An analysis of the user occupational class through Twitter content

Daniel Preoţiuc-Pietro[1] Vasileios Lampos[2] Nikolaos Aletras[2]

[1]Computer and Information Science
University of Pennsylvania

[2]Department of Computer Science
University College London

Penn | World Well-Being Project

UCL

29 July 2015

Motivation

User attribute prediction from text is successful:

- Age (Rao et al. 2010 ACL)
- Gender (Burger et al. 2011 EMNLP)
- Location (Eisenstein et al. 2011 EMNLP)
- Personality (Schwartz et al. 2013 PLoS One)
- Impact (Lampos et al. 2014 EACL)
- Political orientation (Volkova et al. 2014 ACL)
- Mental illness (Coppersmith et al. 2014 ACL)

Downstream applications are benefiting from this:

- Sentiment analysis (Volkova et al. 2013 EMNLP)
- Text classification (Hovy 2015 ACL)

However...

Socio-economic factors (occupation, social class, education, income) play a vital role in language use

(Bernstein 1960, Labov 1972/2006)

No large scale user level dataset to date

Applications:

- sociological analysis of language use
- embedding to downstream tasks (e.g. controlling for socio-economic status)

At a Glance

Our contributions:

- Predicting new user attribute: occupation
- New dataset: user $\longleftrightarrow$ occupation
- 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)

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
    - 1115 Chief Executives and Senior Officials
      Job: chief executive, bank manager
    - 1116 Elected Officers and Representatives
  - 112 Production Managers and Directors
  - 113 Functional Managers and Directors
  - 115 Financial Institution Managers and Directors
  - 116 Managers and Directors in Transport and Logistics
  - 117 Senior Officers in Protective Services
  - 118 Health and Social Services Managers and Directors
  - 119 Managers and Directors in Retail and Wholesale
- 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

5,191 users $\longleftrightarrow$ 3-digit job group

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

Data

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

Feature representation and labels available online

Raw data available for research purposes on request (per Twitter TOS)

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 clusters

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 similiarity 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

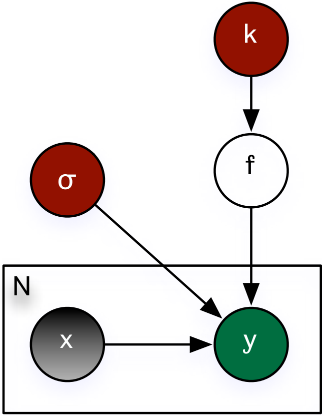
Elegant and powerful framework, with growing popularity in machine learning and application domains

Gaussian Process Graphical Model View

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

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

- $f : \mathcal{R}^D - > R$ is a latent function
- y is a noisy realisation of $f(x)$
- k is the covariance function or 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

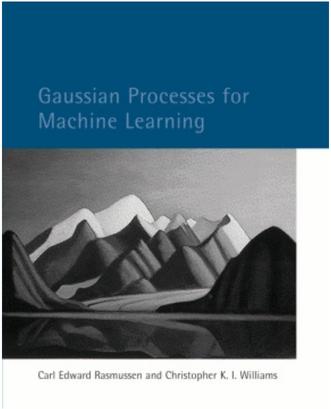
ARD kernel learns feature importance → features most **discriminative** between classes

We learn 9 one-vs-all binary classifiers

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

Gaussian Process Resources

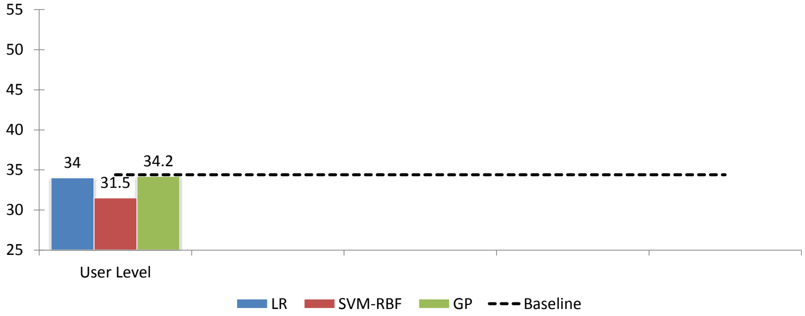
Free book:
`http://www.gaussianprocess.org/gpml/chapters/`



Gaussian Process Resources

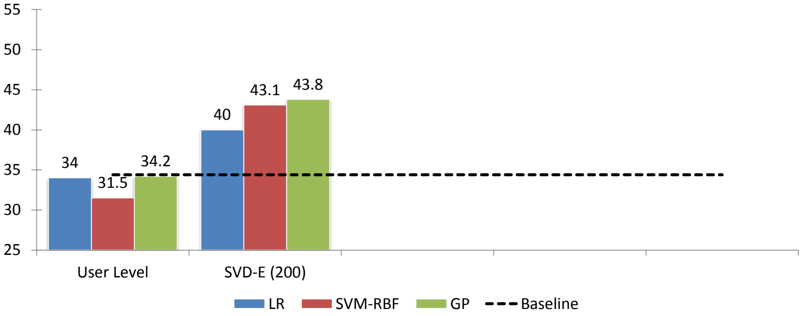
- GPs for Natural Language Processing tutorial (ACL 2014)
  `http://www.preotiuc.ro`
- GP Schools in Sheffield and roadshows in Kampala, Pereira, Nyeri, Melbourne
  `http://ml.dcs.shef.ac.uk/gpss/`
- Annotated bibliography and other materials
  `http://www.gaussianprocess.org`
- GPy Toolkit (Python)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

Prediction



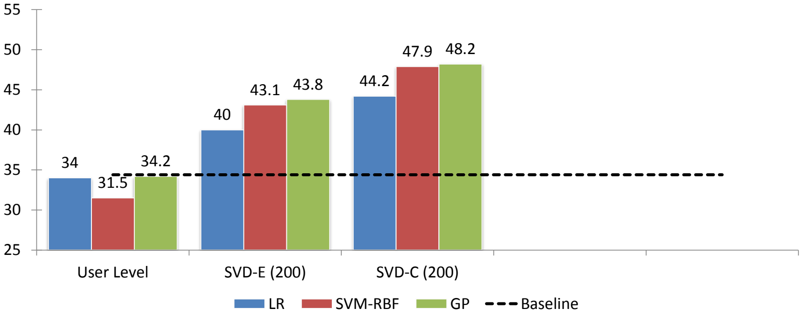
Stratified 10 fold cross-validation

Prediction



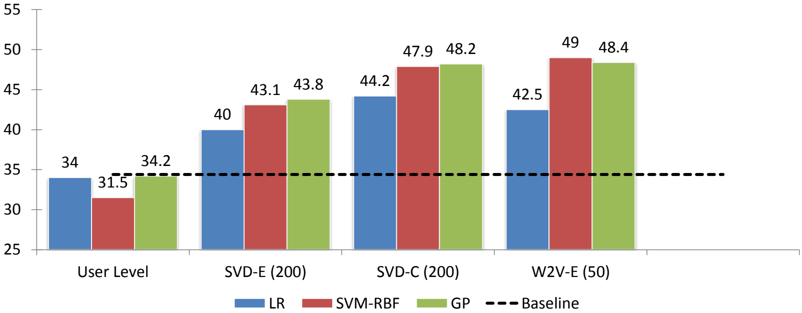
Stratified 10 fold cross-validation

Prediction



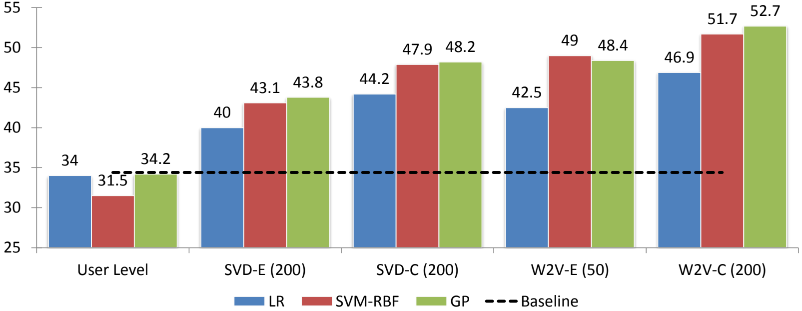
Stratified 10 fold cross-validation

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

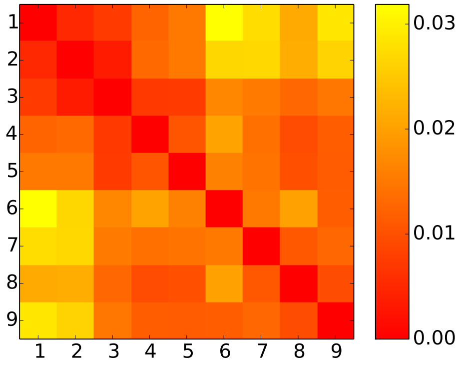
Non-linear methods (SVM-RBF and GP) significantly outperform linear methods

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

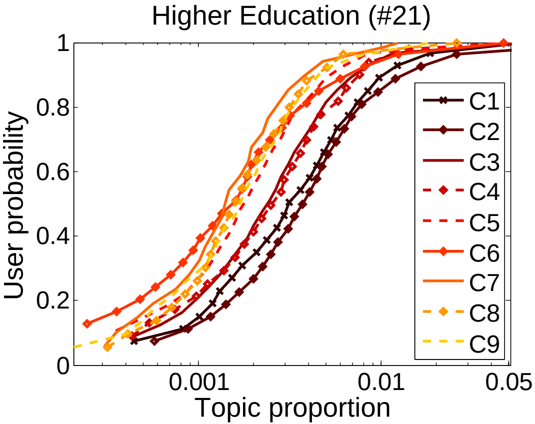
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

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

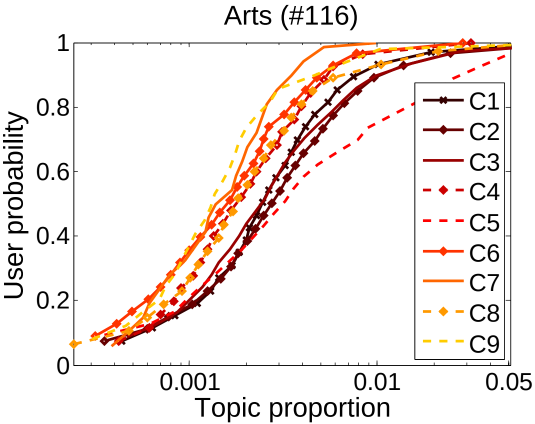
Most predictive Word2Vec 200 clusters as given by Gaussian Process ARD ranking

Feature Analysis - Cumulative density functions



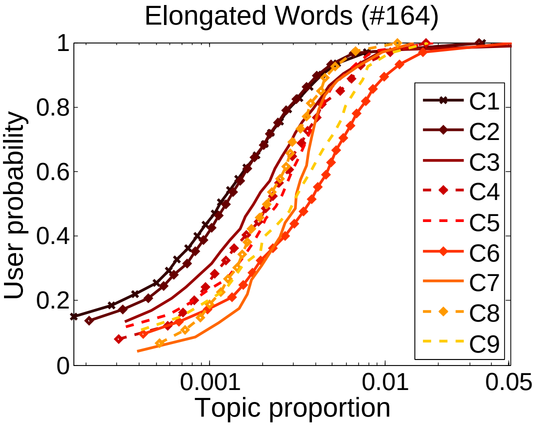
Topic more prevalent → CDF line closer to bottom-right corner

Feature Analysis - Cumulative density functions



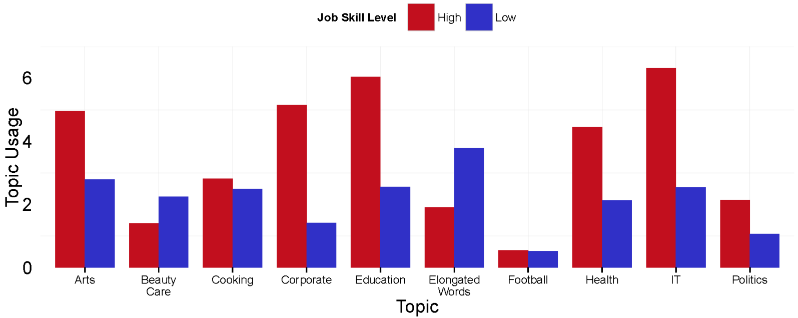
Topic more prevalent → CDF line closer to bottom-right corner

Feature Analysis - Cumulative density functions



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

Feature Analysis



Comparison of mean topic usage between supersets of occupational classes (1-2 vs. 6-9)

Take Aways

User occupation 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