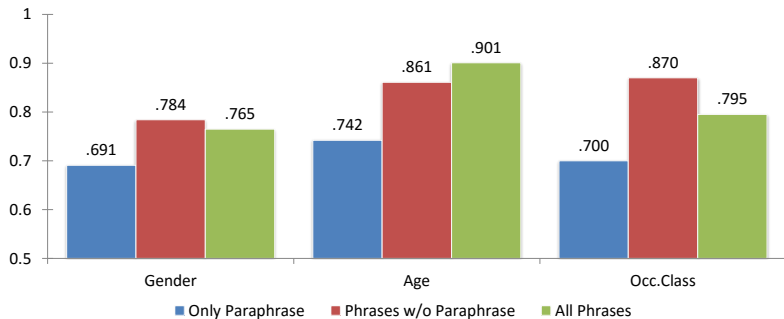
Within a paraphrase pair (w_1, w_2) , the difference $\text{Gender}(w_1) - \text{Gender}(w_2)$ is the stylistic distance.

Examples

Female	≥ 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

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

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** ?

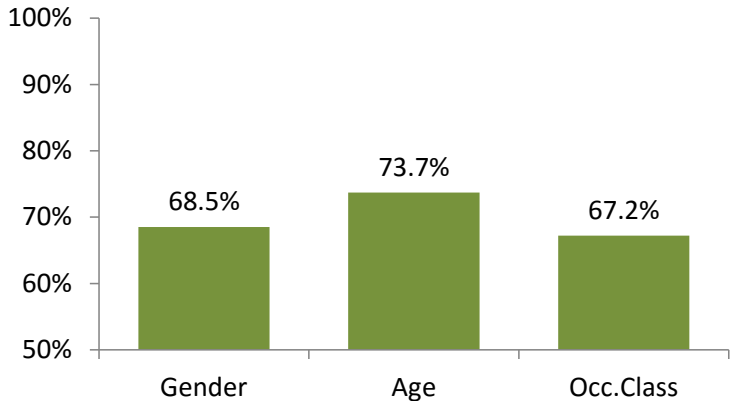
Humour – Wit

Human Perception

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

Humour – **Wit**

Human Perception

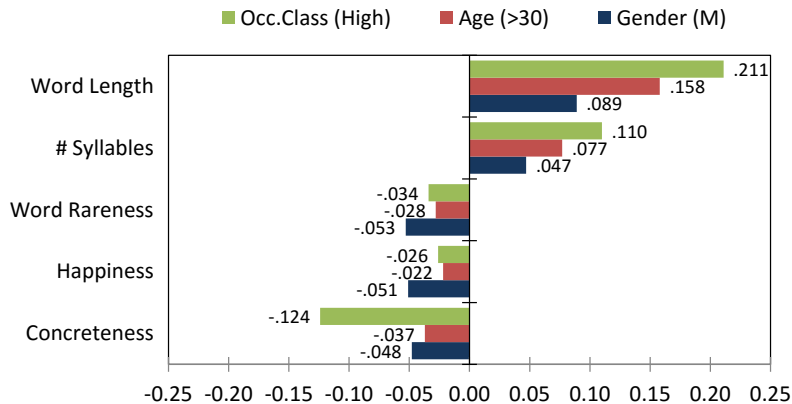


Linguistic Theories

Study which attributes of words in a pair are more preferred by one group:

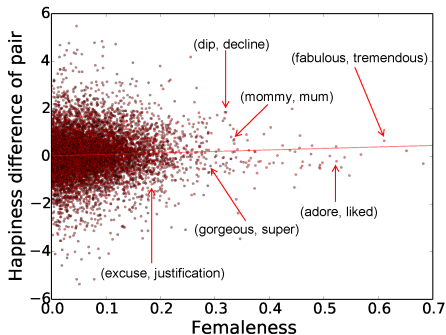
- ▶ Word Length
- ▶ Number of Syllables
- ▶ Word Rareness
Compared to a reference corpus of 400M tweets
- ▶ Perceived happiness
10k words, suicide (0.15) → bacon (0.70) → laughter (1)
- ▶ Concreteness
40k words, spirituality (1) → tiger (5)

Linguistic Theories



Correlation coefficients between paraphrase pair word differences and user group differences in usage.

Gender and Perceived Happiness



Scatter plot and best fit line between paraphrase use differences
and perceived word happiness
Each dot is a paraphrase pair
The first work is more used by females

Take Aways

Stylistic features have important applicability

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

Choices match human intuitions

Collaborators



Thank you!

Resources

Daniel Preoțiuc-Pietro, Vasileios Lampos, Nikolaos Aletras – ACL, 2015
An analysis of the user occupational class through Twitter content.

Daniel Preoțiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, Nikolaos Aletras – PLoS ONE 10(9), 2015
Studying User Income through Language, Behaviour and Affect in Social Media.

Daniel Preoțiuc-Pietro, Wei Xu, Lyle Ungar – AACL, 2016
Discovering User Attribute Stylistic Differences via Paraphrasing.

www.preotiuc.ro