

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Large Scale Personality Classification of Bloggers

Citation for published version:
Iacobelli, F, Gill, A, Nowson, S & Oberlander, J 2011, Large Scale Personality Classification of Bloggers. in S D'Mello, A Graesser, B Schuller & J-C Martin (eds), Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II. Lecture Notes in Computer Science, vol. 6975, Springer-Verlag GmbH, pp. 568-577. https://doi.org/10.1007/978-3-642-24571-8_71

Digital Object Identifier (DOI):
10.1007/978-3-642-24571-8_71

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Affective Computing and Intelligent Interaction

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 19. Mar. 2024

Iacobelli, F., Gill, A., Nowson, S., & Oberlander, J. (2011). Large Scale Personality Classification of Bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J-C. Martin (Eds.), Affective Computing and Intelligent Interaction. (pp. 568-577). (Lecture Notes in Computer Science). Springer Berlin / Heidelberg. 10.1007/978-3-642-24571-8_71

Large Scale Personality Classification of Bloggers

Francisco Iacobelli[1], Alastair J. Gill[2], Scott Nowson[3], and Jon Oberlander[4]

[1] Northeastern Illinois University, Chicago, Illinois 60625 USA.
f-iacobelli@neiu.edu
[2] University of Surrey, Guildford, Surrey GU2 7HX, UK
A.Gill@surrey.ac.uk
[3] Appen Pty Ltd, Chatswood NSW 2067, Australia
snowson@appen.com.au
[4] University of Edinburgh, Edinburgh, EH8 9AB, UK
J.Oberlander@ed.ac.uk

Abstract. Personality is a fundamental component of an individual's affective behavior. Previous work on personality classification has emerged from disparate sources: Varieties of algorithms and feature-selection across spoken and written data have made comparison difficult. Here, we use a large corpus of blogs to compare classification feature selection; we also use these results to identify characteristic language information relating to personality. Using Support Vector Machines, the best accuracies range from 84.36% (openness to experience) to 70.51% (neuroticism). To achieve these results, the best performing features were a combination of: (1) stemmed bigrams; (2) no exclusion of stopwords (i.e. common words); and (3) the boolean, presence or absence of features noted, rather than their rate of use. We take these findings to suggest that both the structure of the text and the presence of common words are important. We also note that a common dictionary of words used for content analysis (LIWC) performs less well in this classification task, which we propose is due to their conceptual breadth. To get a better sense of how personality is expressed in the blogs, we explore the best performing features and discuss how these can provide a deeper understanding of personality language behavior online.

Keywords: Machine Learning, Personality Classification

1 Introduction

Personality traits, which are intimately linked to affect [3], and their detection is of high interest for systems that target users by personalising content (e.g., online stores, recommender systems, social media and search engines; cf. [22]). Personal weblogs (blogs) are a popular way to write freely and express preferences and opinions on anything that is of interest to the author [9], and therefore provide a useful resource for investigating personality. Indeed, personality has been shown to relate to writing style in blogs and more generally (e.g.,[19, 17, 16, 25]), blogger motivation [6], as well as influencing the content that a user prefers to read [23]. However, despite studies indicating lingusitic cues of personality, attempts to classify personality from essays or emails have yielded modest results [11, 4]. In the case of blogs, although classification of author personality

has been successful on small corpora, the performance of features on larger corpora has degraded, possibly as a result of overfitting [14].

Direct comparison between these previous personality classification studies is difficult given inconsistencies in algorithms, feature-selection, and data sources (ranging from speech to essays, emails and blogs). Thus, in this paper we use a very large corpus of blogs with associated author personality information to provide the first systematic comparison of feature sets used by machine learning algorithms in the task of personality classification. In addition, by exploring the features that are more informative for a classifier, we are able to build a deeper picture of how personality behavior is *actually* realised linguistically.

In the following section, we review previous findings of studies exploring and classifying personality and language. In the method section, we describe in more detail the feature and data sets used in this comparison study. We then present our classification results and discussion, where we also present some of the most predictive bigram features relating to the different personality types. The presentation of such features is important, since they can increase our understanding of how personality is expressed in language. We conclude the paper with a summary of our main findings, future directions and pointers to build a better theoretical understanding of personality and its relationship with language.

2 Background

Like other studies relating to personality and language (e.g. [16, 19]), we adopt the five-factor model of personality [2], which describes the following traits on a continuous scale: neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. General behaviors characteristic of high and low scorers for each of the traits are listed in Table 1, along with previous findings for personality and language.

These previous results reported for personality and language relate specifically to written language (although other studies have examined speech, e.g., [12]), and have applied both data-driven words and phrases (e.g., [16, 13]) and lexical features grouped by psychological categories (e.g., [19, 6]). Both approaches have been applied to classification tasks for personality. Studies using psychological groupings of lexical features have adopted the LIWC dictionary (Linguistic Inquiry and Word Count; [18]) to compare baseline classification algorithms to Naïve Bayes and SMO (sequential minimal optimization algorithm for a support vector classifier [24, 20]). Applying such feature groupings to written texts [1] obtained optimal results of around 57-60% accuracy for extraversion and neuroticism using SMO; similar analysis of conversational data resulted in accuracies of around 65% [11].

Other features have included structural and lexical features of corpus of email (approximately 9,800 messages) which were used to classify several dimensions, including the five personality traits [4]. Although accuracies were in the range of 53–57%, this study used the largest corpus for personality classification of which we are aware.

In another case, n-grams were used to assemble features for classifying four of the five personality traits: in a small corpus of blogs, SMO gave the best accuracies (of around 83–93%) [17]. However, when these trained classifiers were applied to a much

Table 1. Personality traits –neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A) and conscientiousness (C)– with behavioral and linguistic characteristics.

Trait	Type	High Score	Low Score
N	Behavior	Emotional instability; anxious; hostile; prone to depression	Emotional stability; calm; less easy upset
	Linguistic	*Use of first person singular and negative emotion words (on essays) [19]; talk of discrepancies, jobs, and physical states (on blogs) [13]; exclusive and inclusive connectives, use of multiple-punctuation expressions (on emails) [16].*	*Use of refererences to other people (on blogs) [13]; more nouns and adverbs (email) [16].*
E	Behavior	Extraverts; warm; assertive; action-oriented; thrill-seeking	Introverts; low key; deliberate; easily stimulated
	Linguistic	*Use of social words, self and other references, positive emotion words, greater certainty (on emails and essays) [16, 19]; greater complexity, conjunctions and adjectives (on email) [16]; present tense verbs, references to communication (on blogs) [13].*	*Use of negations and negative emotion expressions, exclusive, inclusive, causation words, articles (on essays) [19]; greater tentativity (on email) [16]; achievements, discrepancies (on blogs)[13].*
O	Behavior	Appreciate art and ideas; imaginative; aware of feelings	Straightforward interests; conservative; resist change
	Linguistic	*Use of articles, longer words and insight words (on essays) [19]; use longer words, express positive feelings, inclusive words (on blogs) [13].*	*use of first person singular, present tense, and causation words (on essays) [19]; negations, references to school (on blogs) [13]*
A	Behavior	Compassionate; cooperative; considerate; friendly	Suspicious; unfriendly; wary; antagonistic; uncooperative
	Linguistic	*Use of first person singular, positive emotion words (on essays) [19].*	*Use of articles, negative emotion words (on essays) [19]; discrepancies, talk about body states (on blogs) [13].*
C	Behavior	Disciplined; dutiful; persistent; compulsive; perfectionist	Spontaneous; impulsive; achievement less important
	Linguistic	*Use of positive emotion words (on essays) [19].*	*Use of negations, negative emotion, causation, exclusive words, discrepancies (essays) [19]; topics concerned with death (on blogs) [13].*

larger corpus, the accuracies dropped to approximately 55%, which may have been a result of overfitting [14].

As these studies show, a variety of features and algorithms have been applied to personality classification tasks, however their application to different data sets makes comparison difficult (see e.g., [15] which compares the contextuality of blogs against genres of the British National Corpus) . Therefore, in the following study, we use a single, large blog corpus upon which to compare a variety of features for personality classification. We aim to be able to identify the best features for classification and also describe how these features give us new insight into personality as expressed through written language.

3 Method

This study is concerned with the linguistic characteristics of personality (rather than structural or design features of blogs), and therefore compares feature sets derived from 1- and 2-grams. In addition, we include features based upon psychological categories (from LIWC [18]), as they are extensively used in previous studies, to compare them to n-gram derived features.

3.1 Data Preparation

The corpus used was drawn from a large collection of around 3000 bloggers writing over several months. The corpus was processed to give one file per author per month. Each file contains all the postings for each author in each month. HTML tags, embedded and quoted text were removed. Each author completed a self-administered on-line personality questionnaire with five items measuring each of the Big 5 personality types. The items are simple yes/no questions and so personality scoring was rather coarse. The questionnaire gave low, middle and high scores for each trait (for more details see [14])

For inclusion in subsequent analysis, authors had to write a minimum of 1000 words in a month, with any month's contribution capped at 5000 words. When authors contributed in more than one month, their most recent month was used. Following the approach of Argamon et al. [1], only authors who scored high or low on these personality dimensions were included for analysis. Mairesse [11] also tested this approach and reported a 2–3% increase in overall accuracy scores compared to datasets that included middle scorers. However, he suggests that removing the middle scorers "potentially [increases] precision at the cost of reducing recall." Because case data is gathered from online sources, large quantities of data are more likely to result in problems of low precision than low recall.

Finally, to prepare the data for classification, we balanced the size of the high-low groups for each trait, by randomly discarding authors from the larger set to match the number in the smaller set. The number of authors originally within each class, along with the total used in experiments, can be seen in Table 2.

3.2 Feature Selection

For each of the several data sets compared, texts were further processed as follows: (a) words were stemmed using Porter's stemming algorithm [21]; (b) *proper names* (naive

Table 2. Number of authors in each class considered for each personality trait by level. Numbers applied to both high *and* low groups and used for experimentation are in **bold**.

Level	N	E	O	A	C
High	**553**	669	1465	892	884
Low	840	**637**	**137**	**372**	**323**
Total	**1106**	**1274**	**274**	**744**	**646**

detection of continuous sets of words with initial letters capitalised) were replaced by a common token; (c) *laughter* (variants, of different lengths and spellings, of *haha*, *hehe*, etc.) were also replaced by a common token; and (c) *apostrophes* (words containing an apostrophe were tagged).

Data sets were built for each personality trait using all variations of the following features: (a) words window size. Namely single words (size 1) vs. bigrams (size 2) used by five or more authors; and (b) including stop words ("sw") or omitting ("wo") stop words. In addition, each of these features were represented using one of two scores: (i) boolean (score of 1 or 0 to represent the presence or absence of a word) or (ii) importance (TF*IDF scores for each word). Combining features and weighting schemes resulted in 8 data sets per personality trait. In addition, we created one extra data set per trait not based on individual words, but on the psychologically defined categories of the Linguistic Inquiry and Word Count (LIWC) tool [18], as implemented using TAWC [10].

By deriving a number of different data sets, we are able both to compare their features in terms of relevance to the task of personality classification, and also to derive a greater understanding of the linguistic behavior of different personality types. Further, by implementing some of the classification and feature extraction techniques used in analogous data sets [13, 11, 1] we can also begin to understand the relative utility of the different approaches.

After building the data sets, Weka's Cfs-Subset selector [8] with Subset forward selection [7] was applied to each one in order to include only the features that contribute most to accurate classification.

To provide a comparison of features, texts were classified using the LibLinear [5] Support Vector Machines (SVM) classifier in Weka [24] (Weka's wrapper for LibLinear was faster and generally more accurate than the SMO, the standard Weka SVM). Following experimentation on a small training dataset we use the default parameters: $C = 1$; $\epsilon = 0.01$. In each case 10 fold cross validation was used to classify the data sets.

Baseline classification which assigns the majority class (ZeroR) produced 50% in each case since high-low groups were balanced. Classification using Weka's [24] default implementation of Naïve Bayes (NB) were in each case outperformed[5] by the SVM algorithm, and therefore are also omitted.

[5] Statistically significantly in most cases

4 Classification Results and Discussion

Table 3. Accuracy scores by trait and feature sets using SVM. Feature sets were coded as *(b)ool or (f)requency* scoring; *1 or 2* grams; include stopwords (sw) or not (wo). Classification with the LIWC feature set is also included.

Dataset	b-2-sw	b-1-sw	b-1-wo	b-2-wo	f-1-sw	f-1-wo	f-2-sw	f-2-wo	LIWC
N	**70.51**	70.47	70.12	67.77	67.77	67.25	67.83	65.24 ∘	59.56 ∘
E	**71.68**	67.80 ∘	68.40	63.99 ∘	64.84 ∘	65.22 ∘	69.42	64.15 ∘	54.86 ∘
O	**84.36**	81.44	79.14	77.49 ∘	74.74 ∘	74.74 ∘	77.93 ∘	73.77 ∘	56.86 ∘
A	**78.31**	69.98 ∘	69.49 ∘	71.09 ∘	66.01 ∘	64.78 ∘	72.61 ∘	68.07 ∘	61.09 ∘
C	**79.18**	75.17	72.74 ∘	76.41	68.79 ∘	68.34 ∘	73.87 ∘	73.98 ∘	56.11 ∘

∘ statistically significant degradation, $p < 0.05$ compared to **b-2-sw**

Table 3 compares the performance of SVM using the feature sets achieving greatest accuracy (boolean scoring with 2-grams including stop-words; "b-2-sw") with the performance of SVM on the other data sets. The table shows that for openness, when features contain a boolean scoring system, there are no significant differences in accuracy. However, when features contain TF*IDF scores, the accuracy of the classifier becomes significantly worse.

This suggests that openness can be inferred by checking for the occurrence of individual words. The accuracies for conscientiousness suggest that besides boolean scoring resulting in better accuracies, it may be that stop words are relevant for classification. Because the score with bigrams and excluding stopwords produced no significant difference from the accuracy obtained with bigrams including stopwords, it may be that when some structural information is captured (bigrams), stop words become less relevant for classification.

In the case of extraversion, the accuracy produced by bigrams with stop words using boolean scores was not significantly different from bigrams that included stopwords but were scored using TF*IDF. In addition, it was not different from unigrams that excluded stop words. Agreeableness accuracy scores show clearly that the mere presence or absence of bigrams that include stop words (bool-2-sw) produced the most accurate classification. Lastly, the accuracies for neuroticism did not vary significantly except in the case of the use of TF*IDF frequencies on bigrams that did not include stopwords. Also included in Table 3 are results from the data sets based on LIWC's thematic categories. As described earlier, the accuracy across all traits is a significant degradation from that of the best data sets. Considering words in context, as bigrams to a degree allow, is apparently a more reliable indicator of personality than thematic grouping of words. We therefore note that even though these thematic categories appear to be theoretically justifiable, for such a classification task, they appear to overgeneralize.

In sum, we notice that the presence or absence of bigrams including stop words produced the most accurate classification, although this difference was not always significant with respect to other methods.

4.1 Bigrams characteristic of personality

An analysis of the features that best classified our data provides a view of the linguistic features that reveal the bloggers' personalities. In this section we provide a first analysis of these features by presenting (a) their average precision of classification; and (b) the Big 5 personality traits and how they are classified by individual features (bigrams).

To examine which of the bigrams classified high or low scorers of each personality trait, we looked at each bigram ($bigram_i$) within the set of bigrams that best classified each trait (c.f. Section 3.2) and retrieved the set of documents (D_i) that contained it. Then, we counted how many of these documents corresponded to high and low scorers, the highest number determining which score level was more precisely classified by the bigram. We call this the "majority classification" for this bigram. For example, for openness, if the bigram *the hell* was present in 50 documents of which 14 were high scorers and 36 were low scorers, we conclude that the majority classification for this bigram is low scorers. We then compute precision using the following formula: $precision_i = \frac{|tp|}{|tp+fp|}$, where $precision_i$ is the precision score for the i^{th} bigram. In this context, true positives (tp) are the documents from D_i that correspond to the majority classification of $bigram_i$. $|tp|$ is, then, the size of the set of true positives. False positives fp is the set of documents from D_i that do not correspond to the majority classification. In other words, $D_i = tp + fp$. Note that we talk about precision, and not accuracy, because we are measuring fidelity only among the documents that contain each bigram.

Table 4 shows the mean precision for bigrams on each trait on each score level. For example, for neuroticism, the bigrams that best classified low scorers had, on average, a precision of 0.9. As we can see, the bigrams, when present, classified scores for each trait with high precision.

Table 4. Mean precision of classification of bigrams for low and high scorers on each personality trait.

Trait	High	Low
N	0.90	0.90
E	0.86	0.92
O	0.88	0.91
A	0.91	0.86
C	0.88	0.89

We can consider a representative subset of the features that classified each score level on each personality trait. For the present paper we did not analyse the words surrounding the bigrams. Therefore, they cannot be reliably grouped into sub categories within each trait (cf. Table 1). In addition, because the bigrams were stemmed, in some cases it is not possible to determine which words they correspond to. However, in some cases we can reverse the stemming to obtain the exact words (e.g. *onli i* = *only I*) or, at least, a naive, but plausible interpretation (*am excit* = *am excited*, versus other tenses

Table 5. Stemmed bigrams that drive classification

Trait	High	Low
N	hope.thei; punish.for; get.work; onli.problem; you.onli; depress.you; drunk.i; i.wasnt;	mental.togeth; be.sad; am.excit; we'v.had; reflect.on; then.look group.of; chose.to; the.winner
E	more.excit; i.hang; im.at; im.too; b**ch.i; danc.i; love.me; i.miss; you.f**k; wa.f**k; fun.anywai; hear.you; friend.were; love.me; a.club;	wai.so; my.regular; increas.my; my.flower; didn't.need; coupl.year; each.year; bond.slowli; favourit.charact; most.social; other.job;
O	is.beauti; like.s**t; be.held; think.he's; unabl.to; and.fun; danc.and; pick.me; i.lost; the.hell;	to.church; prai.for; at.church; laid.back; mondai.and; not.bad; you.belong; not.exactli; over.time;
A	even.better; of.beauti; compromis.with; hold.you; the.colleg; keep.myself; me.sigh; no.point; from.peopl;	like.it's; comment.about; like.it'; excus.to; later.if; suppos.to; wa.worri; my.offic; sai.thing; goal.is; remain.in; return.of; send.the; unfortun.the; self.interest;
C	and.reliabl; prior.to; succe.in; so.hopefulli; got.caught; the.obviou; do.after; made.for; our.own; of.tear; on.track; to.drag; i.studi; hope.i'm; forget.that; realli.look;	episod.of; be.treat; not.thi; thi.just; pattern.is; real.reason; am.also; i.laugh; how.i'm; dare.to; of.why;

of the verb *excite*). Table 5 presents the subset of bigrams, from those that best classify our data, that are easily translated into the words that likely generated them.

Neuroticism's high and low scorers seem to use some problem talk (*only problem, depressed you, be sad*). The use of these kinds of words had been documented for low extravert scorers only [19]. On the other hand, low scorers of neuroticism use thoughtful words (*reflect on, choose to*). High extraverts use strong curse words (*you f**k, b**ch I, was f**k*), talk (possibly figuratively) about location (*i'm at*), and, and as described in previous literature [16, 19], they use social words and words suggesting positive emotional valence (*dance i, a club, fun anyway, most social*) and more self references than low scorers. In terms of self references, high extraverts use the first person singular more often, whereas low scorers use the possessive *my* more often. Low extraverts seem to use more time related language (*couple years, each year, bond slowly*).

Low openness scorers seem to use words for religious institutions and activities in their blogs (*to church, pray for, at church*). High scorers use weaker cursing than extraverts (*like s**t, the hell*). Also, as previously described [19], our data shows that high agreeableness scorers display more positive words (*even better, of beauty*). Lastly, high conscientiousness scorers seem to use language that denotes planning, outcome and evaluation (*to study, on track, prior to, succeed in*). Low scorers, in contrast, seem to use justification language more often (*real reason, of why*).

Because there were so few bigrams in each personality trait, the classification may overfit the data. However, we consider that given the size of the corpus, these features

provide a reliable insight into the kinds of lexical choices that people with different personality traits make when writing.

Because personality classification is a multi-class classification problem (i.e. personality traits are not mutually exclusive), there are methods that are better suited for this task such as conditional random fields –which consider the influences of the various classes over each other. We also note that future work is likely to harness greater linguistic information from natural language processing tools such as shallow parsers.

In this paper, we did not attempt to compare classification algorithms, but use the best performing one so far and explore the types of features that will lead to better classification and to provide a theoretical insight into the personality of bloggers.

5 Conclusion

In this paper we have presented a systematic examination of feature sets, both data-driven as well as categories derived from psychological dictionaries, to classify personality. Our best results ranged from 70.51% for neuroticism to 84.36% for openness to experience. Choosing bigrams as features yielded the best results. Our LIWC-based results are similar to those of previous studies that used LIWC for this task, but our best results with bigrams significantly improve upon them.

The superior performance of bigrams over word categories suggests that, at least to some degree, language structure is important when classifying personality traits. Similarly, Functional stopwords are also important in this context. Moreover, the presence or absence of features resulted in more accurate classification than frequency related scores for these features.

Based on preliminary analysis, we suggest that the thematic categories of words often used to analyse personality data may be too broad and future work might choose to refine them for this particular task. Future work will also consider classification using different corpora, different features such as topic distributions, and different kind of classifiers such as conditional random fields, in order to gain a better theoretical framework of personality in bloggers.

References

1. Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James 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. Paul T. Costa and Robert R. McCrae. Neo PI-R Professional Manual. *Odessa, FL: Psychological Assessment Resources*, 1992.
3. Michael Eid and Ed Diener. Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, 76(4):662–676, 1999.
4. Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson. Author profiling for english emails. In *10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 262–272, 2007.
5. Rong E. Fan, Kai W. Chang, Cho J. Hsieh, Xiang R. Wang, and Chih J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

6. Alastair J. Gill, Scott Nowson, and Jon Oberlander. What are they blogging about? personality, topic and motivation in blogs. In *ICWSM 2009*, 2009.
7. Martin Gütlein. Large scale attribute selection using wrappers. Master's thesis, Masters thesis, Albert-Ludwigs-Universitat, Freiburg, 2006, 2006.
8. Mark Andrew Hall and Lloyd Smith. Practical feature subset selection for machine learning. In *Proc 21st Australian Computer Science Conference*, pages 181–191, Perth, Australia, 1998. Springer.
9. S. Herring, L. Scheidt, S. Bonus, and E. Wright. Weblogs as a bridging genre. *Information, Technology & People*, 18(2):142–171, 2005.
10. Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. Text analysis as a tool for analyzing conversation in online support groups. In *Extended Abstracts of the 2004 conference on Human Factors and Computing Systems*, pages 1485–1488, 2004.
11. Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, 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.
12. Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862–877, May 2006.
13. Scott Nowson. *The Language of Weblogs: A study of genre and individual differences.* PhD thesis, University of Edinburgh, 2006.
14. Scott Nowson and Jon Oberlander. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the International Conference on Weblogs and Social*, 2007.
15. Scott Nowson, Jon Oberlander, and Alastair J. Gill. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671, 2005.
16. Jon Oberlander and Alastair J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42(3):239–270, 2006.
17. Jon Oberlander and Scott Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of COLING/ACL-06: 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, 2006.
18. James W. Pennebaker and Martha E. Francis. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, 1 edition, August 1999.
19. 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, December 1999.
20. John C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.
21. Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
22. Byron Reeves and Clifford Nass. *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA, 1996.
23. Nicola S. Schutte and John M. Malouff. University student reading preferences in relation to the big five personality dimensions. *Reading Psychology an international quarterly*, 25(4):273–295, October 2004.
24. Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June 2005.
25. T. Yarkoni. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44:363–373, 2010.