Digital Footprints: Envisaging and Analysing Online Behaviour

Giles Oatley and **Tom Crick**[1] and **Mohamed Mostafa**

Abstract. Our long-term research goal is the development of complex (and adaptive) behavioural modelling and profiling using a multitude of online datasets; in this paper we look at suitable tools for use in big social data, specifically here on how to 'envisage' this complex information. We present a novel way of representing personality traits (using the Five Factor model) with behavioural features (fantasy and profanity). We also present some preliminary ideas around developing a scalable solution to modelling behaviour using swear words.

1 Introduction

There are large-scale research efforts in developing new and robust techniques for modelling online behaviour and identity. There exists numerous domains in which it is essential to obtain knowledge about user profiles or models of software applications, including intelligent agents, adaptive systems, intelligent tutoring systems, recommender systems, e-commerce applications and knowledge management systems [32]. The rise of Web 2.0 and social networking has facilitated the publishing of user-generated content on an exponential scale; its analysis is becoming increasingly important (and applicable) to the empirical study of society (and thus societal change).

Big datasets from social networking platforms are now being used for a multitude of purposes, alongside the obvious advertising, marketing and revenue generation; increasingly for government monitoring of citizens[2,3,4], along with covert security, intelligence community and military user profiling. However, the publishing of user-generated content on an exponential scale has significantly changed qualitative and quantitative social research, with its analysis becoming increasingly important to the empirical study of society. There are interesting sociological uses of studying or mining big social data, for instance exploring cyber-physical crowds using location-tagged social networks or the study of personality with large-scale benchmark social datasets and corpora.

However, this "big social data" from social media platforms, for instance social networks, blogs, gaming, shopping and review sites, differs significantly from more traditional/formal sources. With the advent of the social web, there are now orders of magnitude more data available relating to uncensored natural language, requiring the development of new techniques that can meaningful analyse it. This uncensored language is rich in 'unnatural' language (as opposed to 'natural' language, used in formal/traditional published media such as books and newspapers), defined as "*informal expressions, variations, spelling errors...irregular proper nouns, emoticons, unknown words*"[5]. We have been interested in profiling complex behaviours [20], particularly for crime informatics [22, 21] and in this paper we include in our models such bad behaviour that is found in big social data, for example so-called unnatural language with its poor language construction but also context dependent acronyms, jargon, "leetspeak" and swear words or profanity. Leet, also known as eleet or leetspeak, is an alternative alphabet for the English language that is used primarily on the Internet and in geek/cyber communities. It uses various combinations of ASCII characters to replace Latin script. For example, leet spellings of the word "leet" include *1337* and *l33t*; eleet may be spelled *31337* or *3l33t*. See Perea et al. [29] for an discussion of leet from a cognitive processing perspective.

2 Modelling Fantasy and Profanity

2.1 Rude Words: The Language of Pornography

A research project investigating opinions on a range of topics related to pornography usage was carried out; a web-based questionnaire received over five thousand respondents (n=5490). Several of the questions were open-ended, for instance how the person became involved with the subject of pornography, their particular interests and so on, eliciting a number of detailed responses (c.2000 words). From the initial findings [33], the data is ill-structured, with frequent usage of bad grammar and contains a large number of jargon (swear) words relating to pornography and sexuality.

An aim of the original study was the investigation of the usage of fantasy. This resonated with our general interest in determining behaviour from data, and so explored the language characteristics of the answers related specifically to fantasy. We analysed the respondents text using the psycholinguistic databases LIWC and MRC. The Dictionary of Affect in Language (DAL) [35] was also used, due to its specific uses for imagery-based language. We used methods derived from LIWC and MRC to determine personality traits and measures such as formality and deception. We wanted to get a general feel for the level of the text, and to see if there were any correlations between literacy and readability.

Initially we focused on the specific questions that might reveal something about the role of fantasy. For instance, among the many options for the question "*What are your reasons for looking at pornography?*", among the list were the following:

[1] All authors: Department of Computing, Cardiff Metropolitan University, UK; {goatley,tcrick,mmostafa}@cardiffmet.ac.uk

[2] Twitter Transparency Report 2014: `https://transparency.twitter.com/`

[3] Facebook Global Government Requests Report 2014: `https://govtrequests.facebook.com/`

[4] Google Transparency Report 2014: `http://www.google.co.uk/transparencyreport/`

[5] 2nd Unnatural Language Processing Contest, part of the 17th Annual Meeting of the Association for Natural Language Processing (NLP2011): `http://www.anlp.jp/nlp2011/`

(A) “*To see things I might do*”;
(B) “*To see things I can’t do*”;
(C) “*To see things I wouldn’t do*”;
(D) “*To see things I shouldn’t do*”.

The ‘*can’t*’ and ‘*wouldn’t*’ choices clearly indicate respondents utilising pornography more strongly as a form of fantasy. For this we explored the Five Factors personality traits, in particular expecting some correlation with the *Openness to Experience* factor (see Figures 1–4.

	A	B	C	D
A	1			
B	-0.72974	1		
C	-0.46635	-0.06469	1	
D	-0.33821	0.08321	0.091183	1

Table 1. Correlation between question items (where: A=“*To see things I might do*”; B=“*To see things I can’t do*”; C= “*To see things I wouldn’t do*” D=“*To see things I shouldn’t do*”)

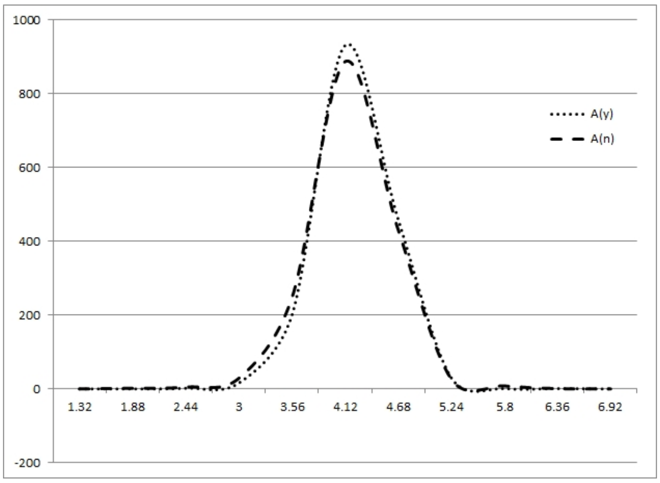

Figure 1. Openness to experience for A(y) (dotted) versus non-A (dashed)

Analysis is ongoing, with the results to be published in the near future; however there appears to be a strong negative correlation between participants who chose “*A. To see things I might do*” versus “*B. To see things I can’t do*”, as originally hypothesised. What was less convincing was our analysis of the Five Factors, and we put this down to the measures we used from [16] being derived from a very different corpus. We are currently concentrating on the lower level features from LIWC, MRC and DAL.

2.2 Disambiguating Profanity

WordNet[6] is a large lexical database of English; nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept, and each synset is interlinked by means of conceptual-semantic and lexical relations. Words that are found in close proximity to one another in the network are semantically disambiguated. WordNet Affect[7], a hierarchical set of emotional categories, and SentiWordNet[8], synsets are assigned sentiment scores (positivity, negativity, objectivity), are built on top of WordNet.

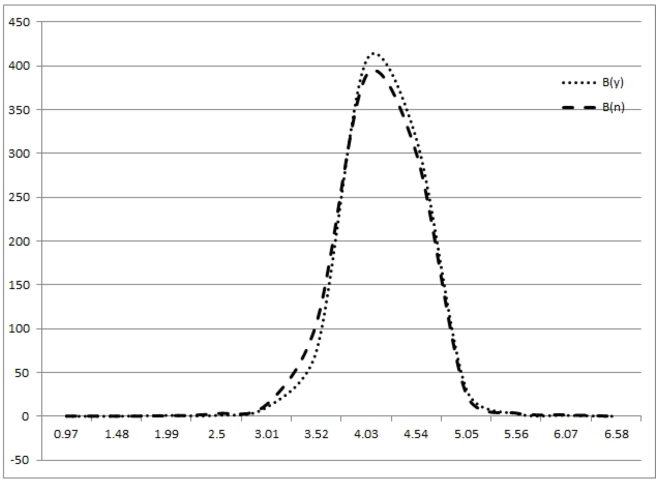

Figure 2. Openness to experience for B(y) (dotted) versus non-A (dashed)

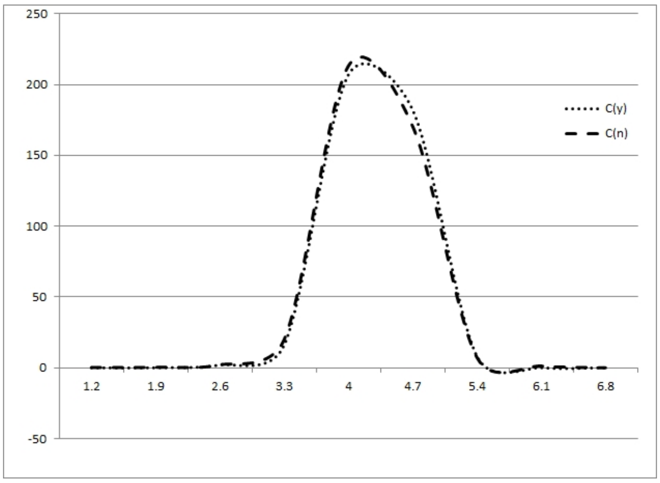

Figure 3. Openness to experience for C(y) (dotted) versus non-A (dashed)

Millwood-Hargrave’s study [17] for Ofcom (formerly, the Broadcasting Standards Commission), the UK’s regulatory and competition authority for the broadcasting, telecommunications and postal industries, in 2000 was designed to test people’s attitudes to swearing and offensive language, and to examine the degree to which context played a role in their reactions. Included in the report were attitudes towards swearing and offensive language ‘in life’, including a range of swear words and terms of abuse. Appendix 2’s ‘list of words’ contained positions of the top swear words (categorised as “*very severe*”, “*fairly severe*”, “*quite mild*” and “*not swearing*”) and their ranking from 1998 to 2000.

[6] http://wordnet.princeton.edu/

[7] http://wndomains.fbk.eu/wnaffect.html

[8] http://sentiwordnet.isti.cnr.it/

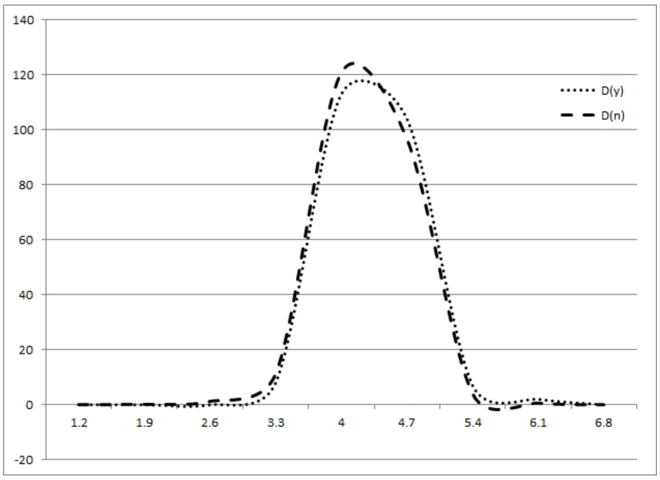

Figure 4. Openness to experience for D(y) (dotted) versus non-A (dashed)

The study of swear words has a longstanding position in linguistics, with the academic journal *Maledicta: The International Journal of Verbal Aggression* running from 1977 until 2005. Maledicta was dedicated to the study of the origin, etymology, meaning, use and influence of vulgar, obscene, aggressive, abusive and blasphemous language. Unfortunately we do not have resources such as databases in the literature; furthermore, WordNet does not contain the range of swear words we encountered in our data and is no use for disambiguating our text. Wikipedia, however, fared much better; but even better than these were Roger's Profanisaurus and Urban Dictionary.

Roger's Profanisaurus[9] is a lexicon of profane words and expressions; the 2005 version (the Profanisaurus Rex), contains over 8,000 words and phrases, with a further-expanded version released in 2007. Unlike a traditional dictionary or thesaurus, the content is enlivened by often pungent or politically incorrect observations and asides intended to provide further comic effect.

Urban Dictionary[10] is a Web-based dictionary that contains nearly eight million definitions as of December 2014. Originally, Urban Dictionary was intended as a peer-reviewed dictionary of slang or cultural words or phrases not typically found in standard dictionaries, with words or phrases on Urban Dictionary having multiple definitions, usage examples and tags.

We created different gazetteers related to rude words; one list was based on Wikipedia entries, and another on lists from Urban Dictionary. The Wikipedia list was created from link text on the Wikipedia porn sub-genre page[11] (link "anchor text" is a typical approach in semantic relatedness studies). This was comprised of 250 words. The Urban Dictionary list was created from the "sex" category[12] (by no means exhaustive – it is a fraction of the pornography-related terms in Urban Dictionary). This was comprised of 156 words. We implemented two metrics for rude words, the key idea of which is to have a simple mathematical model that enables us to estimate the life-history value of a token.

There are numerous other lists of pornographic words, which we compiled from miscellaneous sources; however, we are mainly interested in sources such as Wikipedia and Urban Dictionary as these are maintained by a similar community that uses the words in social networking. In this way we do not have to concern ourselves about this knowledge engineering process, merely concern ourselves about the representation and quality of meaning or definitions. We will in future work make use of the voting scores available on Urban Dictionary, and look to incorporate new resources such as Roger's Profanisaurus.

3 Psycholinguistic Models and Representing Complex Behaviour

Advances in psychology research have suggested it is possible for personality to be determined from digital data [28, 41, 15]. Recent studies [44] have suggested certain keywords and phrases can signal underlying tendencies and that this can form the basis of identifying certain aspects of personality. Extrapolating this suggests that by investigation of an individual's online comments it may be possible to identify individual's personality traits. Initial evidence in support of this hypothesis was demonstrated in 2012 by analysis of Twitter data for indicators of psychotic behaviour [34]. While in the past this has mainly been the textual information contained in blogs, status posts and photo comments [2, 3], there is also a wealth of information in the other ways of interacting with online artefacts. For instance, it is possible to observe the ordering/timings of button clicks of a user. Several researchers have looked at personality prediction (e.g. Five Factor personality traits) based on information in a user's Facebook profile [1, 14] and speech [9, 37], as well as also demonstrating significant correlations with fine affect (emotion) categories such as that of excitement, guilt, yearning, and admiration [18]. There are also several strands of related work based on the benchmark myPersonality Project[13] dataset [7], providing a platform for much-needed comparative studies.

Mairesse et al. [16] highlighted the use of features from the psycholinguistic databases LIWC [27] and MRC [43] to create a range of statistical models for each of the Five Factor personality traits [19, 26].

In previous work [20] we utilised these methods to develop a complex behavioural profile that included 'two faces' to model that we can have several different modes of operation (ego states). We performed our Five Factor analysis, and elaborated two sets of Five Factor results for each user. We chose Chernoff faces [8] for the visual representation. The Five Factors are displayed as five features on a stylised face, where:

- Width of hair represents *Conscientiousness*;
- Width of eyes represents *Agreeableness*;
- Width of nose represents *Openness to experience*;
- Width of mouth represents *Emotional stability*;
- Height of face represents *Extraversion.*

It should be noted that while researchers have continued to work with the Five Factors model, there are well known limitations [13, 25, 4] that are often overlooked by researchers. In particular, it has been criticised for its limited scope, methodology and the absence of an underlying theory. However, attempts to replicate the Big Five in other countries with local dictionaries have succeeded in some countries but not in others [36, 11]. While [10] claim that their Five Factors model "represents basic dimensions of personality", psychologists have identified important trait models, for instance Cattell's 16 Personality Factors [6] and Eysenck's biologically-based theory [12].

[9] http://www.viz.co.uk/profanisaurus.html

[10] http://www.urbandictionary.com/

[11] http://en.wikipedia.org/wiki/List_of_pornographic_sub-genres

[12] http://www.urbandictionary.com/category/sex

[13] http://mypersonality.org/

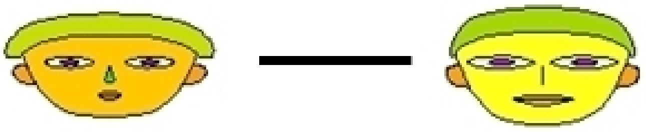

Figure 5. Two faces of a person. Personality traits from the Five Factors model are mapped on a Chernoff face (see later figure for specific trait mappings). Two different faces are drawn from two different linguistic sources, for the same person.

4 Envisaging Information

By analysing the myriad approaches of representing complex information, it is easy to be inspired by Tufte's clarity, precision, and efficiency [40, 39, 38]. We have integrated the profanity and fantasy behavioural features into our Chernoff face representing the Five Factor traits – see Figure 6 – represented on a Chernoff face are the Five Factors plus the additional behaviours for swearing level (darkness of blue colour on face) and fantasy level (amount of 'thought bubbles').

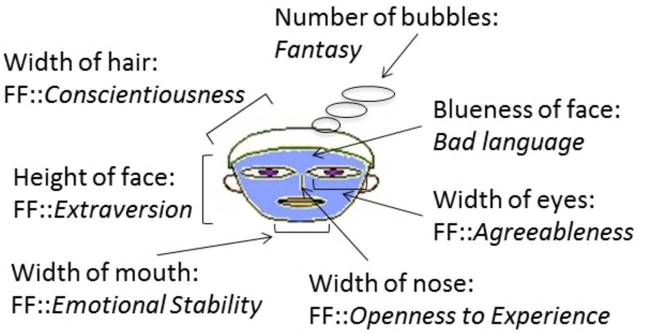

Figure 6. Traits and behaviours. Represented on a Chernoff face are the Five Factors (prepended by FF::) plus the additional behaviours for swearing level (darkness of blue colour on face) and fantasy level (amount of 'thought bubbles').

4.1 Modelling Timelines

Elsewhere we have presented ways to fuse social network (graph) information with geographical information [24, 23], and from spatial statistics there exists methods for space and time such as the Knox and Mantel indices. In this section we look at a method to represent temporal events, something very necessary when developing a behavioural profile.

Our data comes from an online portal for a European Union (EU) international scholarship mobility hosted at a UK university. The case study looked at how people interact with complex online information systems, the online portal for submitting applications. We analysed the document uploading behaviour (also motivation letters, and social media interactions) of the applicants. By examining the upload footprint for the users we determined several classes of behaviour.

There were several thousand applications submitted by over a thousand candidates, applying to 10 EU universities and 10 non-EU universities. Each mobility call has an opening date/time and closing date/time, with occasional extensions given for specific reasons (for instance due to administrative reasons or technical issues with the portal). Applicants are required to submit for their application certain mandatory files, such as motivation letter, passport/identification, curriculum vitae), as well as optional files (supporting documents).

We simplified an applicant's interaction, or timeline, with the portal to include the following milestones: *T0* Registration Time; *T1* First Action; *T2* Last Action; and, *T3* Submission. Additionally we represented an extension to the submission deadline as *T4* Extension. In this way we can represent an applicants interaction as shown in Figure 7, which shows seven example timelines.

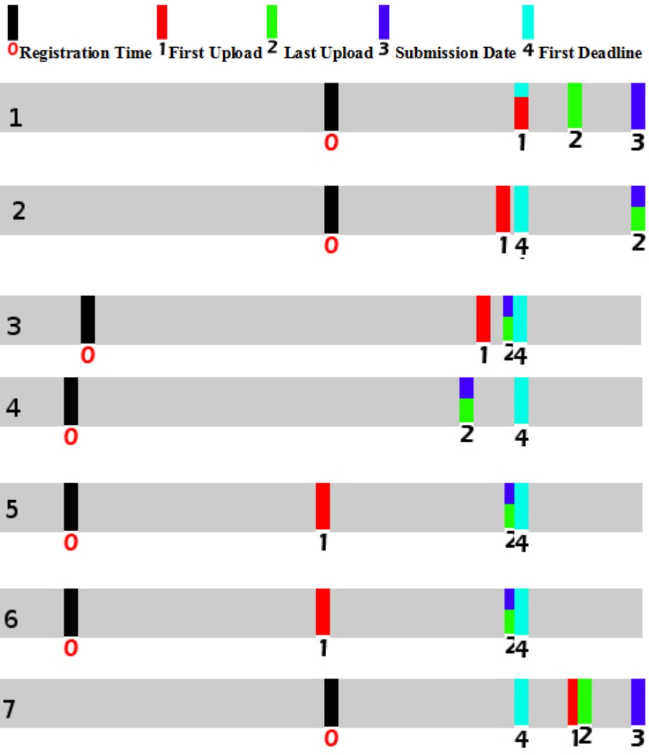

Figure 7. Seven user timelines. *T0* (black bar) is when the applicant first registered with the call. *T1* (red bar) represents when the applicant uploaded their first document, or First Action. *T2* (green bar) represents an applicants' Last Action. *T3* (blue bar) represents the applicants' Submission. *T4* (aquamarine bar) represents the first deadline (certain calls had initial deadlines extended).

Using these milestones we are able to identify interesting behaviours that compare and contract with personality traits and other sources of information. Behaviours such as: how long it was before an applicant became aware of the call, and when they registered; how long after registration did the applicant carry out their first action with the system; how long did they take to complete their application; and, how close to the deadline did they submit their application.

The complete timeline from opening to final close was 125 days. There was an extension from day 112 until day 125. We divided the timeline of the call into five equally spaced segments (S0-S4).

Using these segments we were able to assign the various applicant actions (*T0* Registration, *T1* First Upload, *T2* Last Upload, *T3* Submission) to various time periods. This allowed us to assign applicants to statistically significant categories, and also to add in a few categories from observations. These are shown in the following Table 2; as you can see, a small number of applicants (*n*=4) registered within the segment S1 (20-40% of timeline), and then uploaded all of their documents and submitted within the segment S3 (60-90% of timeline). This is represented by Class A, the first row. Successive rows can be interpreted in the same manner.

Class	*n*	*T0*	*T1*	*T2*	*T3*
A	4	S1	S3	S3	S3
B	14	S2	S2	S2	S2
C	128	S2	S3	S3	S3
D	29	S2	S3	S4	S4
E	678	S3	S3	S3	S3
F	202	S3	S3	S4	S4
G	9	S3	S4	S4	S4
H	54	S4	S4	S4	S4

Table 2. Applicants' timeline actions assigned to segments

We did not want to ascribe a premature alias to the behaviours, as we recognise that there are several possible interpretations; nevertheless, we have used the 'Potential Alias' column in Table 3 to indicate some initial thoughts.

Combining this information with the earlier trait and behaviour model, it could be possible to present several faces along the timeline, or to represent the temporal aspect as a 'clock-type' metaphor, the straight line curved around, surrounding the face. The latter would perhaps be preferable, as we would expect that traits persist through time, but behaviours change. Likewise we would expect the blueness (rudeness) of the Chernoff face to change, and the amount of bubbles (fantasy) to change, but the facial features to remain constant (personality traits).

5 Conclusions and Future Work

The linguistic methods for determining personality traits are still in their infancy, and we have already noted some of the opposition to the lexical hypothesis [4]. Generally, information conveyed by the use of terms in human dialog studied in linguistics follows precise rules; other important rules are now introduced in the philosophy of language, investigating the meanings of terms and their extra-linguistic reference. We would expect that in time these additional information sources, like how people project identities through personal websites [41], judging people by their music preferences [30], personalisation of workspaces [42], etc, will all help with classification.

Further problems related to using social media for classification are that existing NLP tools are known to struggle with unnatural language: "*demonstrated that existing tools for POS tagging, chunking and Named Entity Recognition perform quite poorly when applied to tweets*" [31] and "*showed that [lengthening words] is a common phenomenon in Twitter*" [5], presenting a problem for lexicon-based approaches. These investigations both employed some form of inexact word matching to overcome the difficulties of unnatural language. We have made no attempt to use inexact string matching or to make use of a leetspeak parser. This will form part of future work.

The Web constitutes a world made of a precise formal-social ontology which hardly reflect the complexity of human personality; it is a difficult enterprise to try to mediate between the personal world of humans and the impersonal one of the Web. To assist with the ongoing knowledge modelling problem in this domain we recognise the need to utilise specific lexicons that keep pace with the language used, for instance the use of Urban Dictionary to resolve swear words. We thus need to study how in what precise manner this resource keeps pace with popular culture.

REFERENCES

[1] Mitja D. Back, Juliane M. Stopfer, Simine Vazire, Sam Gaddis, Stefan C. Schmukle, Boris Egloff, and Samuel D. Gosling, 'Facebook Profiles Reflect Actual Personality, Not Self-Idealization', *Psychological Science*, **21**(3), 372–374, (2010).

[2] Benjamin Blamey, Tom Crick, and Giles Oatley, 'R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora', in *Research and Development in Intelligent Systems XXIX, Proceedings of the 32nd SGAI International Conference on Artificial Intelligence (AI-2012)*, pp. 207–212. Springer, (2012).

[3] Benjamin Blamey, Tom Crick, and Giles Oatley, ''The First Day of Summer': Parsing Temporal Expressions with Distributed Semantics', in *Research and Development in Intelligent Systems XXX, Proceedings of the 33rd SGAI International Conference on Artificial Intelligence (AI-2013)*, pp. 389–402, (2013).

[4] Jack Block, 'The Five-Factor Framing of Personality and Beyond: Some Ruminations', *Psychological Inquiry*, **21**(1), 2–25, (2010).

[5] Samuel Brody and Nicholas Diakopoulos, 'Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pp. 562–570, (2011).

[6] Raymond B. Cattell, *The description and measurement of personality*, Harcourt, Brace & World, 1946.

[7] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski, eds. *Workshop on Computational Personality Recognition (Shared Task)*, 2013.

[8] Herman Chernoff, 'The Use of Faces to Represent Points in k-Dimensional Space Graphically', *Journal of the American Statistical Association*, **68**(342), 361–368, (1973).

[9] Cindy Chung and James W. Pennebaker, *Social Communication: Frontiers of Social Psychology*, chapter The psychological function of function words, 343–359, Psychology Press, 2007.

[10] Paul T. Costa and Robert R. McCrae, *Neo PI-R Professional Manual*, Psychological Assessment Resources, 1992.

[11] Filip De Fruyt, Robert R. McCrae, Zsófia Szirmák, and János Nagy, 'The Five-Factor Personality Inventory as a Measure of the Five-Factor Model: Belgian, American, and Hungarian Comparisons with the NEO-PI-R', *Assessment*, **11**(3), 207–215, (2004).

[12] Hans J. Eysenck, *Dimensions of Personality*, Routledge & Kegan Paul, 1947.

[13] Hans J. Eysenck, 'Four ways five factors are not basic', *Personality and Individual Differences*, **13**(6), 667–673, (1992).

[14] Jennifer Golbeck, Cristina Robles, and Karen Turner, 'Predicting personality with social media', in *Proceedings of Human Factors in Computing Systems (CHI'11)*, pp. 253–262. ACM Press, (2011).

[15] Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander, 'Large Scale Personality Classification of Bloggers', in *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, volume 6975 of *Lecture Notes in Computer Science*, pp. 568–577. Springer, (2011).

[16] François Mairesse, Marilyn A. Walker, Matthias R. Mehi, and Roger K. Moore, 'Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text', *Journal of Artificial Intelligence Research*, **30**, 457–500, (2007).

[17] Andrea Millwood-Hargrave, 'Delete expletives?', Technical report, Research undertaken jointly by the Advertising Standards Authority, British Broadcasting Corporation, Broadcasting Standards Commission and the Independent Television Commission, (2000).

[18] Saif M. Mohammad and Svetlana Kiritchenko, 'Using Nuances of Emotion to Identify Personality', in *Proceedings of the ICWSM Workshop on Computational Personality Recognition*, (2013).

[19] Warren T. Norman, 'Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rat-

Class	Description	Potential Alias
A	Register early, and take some time to upload documents, but submit with plenty of time before deadline	EverythingEarly
B	Register reasonably early, but then upload documents and submit straight after with plenty of time before deadline, making no amendments	QuiteEarlyAndQuick
C	Similar to Class B, but submitting more slowly	Cautious
D	Registers reasonably early, and then takes time to upload, and only submits at the last days	VeryCautious
E	Latecomer to registration, but then uploads and submits quickly thereafter	Cautious
F	Latecomer to registration, but then uploads and submits slowly	Cautious
G	Latecomer to registration, but delays uploading and submission to last days	Cautious
H	Does everything at the last days, from registration to submission	EverythingLastMinute

Table 3. Description of each class

ings', *Journal of Abnormal and Social Psychology*, **66**(6), 574–583, (1963).

[20] Giles Oatley and Tom Crick, 'Changing Faces: Identifying Complex Behavioural Profiles', in *Proceedings of 2nd International Conference on Human Aspects of Information Security, Privacy and Trust (HAS 2014)*, volume 8533 of *Lecture Notes in Computer Science*, pp. 282–293. Springer, (2014).

[21] Giles Oatley and Tom Crick, 'Exploring UK Crime Networks', in *2014 International Symposium on Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI 2014)*. IEEE Press, (2014).

[22] Giles Oatley and Tom Crick, 'Measuring UK Crime Gangs', in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE Press, (2014).

[23] Giles Oatley, Tom Crick, and Ray Howell, 'Data Exploration with GIS Viewsheds and Social Network Analysis', in *Proceedings of 23rd GIS Research UK Conference (GISRUK 2015)*, (2015). (in press).

[24] Giles Oatley, Kenneth McGarry, and Brian Ewart, 'Offender Network Metrics', *WSEAS Transactions on Information Science & Applications*, **12**(3), 2440–2448, (2006).

[25] Sampo V. Paunonen and Douglas N. Jackson, 'What is beyond the Big Five? Plenty!', *Journal of Personality*, **68**(5), 821–836, (2000).

[26] Dean Peabody and Lewis R. Goldberg, 'Some determinants of factor structures from personality-trait descriptor', *Journal of Personality and Social Psychology*, **57**(3), 552–567, (1989).

[27] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count. Erlbaum Publishers, 2001.

[28] James W. Pennebaker and Laura A. King, 'Linguistic styles: language use as an individual difference', *Journal of Personality and Social Psychology*, **77**(6), 1296–1312, (1999).

[29] Manuel Perea, Jon A. Dunabeitia, and Manuel Carreiras, 'R34D1NG W0RD5 W1TH NUMB3R5', *Journal of Experimental Psychology: Human Perception and Performance*, **34**(1), 237–241, (2008).

[30] Peter J. Rentfrow and Samuel D. Gosling, 'Message in a Ballad: The Role of Music Preferences in Interpersonal Perception', *Psychological Science*, **17**(3), 236–242, (2006).

[31] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni, 'Named entity recognition in tweets: an experimental study', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pp. 1524–1534, (2011).

[32] Silvia Schiaffino and Analía Amandi, 'Intelligent User Profiling', in *Artificial Intelligence: An International Perspective*, volume 5640 of *Lecture Notes in Computer Science*, pp. 193–216. Springer, (2009).

[33] Clarissa Smith, Feona Attwood, and Martin Barker. pornresearch.org Prelininary Findings. Available from: `http://www.pornresearch.org/Firstsummaryforwebsite.pdf`, 2013.

[34] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park, 'Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets', in *Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA 2012)*. IEEE Press, (2012).

[35] Kevin Sweeney and Cynthia Whissell, 'A dictionary of affect in language: I, establishment and preliminary validation', *Perceptual and Motor Skills*, **59**(3), 695–698, (1984).

[36] Zsófia Szirmák and Boele De Raad, 'Taxonomy and structure of Hungarian personality traits', *European Journal of Personality*, **8**(2), 95–117, (1994).

[37] Yla R. Tausczik and James W. Pennebaker, 'The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods', *Journal of Language and Social Psychology*, **29**(1), 24–54, (2010).

[38] Edward R. Tufte, *Envisioning Information*, Graphics Press USA, 1990.

[39] Edward R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press USA, 1997.

[40] Edward R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press USA, 2nd edn., 2001.

[41] Simine Vazire and Samuel D. Gosling, 'e-Perceptions: Personality Impressions Based on Personal Websites', *Journal of Personality and Social Psychology*, **87**(1), 123–132, (2004).

[42] Meredith Wells and Luke Thelen, 'What Does Your Workspace Say about You? The Influence of Personality, Status, and Workspace on Personalization', *Environment and Behavior*, **34**(3), 300–321, (20062).

[43] Michael Wilson, 'The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2.00', *Behavior Research Methods, Instruments & Computers*, **20**(1), 6–10, (1988).

[44] Michael Woodworth, Jeffrey Hancock, Stephen Porter, Robert Hare, Matt Logan, Mary Ellen OToole, and Sharon Smith, 'The Language of Psychopaths: New Findings and Implications for Law Enforcement', *FBI Law Enforcement Bulletin*, (July 2012).