Inferring the Socioeconomic Status of Social Media Users based on Behaviour & Language

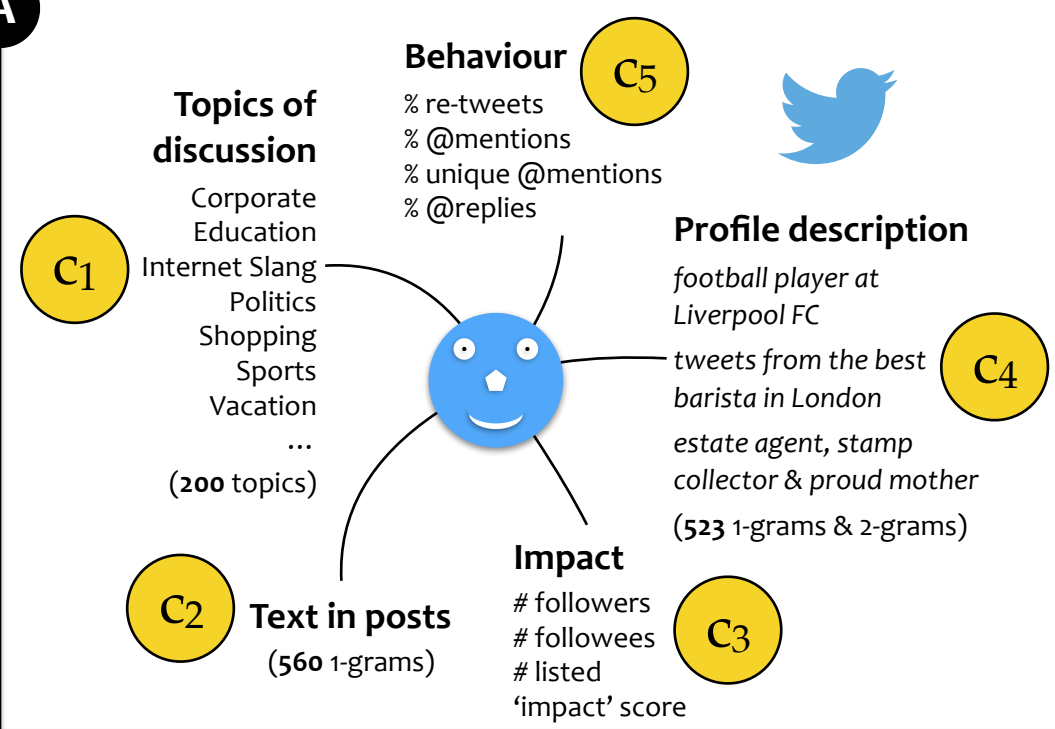
Vasileios Lampos, Nikolaos Aletras, Jens K. Geyti, Bin Zou & Ingemar J. Cox

UCL

Summary. *We present a method for determining the* ***socioeconomic status*** *of a social media (Twitter) user. Initially, we formulate a 3-way classification task, where users are classified as having an* ***upper***, ***middle*** *or* ***lower*** *socioeconomic status. A nonlinear learning approach using a composite* ***Gaussian Process*** *kernel provides a classification accuracy of* ***75%****. By turning this task into a binary classification – upper vs. medium and lower class – the proposed classifier reaches an accuracy of* ***82%****.*

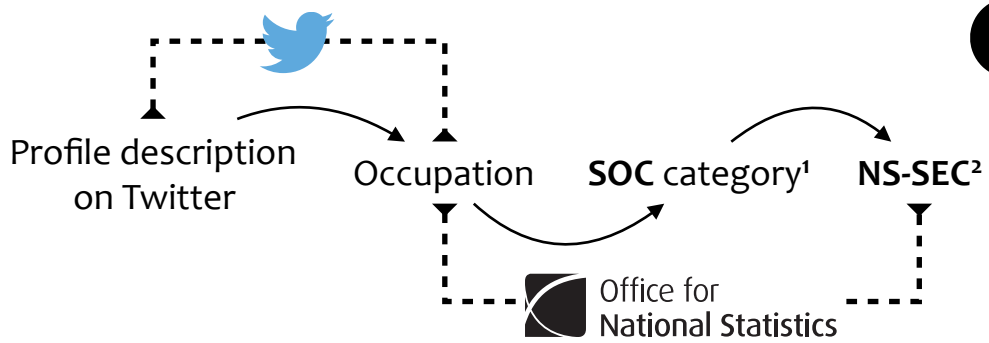
A

Twitter user attributes (feature categories)



B

How is a user profile mapped to a socioeconomic status?



1. **S**tandard **O**ccupational **C**lassification: 369 job groupings
2. **N**ational **S**tatistics **S**ocio-**E**conomic **C**lassification: Map from the job groupings in SOC to a socioeconomic status, *i.e.* {*upper*, *middle* or *lower*}

C

Data sets

T1: 1,342 Twitter user profiles, 2 million tweets, from February 1, 2014 to March 21, 2015; profiles are labelled with a socio-economic status

T2: 160 million tweets, sample of UK Twitter, same date range with T1, used to learn a set of 200 latent topics

D

Topics (word clusters) are formed by applying **spectral clustering** on daily word frequencies in **T2.**

Examples of topics with word samples

Corporate: *#business, clients, development, marketing, offices*

Education: *assignments, coursework, dissertation, essay, library*

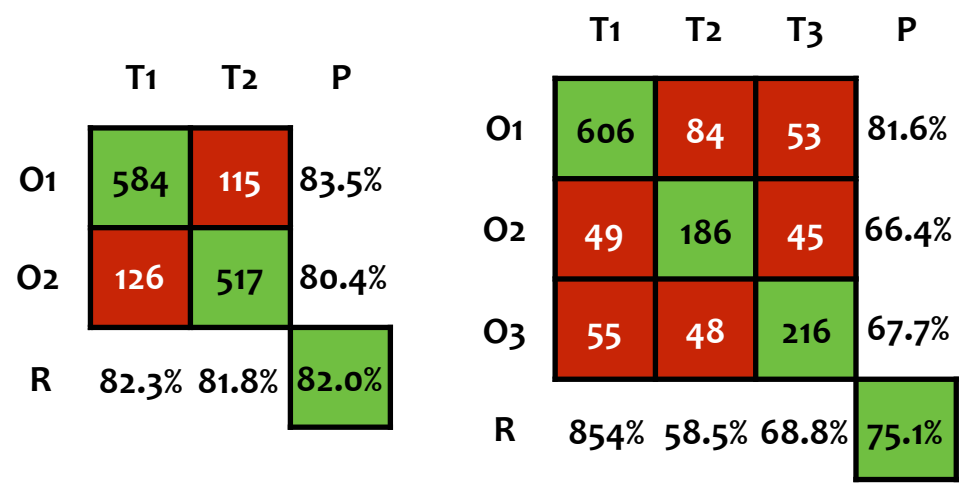
Internet Slang: *ahahaha, awwww, hahaa, hahahaha, hmmmm*

Politics: *#labour, #politics, #tories, conservatives, democracy*

Shopping: *#shopping, asda, bargain, customers, market, retail*

Sports: *#football, #winner, ball, bench, defending, footballer*

Confusion matrices (aggregate)



O = output (inferred), **T** = target, **P** = precision, **R** = recall
{1, 2, 3} = {upper, middle, lower} socioeconomic status

E

Formulating a Gaussian Process classifier

Definition:

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

$$f : \mathbb{R}^d \to \mathbb{R} \quad \mathbf{x} \in \mathbb{R}^d$$

Kernel formulation:

$$k(\mathbf{x}, \mathbf{x}') = \left(\sum_{n=1}^{C} k_{\mathrm{SE}}(\mathbf{c}_n, \mathbf{c}'_n) \right) + k_{\mathrm{N}}(\mathbf{x}, \mathbf{x}')$$

where

$$\mathbf{x} = \{\mathbf{c}_1, \ldots, \mathbf{c}_C\},\ C = 5$$

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \theta^2 \exp\left(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / (2\ell^2)\right)$$

$$k_{\mathrm{N}}(\mathbf{x}, \mathbf{x}') = \theta_{\mathrm{N}}^2 \times \delta(\mathbf{x}, \mathbf{x}')$$

F

Classification performance (10-fold CV)

Classification	Accuracy (%)	Precision (%)	Recall (%)	F1
2-way	82.05 (2.4)	82.2 (2.4)	81.97 (2.6)	.821 (.03)
3-way	75.09 (3.3)	72.04 (4.4)	70.76 (5.7)	.714 (.05)

Conclusions. ***(a)*** *First approach for inferring the socioeconomic status of a social media user,* ***(b)*** *75% & 82% accuracy for the 3-way and binary classification tasks respectively, and* ***(c)*** *future work is required to evaluate this framework more rigorously and to analyse underlying qualitative properties in detail.*

Download the data set

Funded by the EPSRC IRC project "i-sense"

EPSRC Pioneering research and skills

i-sense

Selected References

Lampos et al. *Predicting and Characterising User Impact on Twitter*. EACL, 2014.

Preotiuc-Pietro et al. *An analysis of the user occupational class through Twitter content*. ACL, 2015.

Preotiuc-Pietro et al. *Studying User Income through Language, Behaviour and Affect in Social Media*. PLoS ONE, 2015.

Rasmussen and Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.