Identifying more bloggers

Towards large scale personality classification of personal weblogs

Scott Nowson*
Centre for Language Technology
Macquarie University
Sydney, Australia
snowson@ics.mq.edu.au

Jon Oberlander
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh, UK
j.oberlander@ed.ac.uk

Abstract

We report new results on the relatively novel task of automatic classification of blog author personality. Promisingly high classification accuracies have recently been reported for four important personality traits (Extraversion, Neuroticism, Agreeableness and Conscientiousness). But the blog corpus used in that work required careful preparation, and was consequently quite small (with less than a hundred authors; and less than half a million words). Here, we provide an initial report on the classification accuracies that can be achieved when classifiers conditioned on the small corpus are applied to a larger, automatically-acquired blog corpus, using lower-granularity personality data and substantially less manual preparation (with over a thousand bloggers, and approximately five million words). Predictably, results on the larger corpus are not as impressive as those on the smaller; nevertheless, they point the way forward for further work.

Keywords

Personality; Computational linguistics; Automatic classification; Corpus analysis

1. Introduction

The existence of ICWSM proves that academic interest in weblogs continues to grow. Weblogs can be studied for at least two reasons. First, it might be to uncover interesting information about both weblogs as multimedia texts and webloggers as online authors and conversationalists. Secondly, it might be because they provide a rich, ready and revealing source of highly varied text written by individuals who also choose to publish useful collateral information about themselves. The current work studies weblogs (and more specifically, personal weblogs, or 'blogs') for the second reason. We are primarily interested in individual differences and how they are revealed in language use. Personal weblogs, as a genre as yet relatively unrestricted by rules or common expectations, offer authors considerable personal freedom, and hence, much variation in style is visible.

But there are practical applications for this work. Within computational linguistics, a strand of recent work has attempted sentiment analysis and classification for instance. Pang and Lee [26] and Turney [32] have addressed the thumbs up/thumbs down decision: can the sentiment orientation (positive or negative) of a product review be accurately estimated from the text (in its entirety or in summary)? Sentiment analysis on blogs has already been attempted. Mishne [20] reports on the task of classifying the primary mood of weblog postings. Oberlander and Nowson [25] report on the task classifying the personality of bloggers from their postings. On the most straightforward binary classification task (see Section 2.3), they achieve accuracies of between 75% and 84%, against a (majority) baseline accuracy usually around 50%. They conclude that "if we spot a thumbs-up review in a weblog, we should be able to check other text in that weblog, and tell whose thumb it is; or more accurately, what *kind* of person's thumb it is, anyway. And that in turn should help tell us how high the thumb is really being held."

However, as we shall shortly describe, [25] use a small, carefully constructed corpus for their work, and one might doubt whether the results will scale up to the blogosphere proper, both because their original corpus may be unrepresentative, and because the method requires too much careful text processing to be practical on large collections of blogs. We have therefore assembled a new, substantially larger corpus for research purposes. A natural next step, then, is to ask what levels of classification accuracies can be achieved when classifiers conditioned on the original corpus are applied to classification tasks on the new corpus. This paper addresses this question—using only a subset of our new blog corpus—and points forward to the next steps we will be taking.

2. Background

2.1 Personality

The pioneering work of Cattell [6] led to the isolation of 16 primary personality factors. Later work on secondary factors led to Costa and McCrae's five-factor model [8], which is closely related to the 'Big Five' models which emerged from lexical research [10, 12]. Each personality factor gives a continuous dimension for scoring. They can be defined by their facets [18]: *Neuroticism* (anxiety, angry hostility, depression, self-consciousness, impulsiveness, and vulnerability); *Extraversion* (warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotion); *Openness to experience* (fantasy, aesthetics, feelings, actions, ideas, and values); *Agreeableness* (trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness); and *Conscientiousness* (competence, order, dutifulness, achievement striving, self-discipline, and deliberation)

*Formerly at University of Edinburgh.

2.2 Language and style

There is a respectable body of work investigating the relationship between language and personality (eg. [30, 9]). However, language generally means speech, while for personality only Extraversion, and to a lesser extent Neuroticism, have been studied at any length. Looking at writing using a full set of personality traits, Pennebaker and colleagues secured significant results using the Linguistic Inquiry and Word Count text analysis program [27]. Primarily, this tool calculates the relative frequencies of word-stems in pre-defined semantic and syntactic categories. It shows, for instance, that high Neuroticism scorers use: more first person singular and negative emotion words; and fewer articles and positive emotion words [28]. Building on this, Oberlander and Gill recently used a bottom-up stratified corpus comparison technique, which shows, for instance, that high Neuroticism scorers tend to use collocations involving multiple punctuation, articles, inclusives and conjunction [24].

Language has been explored within the context of blogs to investigate similar concepts: gender-based language has been studied in the weblogs of teenagers [14] and comments made to weblog posts [15], and explored alongside age [31]; Mood has been explored in weblogs as a response to a traumatic event [7] and for identifying trends [20, 2]. In exploring the differences between *filter* and *personal* weblogs, Herring and Paolillo [13] have shown the former to use language considered more characteristically male, while the latter showed more female language traits, regardless of author gender. Linguistic analysis has also shown personal blogs to be less contextual than email [23] though they share a similar factor structure [11].

2.3 Classification

Perhaps the most relevant work here is the small but growing collection on the automatic classification of personality (to which this paper is an addition). Argamon et al. [1] focused on Extraversion and Neuroticism, dividing Pennebaker and King's [28] population into just the top- and bottom-third scorers on a dimension, discarding the middle third. Employing various feature sets, including function words, and systemic functional grammar analysis, they report a small improvement over the random baseline for binary classification accuracy.

Mairesse and Walker [16, 17] used features drawn from the LIWC, and for speech prosody information and utterance types to classify personality in corpora also from Pennebaker and colleagues [28, 19]. Not only did they investigate an even high/low split, but they also compared self-rated scores of personality to observer ratings. A number of their results proved statistically significant, and they showed that observer ratings can often prove more accurate to model than self-scores.

As mentioned earlier, Oberlander and Nowson [25] used simple language models based on n-grams to explore a limited corpus of personality labeled weblogs. As this work is the basis for the current study, more details will be given in the course of this paper. The main point for now is that on a task similar to that investigated by Argamon et al., they gained accuracies of between 75% and 84%, depending on personality dimension—the best result being for Neuroticism.[1] But such results were gained on a relatively small, carefully processed blog corpus. Will it scale up?

[1] They report even higher accuracies, given automatic feature selection, but this is likely due to overfitting.

3. Blog Corpora

In this paper we are working with two separate corpora of weblogs: the first, collected for [21] is referred to as the original corpus (OC); the second, reported for the first time here, is referred to as the new corpus (NC).

3.1 Small and clean

In the original corpus, the emphasis was placed on the quality of the data. This required that participants provided accurate personality scores, and that text be as clean as possible. Participants were recruited directly via e-mail to suitable candidates, and indirectly by word-of-mouth: many participants blogged the study. Participants were first required to answer sociobiographic and personality questionnaires. The personality instrument has specifically been validated for online completion [4]. It was derived from the 50-item IPIP implementation of Costa and McCrae's [8] revised NEO personality inventory; participants rate themselves on 41-items using a 5-point Likert scale. This provides scores for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness.

Participants were also requested to submit one month's worth of prior postings. The month was specified to reduce the effects of participant choosing what they considered their 'best' or 'preferred' month. Raw submissions were marked-up using XML so as to automate extraction of the desired text. Text was marked by post type, such as purely personal, commentary-like reporting of external matters, or direct posting of internet memes such as quizzes, and features such as links or quotes were labeled. The corpus consisted of text from 71 participants (47 females, 24 males; average ages 27.8 and 29.4, respectively). In order to explore individual difference as much as possible only the authored text marked as 'personal' from each weblog was extracted, approximately 410,000 words. To eliminate undue influence of particularly verbose individuals, the size of each weblog file was truncated at the mean word count plus 2 standard deviations.

3.2 Big and dirty

The approach described above required participants to voluntarily complete a detailed personality questionnaire, and data annotation of just 71 texts was considerably time consuming. In considering sources of more data the amount of cleaning that can realistically be performed needs to be considered.

An internet meme was identified that was equivalent to a personality test, which had been taken by bloggers numbering in the thousands. The questionnaire consists of five items for each of the five factors of personality employed in the previous study. The items are simple yes/no questions and so personality scoring is far more coarse: 4-5 yes answers is labeled high; 2-3 is labeled medium; and 0-1 low. Despite the untraceable origins and non-validated nature of the questionnaire, the items appear to be fairly standard markers for the big 5 model of personality [5].

All bloggers who took this test were identified by their linking to the source. Once URLs were identified, it was possible to acquire blog text for February through June from Nielsen BuzzMetrics' blog data. This data is tagged in a manner similar to the ICWSM corpus.

For this initial exploration of the NC data, we are using only the text written in February. With data running into thousands of bloggers and tens of thousands of posts, the manual cleaning conducted on the OC is not possible. However, there are some obviously dirty elements that can be removed:

- In terms of personal text, the majority of noise comes from the posting of memes. The majority of sites that provide these deliver your results with HTML table code to copy and paste onto your site. Anything within a post between <table> tags is removed.
- Non-author text is also a problem; in an attempt to reduce this form of noise, text within <blockquote> tags is removed.
- All other HTML tags are then removed.
- Many memes take the form of lists not contained within tags (e.g. 30 things about me; Top Ten Favourite Blog Conferences). These are removed by identifying sequences of four consecutive numbers within a text.

Note that there is no guarantee that these approaches remove all instances, nor that they do not remove text unnecessarily. With the volumes of data we are dealing with however, it was decided that some easily replicable, automatic pre-processing was better than potentially more reliable, but harder to replicate pre-processing requiring human intervention. And it is very likely better than none.

A further consideration is that of length. Overly verbose texts can exert undue influence, and those with very little text in which to discover features result in sparse data. Therefore it was decided only to explore blogs that provided more than 1000 words over a month, but to cap those blogs with more than 5000 words at the 5000 mark.

With these measures employed, this February-only version of NC consists of 1672 bloggers, and 4.8M words (mean = 2878).

3.3 Open blogger hypothesis

It is a common misconception that bloggers are exhibitionist narcissists. In terms of personality traits, there is a general assumption that they are Extraverts. As we have previously reported [22], there did not appear to be any bias in the OC for Extraversion. However, there was a significant bias on the Openness dimension in favour of higher scores. So much so, that there were no subjects who could realistically be considered low scorers. Due to the stratified approach of our analysis (see section 4.3), this meant it was not possible to explore this dimension further. However, without a collecting suitably comparable data on non-bloggers, it was not possible to say with authority that bloggers are generally Open individuals; it could merely be an artifact of those individuals who chose to submit to the analysis.

However, it is interesting to note the distribution of Openness scores within the NC. We find that whilst all other traits have a reasonable approximation of a normal distribution,[2] Openness seems to follow the findings of the previous study. As can be seen in figure 1, low scorers make up just 5% of the population, while high scorers account for 62% (leaving the remaining 33% in the middle). Again, this by no means proves the Open blogger hypothesis, but certainly appears to lend it support.

[2] At least as reasonable as one can expect from just three classes.

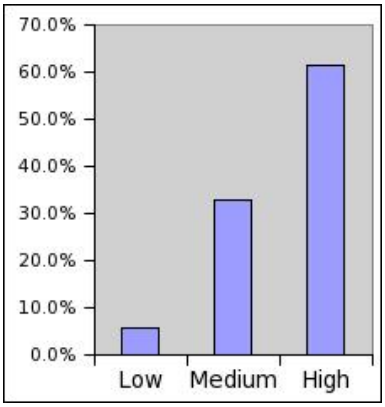

Fig. 1: *Distribution of Openness scores in NC*

4. Methodology

The main aim of this study is to explore the application of language models and classifiers developed using the cleaner and more reliably defined OC data to the more coarse-grained and noisier NC data. In this section we first describe the language models constructed from our previous analysis [25] and how these are to be applied in the current situation. The mechanics of the classification to be conducted will then be described. This leads to a discussion of the various training sets of the OC, which are used to construct our classifiers.

4.1 Language models

There are many potential features that can be used for text classification. Our previous analysis [25] used essentially word-based bi- and tri-grams. It should be noted that some generalisations were made in this analysis: all proper nouns were identified via CLAWS tagging using the WMatrix tool [29], and replaced with a single marker (NP1); punctuation was collapsed into a single marker (<*p*>); and additional tags correspond to non-linguistic features of blogs—for instance, <*SOP*> and <*EOP*> were used the mark the start and end of individual blog posts. Word n-gram approaches provide a large feature space with which to work. But in the interests of computational tractability, it is useful to reduce the size of the feature set. There are automatic approaches to feature selection which rely on a statistical analysis of those features which best aid classification. However, our work is motivated by the study of personality, and so ideally we wish to explore those features which can be shown to relate well to traits. We employ increasing restrictive approaches to feature set reduction to create our language model; for each dimension, we have four potential models:

I The least restricted feature set consists of the n-grams most commonly occurring within the OC. Therefore, the feature set for each personality dimension is to be drawn from the same pool. The difference lies in how many features are selected: this will match that of the next level of restriction.

	I	II	III	IV
N	747	747	169	22
E	701	701	167	11
A	823	823	237	36
C	704	704	197	22

Table 1: *Number of n-grams per model*

II The next set includes only those n-grams which were distinctive for the two extremes (high and low) of each personality trait. Only features with a corpus frequency $\geq$5 are included to allow accurate calculation of log-likelihood G^2 statistics [29]. Distinct collocations are identified via a three way comparison between the H and L groups in training set 1 (see section 4.3.1) and a third, neutral group. This neutral group contains all those individuals who fell in the medium group (M) for *all four traits in the study*; the resulting group was of comparable size to the H and L groups for each trait. Hence, this approach selects features using only a *subset* of the corpus. N-gram software was used to identify and count collocations within a sub-corpus [3]. For each feature found, its frequency and relative frequency are calculated. This permits relative frequency ratios and log-likelihood comparisons to be made between High-Low, High-Neutral and Low-Neutral. Only features that prove distinctive for the H or L groups with a significance of $p < .01$ are included in the feature set.

III The next set takes into account the possibility that, for a group used in Model-II, an n-gram may be used relatively frequently, but only because a small number of authors in a group use it very frequently, while others in the same group use it not at all. For example a single author might use the same catchphrase in every post. To enter the Model-III set, an n-gram meeting the Model-II criteria must also be used by at least 50%[3] of the individuals within the subgroup for which it is reported to be distinctive.

IV While Model-III guards against excessive individual influence, it may abstract too far from the fine-grained variation *within* a personality trait. The final manual set therefore includes only those n-grams that meet the Model-II criteria with $p < .001$, meet the Model-III criteria, and also correlate significantly ($p < .05$) with individual personality trait scores.

Since different sub-groups are considered for each trait, the feature sets which meet the increasingly stringent criteria vary in size. Table 1 lists the size of each of the model feature sets for each of the four personality traits. Note again that the number of n-grams selected from the most frequent in the corpus for Model-I matches the size of the set for Model-II.

These feature sets were derived from the OC data, and define the different language models to be used to explore the relationship between personality and language. A similar n-gram counting approach [3] was used to calculate statistics for the NC data. Relative frequencies were extracted for just those n-grams in each of the 16 models.

[3] Conservatively rounded down in the case of an odd number of subjects.

4.2 Classification

The current paper is the second step in our work on personality classification, and is a direct follow up to our previous efforts [25]. With this in mind, classification methodology will be similarly simple. Simple comparisons performed previously showed that naïve bayes (NB) outperformed support vector machines (SVM) on the majority of our tasks. We consider refining the SVM parameters as an important task for future work, but here we again employ NB, (as implemented in the WEKA toolkit [33]) so as best to compare results with our previous work.

In this work, we test both the suitability of the language models (described in section 4.1) and classifiers derived from clean data (described in section 4.3) for classifying dirty data. These are distinct because the language models were derived from a subset of the OC data, whereas the classifiers rely on the full dataset. The current experiment is therefore carried out in two stages: the first simply applies the old language models to the NC data; the second applies the classifiers.

As described in section 3.2 the personality data we are looking to classify has three classes: low, medium and high. This lends itself well to two classification tasks: the easier binary task, distinguishing between high and low scorers (here, we call it 'lh'); and the harder 3-class task, classifying over all groups (here, we call it 'lmh'). In the first (model) analysis, the n-grams used in the feature sets are chosen on the basis of the OC data, but 10-fold cross validation is used on the n-gram relative frequencies from the NC data; thus training and testing on the same dataset, but no further tuning of the language models.

The second analysis explore the classifiers from the OC data. In order to do this, the OC data is used to train the automatic classifier which will then be tested on the NC. This is, of course, a less than desirable balance: one would typically train on 2/3 of data points. Here, training will be on at most just over 4% of the total subject base. The personality data from the OC data to be used as training, as described in section 3.1, is much finer-grained than the data to be used for testing. This finer granularity leads to a number of ways of stratifying the corpus, thus creating a number of possible training sets. These will be described in the next section.

4.3 Training Sets

For any blog in the OC, we have available the scores of its author on four continuous personality dimensions. However, the NC is clearly stratified into just low, medium and high groups. Whilst the coarser-grained data is less flexible, it is suitable for machine classification. The simplest task is a binary classification between high and low, with the harder task incorporating the medium group. In order to compare results between the datasets, however, it is necessary to divide the OC authors similarly. This can be done in a number of ways, and this section describes the various training sets to be used to investigate the classifiers of the OC.

4.3.1 Binary classification

The aim of training with these sets is to construct a classifier which can distinguish authors as either high or low scorers of a personality trait. There are a number of ways to split a continuous dimension into binary classes:

1. The simplest approach is to keep the high and low groups as far apart as possible: high scorers (H) are those whose

scores fall above 1 SD above the mean; low scorers (L) are those whose scores fall below 1 SD below the mean.

2. Training set 1 creates distinct groups, at the price of excluding over 50% of the OC from the experiment. To include more of the corpus, parameters are relaxed: the high group (HH) includes anyone whose score is above .5 SD above the mean; the low group (LL) is similarly placed below.

3. The most obvious task (but not the easiest) arises from dividing the OC in half about the mean score. This creates high (HHH) and low (LLL) groups, covering the entire population. Inevitably, some high scorers will actually have scores much closer to those of low scorers than to others from their own class.

These sub-groups are tabulated in Table 2, giving the size of each group within each trait. Note that in training set 2, the standard-deviation-based divisions contain very nearly the top third and bottom third of the population for each dimension. Hence, training set 2 is closest in proportion to the division by thirds used by Argamon et al. [1].

	Lowest	...	Highest
1	L	–	H
2	LL	–	HH
3	LLL		HHH
N1	12	–	13
N2	25	–	22
N3	39		32
E1	11	–	12
E2	23	–	24
E3	32		39
A1	11	–	13
A2	22	–	21
A3	34		37
C1	11	–	14
C2	17	–	27
C3	30		41

Table 2: *Binary training set groups: division method and author numbers. N = Neuroticism; E = Extraversion; A = Agreeableness; C = Conscientiousness*

4.3.2 Three-class classification

In our previous study ([25]), we reported results from 4 further tasks representing multi-class classification. Though only two of these are suitable for comparison with the NC data, for consistency we retain their numbering (5 and 6) for the training sets here.

5. Takes the greatest distinction between high (H) and low (L) groups from training set 1, and includes the medium (M) scorers.

6. Similarly, following training set 2, this uses the larger high (HH) and low (LL) groups, with those between forming a smaller medium (m) group.

These sub-groups are tabulated in Table 3, giving the size of each group within each trait.

	Lowest	...	Highest
5	L	M	H
6	LL	m	HH
N5	12	46	13
N6	25	24	22
E5	11	48	12
E6	23	24	24
A5	11	47	13
A6	22	28	21
C5	11	46	14
C6	17	27	27

Table 3: *3-class training set groups: division method and author numbers. N = Neuroticism; E = Extraversion; A = Agreeableness; C = Conscientiousness*

5. Results

As described in the previous sections, there are two stages to the analysis to be reported here. The first explores the use of just the four language models (section 4.1) for 2- and 3-class classification of the NC data. The second analysis is similar, only the classifier is trained on OC data: as noted in section 4.3 there are a number of ways of stratifying the subjects. These different training sets are reported separately.

5.1 Model Analysis

This section reports the results of the language model analysis which uses 10-fold cross validation on the NC data for binary (-lh) and 3-class (-lmh) classification for each personality trait. Table 4 shows the raw accuracies for each model, with the most successful highlighted in bold.

Training set	Mod. I	Mod. II	Mod. III	Mod. IV
N-lh	**59.2**	50.9	50.4	53.0
N-lmh	38.2	34.5	35.3	**39.0**
E-lh	**56.0**	54.1	51.9	48.0
E-lmh	33.6	32.5	34.7	**38.6**
A-lh	53.2	**56.3**	53.4	55.2
A-lmh	32.7	**35.6**	35.0	32.8
C-lh	59.0	54.6	60.1	**66.4**
C-lmh	36.6	35.5	37.9	**41.2**

Table 4: *Naïve Bayes performance with four language models for 2- and 3-class classification of personality (random base line for -lh=50%, for -lmh=33.3%)*

The first, unsurprising, observation is that the results are uniformly poorer than those we previously derived on the OC itself. And in every case, the easier binary (lh) classifications show greater improvement over the baseline than the 3-class cases. Somewhat surprisingly, the best of the results for Neuroticism and Extraversion are achieved using Model-I. As described in section 4.1, this model was created solely from the most frequent n-grams of the corpus, with no statistical relationship to personality. However, for the 3-class (lmh) classifications of Neuroticism and Extraversion, it is Model-IV—the smallest and most restricted language model—which proves the most successful.

For Agreeableness and Conscientiousness, there is more

consistency, and more satisfactory results. Model-II provides the best results for both levels of classification of Agreeableness, while Model-IV results in the highest raw classification accuracies overall, for Conscientiousness. Compared with a random baseline of 50% on a binary task, 66.4% is somewhat respectable.

5.2 Classifier Analysis

This section reports the results of the classifier analysis. This trains the automatic classifiers on the training sets of the OC and tests these with NC data. For each personality trait, there are three binary training sets (1–3) and two 3-class (5 and 6). Table 5 shows the raw accuracies for each model, with the most successful highlighted in bold. The model analysis results from table 4 are also included, in italics, for ease of comparison.

Training set	Mod. I	Mod. II	Mod. III	Mod. IV
N-lh	*59.2*	*50.9*	*50.4*	*53.0*
N1	49.0	51.7	**56.3**	54.7
N2	51.3	50.8	50.8	**53.9**
N3	49.5	53.3	53.7	**55.8**
N-lmh	*38.2*	*34.5*	*35.3*	*39.0*
N5	39.6	**40.2**	39.7	35.7
N6	35.2	33.7	33.6	**36.2**
E-lh	*56.0*	*54.1*	*51.9*	*48.0*
E1	**55.4**	50.6	50.7	51.0
E2	52.6	**53.0**	52.2	51.4
E3	52.6	**52.8**	51.3	50.6
E-lmh	*33.6*	*32.5*	*34.7*	*38.6*
E5	**44.2**	42.1	42.6	36.1
E6	35.5	33.2	34.5	**36.5**
A-lh	*53.2*	*56.3*	*53.4*	*55.2*
A1	47.9	**61.6**	56.0	49.6
A2	44.9	**52.7**	46.0	45.7
A3	49.5	50.9	49.6	**52.9**
A-lmh	*32.7*	*35.6*	*35.0*	*32.8*
A5	44.8	46.2	**46.6**	41.6
A6	34.5	35.8	**36.6**	32.9
C-lh	*59.0*	*54.6*	*60.1*	*66.4*
C1	60.0	**64.2**	60.8	49.3
C2	59.6	**64.8**	59.4	51.7
C3	**59.1**	57.7	54.6	56.6
C-lmh	*36.6*	*35.5*	*37.9*	*41.2*
C5	46.4	47.2	**47.4**	36.8
C6	**38.4**	37.0	38.0	34.3

Table 5: *Naïve Bayes performance with four language models for 2- and 3-class classification trained on clean data (random base line for -lh,1-3=50%, for -lmh,5,6=33.3%)*

Consider first the binary classifiers. Except for Conscientiousness, OC training set 1 is best for binary classifiers. However, language models derived from OC but not tuned on OC data outperform these classifiers on three of the four dimensions. The exception is Agreeableness where training set 1 (which used only the most extreme scorers in the original corpus) and Model-II combine to give an accuracy of 61.6%, 11.6% above a random baseline. Training set 1 gives classifiers that always outperform those using training set 3; and its classifiers also outperform those using training set 2 in three out of four comparisons. Model-II gives the best feature set overall.

For 3-way classifiers, training set 5 (with small extreme groups and a large mid-group) always gives better performance than training set 6. Moreover, these classifiers outperform those derived by tuning the OC language models on NC. Model-II is no longer the best, however; Model-III is superior. The best 3-way result is for Conscientiousness, where training set 5 and Model-III combine to give an accuracy of 47.4%, 14.1% above a random baseline.

Although absolute accuracies are better for binary than for 3-way classification, compared to a random baseline, the relative improvement appears better for 3-way classification, particularly for Conscientiousness.

5.3 Discussion

These results overall are predictably poorer than those where language models and training sets derived from OC were tuned, trained and cross-validation tested on OC [25]. However, they show at least some promise. From the model analysis, it appears that Model-IV, the smallest and most restrictive, shows the greatest relationship with personality in general. Similarly, from the classifier analysis, it appears that training sets 1 and 5, those with the most extreme (most restricted) personality groups, also generalise best. Or perhaps it would be more accurate to say that these models and training sets are merely the least bad among a poor bunch.

Obviously, if we build language models from scratch, based on what is found in NC, we should expect better results; and similarly, if we use training sets derived from NC, the picture should be brighter. In future work, we will explore these options, and take advantage of our larger corpus to, for instance, train on February data, and test on March data. In addition, we expect to experiment with a broader range of learning algorithms than naïve bayes, focusing particularly on support vector machines. And for future testing, we will set out to used balanced test sets rather than relying on random baselines.

6. Conclusions

Despite the comparatively disappointing results, it looks for now as if the more automatic processing required to handle the larger corpus has not introduced so much noise that personality classification has become hopeless. Indeed, it is our most finely tuned models and classifiers that seem to suffer least in the scaling up procedure. So, big and dirty may be difficult, but it is not impossible.

Acknowledgments

We gratefully acknowledge the data collection assistance provided by Natalie Glance and Matt Hurst at Nielsen BuzzMetrics: as should be obvious, this study would have been impossible without it. The first author was funded through earlier stages of this work by the UK Economic and Social Research Council, and latterly by an award from the Edinburgh-Stanford Link.

References

[1] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the*

Interface and the Classification Society of North America, 2005.

[2] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *11th Meeting of the European Chapter of the Association for Computational Linguistics*. EACL, 2006.

[3] S. Banerjee and T. Pedersen. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, 2003.

[4] T. Buchanan. Online implementation of an ipip five factor personality inventory. Available at `http://users.wmin.ac.uk/~buchant/wwwffi/introduction.html`, *accessed 07/10/06*, 2001.

[5] T. Buchanan. Personal Correspondence, 2006. received 23/11/06.

[6] R. B. Cattell. *Description and measurement of personality*. George Harrap, London, 1946.

[7] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker. Linguistic markers of psychological change surrounding september 11. *Psychological Science*, 15:687–693, 2004.

[8] P. T. Costa and R. R. McCrae. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Odessa, FL: Psychological Assessment Resources, 1992.

[9] J.-M. Dewaele and A. Furnham. Extraversion: The unloved variable in applied linguistic research. *Language Learning*, 49:509–544, 1999.

[10] J. Digman. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41:417–440, 1990.

[11] A. J. Gill, S. Nowson, and J. Oberlander. Language and personality in computer-mediated communication: A cross-genre comparison. *Journal of Computer Mediated Communication*, submitted.

[12] L. R. Goldberg. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34, 1993.

[13] S. C. Herring and J. C. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2004.

[14] D. Huffaker and S. L. Calvert. Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication*, 10(2), 2005.

[15] T. L. M. Kennedy, J. S. Robinson, and K. Trammell. Does gender matter? Examining conversations in the blogosphere. Paper presented at Internet Research 6.0: Internet Generations, Chicago, IL, October 5-9, 2005.

[16] F. Mairesse and M. Walker. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, New York City, 2006.

[17] F. Mairesse and M. Walker. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci)*, pages 543–548, Vancouver, 2006.

[18] G. Matthews, I. J. Deary, and M. C. Whiteman. *Personality Traits*. Cambridge University Press, Cambridge, 2nd edition, 2003.

[19] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90:862–877, 2006.

[20] G. Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, 2005.

[21] S. Nowson. *The Language of Weblogs: A study of genre and individual differences*. PhD thesis, University of Edinburgh, 2006.

[22] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. *AAAI Spring Symposium, Computational Approaches to Analysing Weblogs*, Stanford University., 2006.

[23] S. Nowson, J. Oberlander, and A. J. Gill. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671, Hillsdale, NJ, 2005. Lawrence Erlbaum Associates.

[24] J. Oberlander and A. J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42:239–270, 2006.

[25] J. Oberlander and S. Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, Sydney, Australia, 2006.

[26] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[27] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic Inquiry and Word Count 2001*. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.

[28] J. W. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312, 1999.

[29] P. Rayson. *Wmatrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University, 2003.

[30] K. Scherer. Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social Markers in Speech*, pages 147–209. Cambridge University Press, Cambridge, 1979.

[31] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 2006.

[32] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 417–424, 2002.

[33] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.