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

Vasileios Lampos[1,*], Nikolaos Aletras[1], Jens K. Geyti[1], Bin Zou[1], and Ingemar J. Cox[1,2]

[1] Department of Computer Science, University College London, UK
[2] Department of Computer Science, University of Copenhagen, Denmark
[*] v.lampos@ucl.ac.uk

Abstract. This paper presents a method to classify social media users based on their socioeconomic status. Our experiments are conducted on a curated set of Twitter profiles, where each user is represented by the posted text, topics of discussion, interactive behaviour and estimated impact on the microblogging platform. Initially, we formulate a 3-way classification task, where users are classified as belonging to an upper, middle or lower socioeconomic class. A nonlinear, generative learning approach using a composite Gaussian Process kernel provides significantly better classification accuracy (75%) than a competitive linear alternative. By turning this task into a binary classification – upper vs. medium and lower class – the proposed classifier reaches an accuracy of 82%.

Keywords: social media; Twitter; user profiling; socioeconomic status; classification; Gaussian Process

1 Introduction

Online information has been used in recent research to derive new or enhance our existing knowledge about the *physical* world. Some examples include the use of social media or search query logs to model financial indices [1], understand voting intentions [10] or improve disease surveillance [8,4,9]. At the same time, complementary studies have focused on characterising individual users or specific groups of them. It has been shown that user attributes, such as age [15], gender [2], impact [7], occupation [14] or income [13], can be inferred from Twitter profiles. This automatic and often large-scale information extraction has commercial and research applications, from improving personalised advertisements to facilitating answers to various questions in the social sciences.

This paper presents a method for classifying social media users according to their socioeconomic status (SES).[1] SES can be broadly defined as one's access to financial, social, cultural, and human capital resources; it also includes additional components such as parental and neighbourhood properties [3]. We focused our work on the microblogging platform of Twitter and formed a new data set of user profiles together with a SES label for each one of them. To

[1] To the best of our knowledge, this task has not previously been reported.

map users to a SES, we utilised the Standard Occupation Classification (SOC) hierarchy, a broad taxonomy of occupations attached to socioeconomic categorisations in conjunction with the National Statistics Socio-Economic Classification (NS-SEC) [5,17].

Users are represented by a broad set of features reflecting their behaviour and impact on the social platform. The classification task uses a nonlinear, kernelised method that can more efficiently capture the divergent feature categories. Related work has looked into different aspects of this problem, such as inferring the job category [14] or the income (as a regression task [13]) of social media users. Similarly to our work here, nonlinear methods showed better performance in these tasks as well. However, the previously proposed models did not jointly explore the various sets of features, as the method of this paper suggests. The proposed classifier achieves a strong performance in both 3-way and binary classification scenarios.

2 Data Set and Task Description

Our analysis is conducted on a set of $1,342$ Twitter user profiles located in the UK[2] and their corresponding tweets from February 1, 2014 to March 21, 2015 inclusive ($2,082,651$ tweets in total; denoted by $\mathcal{D}_1$). The user selection was performed by searching for occupation mentions in the profile description field of a pool of approximately $100,000$ UK Twitter users. An extensive taxonomy of occupations was obtained from the SOC hierarchy. We have manually supervised this process, removing accounts where the assigned occupation was incorrect or uncertain. Accounts that were not related to individuals (e.g. representing an organisation) were not considered.

We have also created an additional data set by randomly sampling all UK tweets posted in the same exact period as $\mathcal{D}_1$ ($159,101,560$ tweets in total; denoted by $\mathcal{D}_2$). $\mathcal{D}_2$ was used to automatically compile a set of latent topics that Twitter users were communicating about.

From $\mathcal{D}_1$, we extracted the following five user **feature categories**:

- $\mathbf{c}_1$: Platform-based **behaviour** as represented by the proportion (over the total number of tweets) of retweets, mentions of, unique mentions of and replies to other user accounts.
- $\mathbf{c}_2$: Platform **impact** expressed by the number of accounts followed (followees), followed by (followers), times listed (bookmarked) as well as a user impact score (defined in [7]) that combines the previous metrics.
- $\mathbf{c}_3$: Keywords (1-grams and 2-grams) present in a user's **profile description**.
- $\mathbf{c}_4$: The frequency of the 1-grams present in a user's **tweets**. The frequency of a 1-gram x for a user i is defined as $z_i = |x_i|/N_i$, where N_i denotes the total number of tweets for i and $|x_i|$ is the number of appearances of x in them.
- $\mathbf{c}_5$: A frequency distribution across a set of 200 latent **topics** represented by clusters of 1-grams. The frequency of a topic τ for a user i is defined as $\tau_i = \sum_\tau z_i^\tau$, where z_i^τ denotes the frequency of a 1-gram that belongs to the cluster of 1-grams (topic) τ.

[2] Inferred from the location name provided in the user profile description.

Table 1. 1-gram samples from a subset of the 200 latent topics (word clusters) extracted automatically from Twitter data ($\mathcal{D}_2$).

Topic	Sample of 1-grams
Corporate	#business, clients, development, marketing, offices, product
Education	assignments, coursework, dissertation, essay, library, notes, studies
Family	#family, auntie, dad, family, mother, nephew, sister, uncle
Internet Slang	ahahaha, awwww, hahaa, hahahaha, hmmmm, looooool, oooo, yay
Politics	#labour, #politics, #tories, conservatives, democracy, voters
Shopping	#shopping, asda, bargain, customers, market, retail, shops, toys
Sports	#football, #winner, ball, bench, defending, footballer, goal, won
Summertime	#beach, #sea, #summer, #sunshine, bbq, hot, seaside, swimming
Terrorism	#jesuischarlie, cartoon, freedom, religion, shootings, terrorism

The dimensionality of user attributes $\mathbf{c}_3$ and $\mathbf{c}_4$, after filtering out stop words and n-grams occurring less than two times in the data, was equal to 523 (1-grams plus 2-grams) and 560 (1-grams) respectively. Thus, a Twitter user in our data set is represented by a $1,291$-dimensional feature vector.

We applied spectral clustering [12] on $\mathcal{D}_2$ to derive 200 (hard) clusters of 1-grams that capture a number of latent topics and linguistic expressions (e.g. 'Politics', 'Sports', 'Internet Slang'), a snapshot of which is presented in Table 1. Previous research has shown that this amount of clusters is adequate for achieving a strong performance in similar tasks [7,13,14]. We then computed the frequency of each topic in the tweets of $\mathcal{D}_1$ as described in feature category $\mathbf{c}_5$.

To obtain a SES label for each user account, we took advantage of the SOC hierarchy's characteristics [5]. In SOC, jobs are categorised based on the required skill level and specialisation. At the top level, there exist 9 general occupation groups, and the scheme breaks down to sub-categories forming a 4-level structure. The bottom of this hierarchy contains more specific job groupings (369 in total). SOC also provides a simplified mapping from these job groupings to a SES as defined by NS-SEC [17]. We used this mapping to assign an upper, middle or lower SES to each user account in our data set. This process resulted in 710, 318 and 314 users in the upper, middle and lower SES classes, respectively.[3]

3 Classification Methods

We use a composite Gaussian Process (GP), described below, as our main method for performing classification. GPs can be defined as sets of random variables, any finite number of which have a multivariate Gaussian distribution [16]. Formally, GP methods aim to learn a function $f : \mathbb{R}^d \to \mathbb{R}$ drawn from a GP prior given the inputs $\mathbf{x} \in \mathbb{R}^d$:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \,, \tag{1}$$

[3] The data set is available at http://dx.doi.org/10.6084/m9.figshare.1619703.

Table 2. SES classification mean performance as estimated via a 10-fold cross validation of the composite GP classifier for both problem specifications. Parentheses hold the SD of the mean estimate.

Num. of classes	Accuracy	Precision	Recall	F-score
3	75.09% (3.28%)	72.04% (4.40%)	70.76% (5.65%)	.714 (.049)
2	82.05% (2.41%)	82.20% (2.39%)	81.97% (2.55%)	.821 (.025)

where $m(\cdot)$ is the mean function (here set equal to 0) and $k(\cdot,\cdot)$ is the covariance kernel. We apply the squared exponential (SE) kernel, also known as the radial basis function (RBF), defined as $k_{\mathrm{SE}}(\mathbf{x},\mathbf{x}') = \theta^2 \exp\left(-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\ell^2)\right)$, where θ^2 is a constant that describes the overall level of variance and ℓ is referred to as the characteristic length-scale parameter. Note that ℓ is inversely proportional to the predictive relevancy of $\mathbf{x}$ (high values indicate a low degree of relevance). Binary classification using GPs 'squashes' the real valued latent function $f(\mathbf{x})$ output through a logistic function: $\pi(\mathbf{x}) \triangleq \mathrm{P}(y=1|\mathbf{x}) = \sigma(f(\mathbf{x}))$ in a similar way to logistic regression classification. In binary classification, the distribution over the latent f_* is combined with the logistic function to produce the prediction $\bar{\pi}_* = \int \sigma(f_*)\mathrm{P}(f_*|\mathbf{x},\mathbf{y},x_*)df_*$. The posterior formulation has a non-Gaussian likelihood and thus, the model parameters can only be estimated. For this purpose we use the Laplace approximation [16,18].

Based on the property that the sum of covariance functions is also a valid covariance function [16], we model the different user feature categories with a different SE kernel. The final covariance function, therefore, becomes

$$k(\mathbf{x},\mathbf{x}') = \left(\sum_{n=1}^{C} k_{\mathrm{SE}}(\mathbf{c}_n,\mathbf{c}'_n)\right) + k_{\mathrm{N}}(\mathbf{x},\mathbf{x}')\,, \tag{2}$$

where $\mathbf{c}_n$ is used to express the features of each category, i.e., $\mathbf{x} = \{\mathbf{c}_1,\ldots,\mathbf{c}_C,\}$, C is equal to the number of feature categories (in our experimental setup, $C=5$) and $k_{\mathrm{N}}(\mathbf{x},\mathbf{x}') = \theta_{\mathrm{N}}^2 \times \delta(\mathbf{x},\mathbf{x}')$ models noise (δ being a Kronecker delta function). Similar GP kernel formulations have been applied for text regression tasks [7,9,11] as a way of capturing groupings of the feature space more effectively.

Although related work has indicated the superiority of nonlinear approaches in similar multimodal tasks [7,14], we also estimate a performance baseline using a linear method. Given the high dimensionality of our task, we apply logistic regression with elastic net regularisation [6] for this purpose. As both classification techniques can address binary tasks, we adopt the one–vs.–all strategy for conducting an inference.

4 Experimental Results

We assess the performance of the proposed classifiers via a stratified 10-fold cross validation. Each fold contains a random 10% sample of the users from each of the three socioeconomic statuses. To train the classifier on a balanced data set, during training we over-sample the two less dominant classes (middle and lower), so that they match the size of the one with the greatest representation (upper).

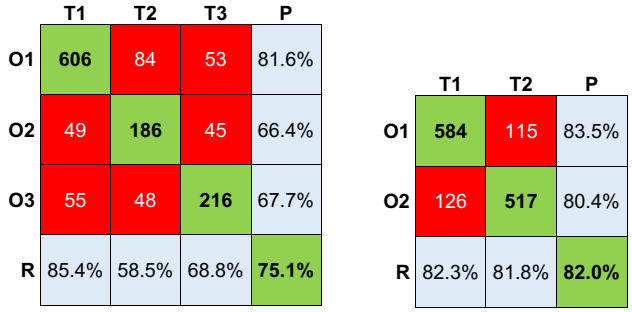

Fig. 1. The cumulative confusion matrices for the 3-way (left) and binary (right) classification tasks. Columns contain the **T**arget class labels and rows the **O**utput ones. The row and column extensions respectively specify the **P**recision and **R**ecall per class. The numeric identifiers (1–3) are in descending SES order (upper to lower).

We have also tested the performance of a binary classifier, where the middle and lower classes are merged. The cumulative confusion matrices (all data from the 10 folds) for both classification scenarios and the GP-based classifier are presented in Fig. 1. Table 2 holds the respective mean performance metrics. The mean accuracy of the 3-way classification obtained by the GP model is equal to 75.09% (SD = 3.28%). The regularised logistic regression model yielded a mean accuracy of 72.01% (SD = 2.45%). A two sample t-test concluded that the 3.08% difference between these mean performances is statistically significant ($p = 0.029$). The precision and recall per class are reported in the row and column extensions of the confusion matrices respectively. It is evident that it is more difficult to correctly classify users from the middle class (lowest precision and recall). The binary classifier is able to create a much better class separation, achieving a mean accuracy of 82.05% (SD = 2.41%) with fairly balanced precision and recall among the classes.

Looking at the occupation titles of the users, where false negatives occurred in the 3-way classification, we identified the following jobs as the most error-prone: 'sports players' for the upper class, 'photographers', 'broadcasting equipment operators', 'product/clothing designers' for the middle class, 'fitness instructors' and 'bar staff' for the lower class. Further investigation is needed to fully understand the nature of these errors. However, we note that SES is influenced by many factors, including income, education and occupation. In contrast, our classifier does not explicitly consider either income or education, and this may limit accuracy.

5 Conclusions and Future Work

We have presented the first approach for inferring the socioeconomic status of a social media user based on content (text, topics) and behaviour (interaction, impact). As in previous case studies [7,14], the multimodal feature space favoured a nonlinear classifier. Our method yielded strong accuracy in both 3-way (75%) and binary (82%) classification scenarios. The absence of a definitive gold stan-

dard for training and evaluating, i.e. a confirmed SES that represents each user rather than a simplified estimate of it through the SOC taxonomy, is the main limitation for this line of research. Future work should focus on the construction of a stronger evaluation framework, as well as improved classification algorithms. Nevertheless, we hope that the method outlined here will facilitate subsequent research in the domains of computational social science and digital health.

Acknowledgements. This work has been supported by the EPSRC grant EP/K031953/1 ("Early-Warning Sensing Systems for Infectious Diseases").

References

1. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comp. Sci. 2(1), 1–8 (2011)
2. Burger, D.J., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. pp. 1301–1309. In EMNLP (2011)
3. Cowan, C.D., et al.: Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress: A Theoretical Foundation. Tech. Report, National Center for Education Statistics (2003)
4. Culotta, A.: Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. pp. 115–122. In SMA (2010)
5. Elias, P., Birch, M.: SOC2010: Revision of the Standard Occupational Classification. Economic and Labour Market Review 4(7), 48–55 (2010)
6. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33(1), 1–22 (2010)
7. Lampos, V., Aletras, N., Preoţiuc-Pietro, D., Cohn, T.: Predicting and Characterising User Impact on Twitter. pp. 405–413. In EACL (2014)
8. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the Social Web. pp. 411–416. In CIP (2010)
9. Lampos, V., Miller, A.C., Crossan, S., Stefansen, C.: Advances in nowcasting influenza-like illness rates using search query logs. Sci. Rep. 5, 12760 (2015)
10. Lampos, V., Preoţiuc-Pietro, D., Cohn, T.: A user-centric model of voting intention from Social Media. pp. 993–1003. In ACL (2013)
11. Lampos, V., Yom-Tov, E., Pebody, R., Cox, I.: Assessing the impact of a health intervention via user-generated Internet content. Data Min. Knowl. Disc. 29(5), 1434–1457 (2015)
12. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. 17(4), 395–416 (2007)
13. Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N.: Studying User Income through Language, Behaviour and Affect in Social Media. PLoS ONE 10(9), e0138717 (2015)
14. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content . pp. 1754–1764. In ACL (2015)
15. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying Latent User Attributes in Twitter. pp. 37–44. In SMUC (2010)
16. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
17. Rose, D., Pevalin, D.: Re-basing the NS-SEC on SOC2010: A Report to ONS. Tech. Report, University of Essex (2010)
18. Williams, C.K.I., Barber, D.: Bayesian classification with Gaussian Processes. IEEE T. Pattern Anal. 20(12), 1342–1351 (1998)