Exploring Stylistic Variation with Age and Income on Twitter

Lucie Flekova[*]
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Lyle Ungar and **Daniel Preoţiuc-Pietro**
Computer & Information Science
University of Pennsylvania
ungar@cis.upenn.edu
danielpr@sas.upenn.edu

Abstract

Writing style allows NLP tools to adjust to the traits of an author. In this paper, we explore the relation between stylistic and syntactic features and authors' age and income. We confirm our hypothesis that for numerous feature types writing style is predictive of income even beyond age. We analyze the predictive power of writing style features in a regression task on two data sets of around 5,000 Twitter users each. Additionally, we use our validated features to study daily variations in writing style of users from distinct income groups. Temporal stylistic patterns not only provide novel psychological insight into user behavior, but are useful for future research and applications in social media.

1 Introduction

The widespread use of social media enables researchers to examine human behavior at a scale hardly imaginable before. Research in text profiling has recently shown that a diverse set of user traits is predictable from language use. Examples range from demographics such as age (Rao et al., 2010), gender (Burger et al., 2011; Bamman et al., 2014), popularity (Lampos et al., 2014), occupation (Preoţiuc-Pietro et al., 2015a) and location (Eisenstein et al., 2010) to psychological traits such as personality (Schwartz et al., 2013) or mental illness (De Choudhury et al., 2013) and their interplay (Preotiuc-Pietro et al., 2015). To a large extent, the prominent differences captured by text are topical: adolescents post more about school, females about relationships (Sap et al., 2014) and sport fans about their local team (Cheng et al., 2010). Writing style and readability offer a different insight into who the authors are. This can help applications such as cross-lingual adaptations without direct translation, for text simplification closely matching the reader's age, level of education and income or tailored to the specific moment the document is presented. Recently, Hovy and Søgaard (2015) have shown that the age of the authors should be taken into account when building and using part-of-speech taggers. Likewise, socioeconomic factors have been found to influence language use (Labov, 2006). Understanding these biases and their underlying factors in detail is important to develop NLP tools without socio-demographic bias.

Writing style measures have initially been created to be applied at the document level, where they are often used to assess the quality of a document (Louis and Nenkova, 2013) or a summarization (Louis and Nenkova, 2014) , or even to predict the success of a novel (Ashok et al., 2013). In contrast to these document-level studies, we adopt a user-centric approach to measuring stylistic differences. We examine writing style of users on Twitter in relation to their age and income. Both attributes should be closely related to writing style: users of older age write on average more standard-conform (up to a certain point), and higher income is an indicator of education and conscientiousness (Judge et al., 1999), which determines writing style. Indeed, many features that aim to measure the complexity of the language use have been developed in order to study human cognitive abilities, e.g., cognitive decline (Boyé et al., 2014; Le et al., 2011).

The relationship between age and language has been extensively studied by psychologists, and more recently by computational linguists in various corpora, including social media. Pennebaker et al. (2003) connect language use with style and personality, while Schler et al. (2006) automatically

[*] Project carried out during a research stay at the University of Pennsylvania

classified blogs text into three classes based on self-reported age using part-of-speech features. Johannsen et al. (2015) uncover some consistent age patterns in part-of-speech usage across languages, while Rosenthal and McKeown (2011) studies the use of Internet specific phenomena such as slang, acronyms and capitalisation patterns. Preoţiuc-Pietro et al. (2016) study differences in paraphrase choice between older and younger Twitter users as a measure of style. Nguyen et al. (2013) analyzed the relationship between language use and age, modelled as a continuous variable. They found similar language usage trends for both genders, with increasing word and tweet length with age, and an increasing tendency to write more grammatically correct, standardized text. Such findings encourage further research in the area of measuring readability, which not only facilitates adjusting the text to the reader (Danescu-Niculescu-Mizil et al., 2011), but can also play an important role in identifying authorial style (Pitler and Nenkova, 2008). Davenport and DeLine (2014) report negative correlation between tweet readability (i.e., simplicity) and the percentage of people with college degree in the area. Eisenstein et al. (2011) employ language use as a socio-demographic predictor.

In this paper we analyze two data sets of millions of tweets produced by thousands of users annotated with their age and income. We define a set of features ranging from readability and style to syntactic features. We use both linear and non-linear machine learning regression methods to predict and analyze user income and age. We show that writing style measures give large correlations with both age and income, and that writing style is predictive of income even beyond age. Finally, Twitter data allows the unique possibility to study the variation in writing with time. We explore the effects of time of day in user behavior dependent in part on the socio-demographic group.

2 Data

We study two large data sets of tweets. Each data set consists of users and their historical record of tweet content, profile information and trait level features extracted with high precision from their profile information. All data was tokenized using the Trendminer pipeline (Preoţiuc-Pietro et al., 2012), @-mentions and URL's collapsed, automatically filtered for English using the *langid.py* tool (Lui and Baldwin, 2012) and part-of-speech tagged using the ArkTweet POS tagger (Gimpel et al., 2011).

Income ($\mathcal{D}_1$) First, we use a large data set consisting of 5,191 Twitter users mapped to their income through their occupational class. This data set, introduced in (Preoţiuc-Pietro et al., 2015a; Preoţiuc-Pietro et al., 2015b), relies on a standardised job classification taxonomy (the UK Standard Occupational Classification) to extract job-related keywords, search user profile fields for users having those jobs and map them to their mean UK income, independently of user location. The final data set consists of 10,796,836 tweets.

Age ($\mathcal{D}_2$) The age data set consists of 4,279 users mapped to their age from (Volkova and Bachrach, 2015). The final data set consists of 574,095 tweets.

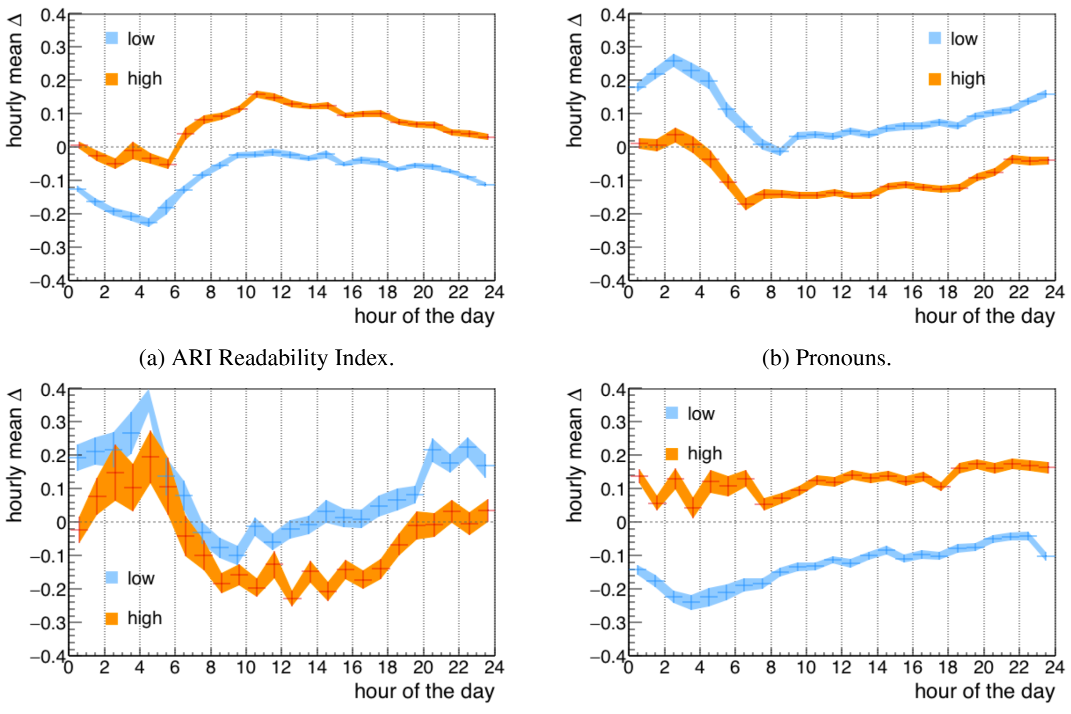
3 Features

We use a variety of features to capture the language behavior of a user. We group these features into:

Surface We measure the length of tweets in words and characters, and the length of words. As shorter words are considered more readable (Gunning, 1969; Pitler and Nenkova, 2008), we also measure the ratio of words longer than five letters. We further calculate the type-token ratio per user, which indicates the lexical density of text and is considered to be a readability predictor (Oakland and Lane, 2004). Additionally we capture the number of positive and negative smileys in the tweet and the number of URLs.

Readability After filtering tweets to contain only words, we use the most prominent readability measures per user: the Automatic Readability Index (Senter and Smith, 1967), the Flesch-Kincaid Grade Level (Kincaid et al., 1975), the Coleman-Liau Index (Coleman and Liau, 1975), the Flesch Reading Ease (Flesch, 1948), the LIX Index (Anderson, 1983), the SMOG grade (McLaughlin, 1969) and the Gunning-Fog Index (Gunning, 1969). The majority of those are computed using the average word and sentence lengths and number of syllables per sentence, combined with weights.

Syntax Researchers argue about longer sentences not necessarily being more complex in terms of syntax (Feng et al., 2009; Pitler and Nenkova, 2008). However, advanced sentence parsing on Twitter remains a challenging task. We thus limit ourselves in this study to the part-of-speech (POS)



(a) ARI Readability Index.

(b) Pronouns.

(c) Interjections.

(d) Named Entities.

Figure 1: Temporal patterns for groups of lowest (blue) and highest (orange) income users in our data set. X-axis shows the course of 24 hours in normalized time of day. Y-axis shows a normalized difference of the hourly means from the overall mean feature value. Width of a line shows the standard error.

information. In previous work on writing style (Pennebaker et al., 2003; Argamon et al., 2009; Rangel et al., 2014), a text with more nouns and articles as opposed to pronouns and adverbs is considered more formal. We thus measure the ratio of each POS using the universal tagset (Petrov et al., 2012).

Style We implemented a contextuality measure, based on the work of Heylighen and Dewaele (2002), which assesses explicitness of the text based on the POS used and serves as a proxy for formality. Using Stanford Named Entity Recognizer (Finkel et al., 2005), we measure the proportion of named entities (3-classed) to words, as their presence potentially decreases readability (Beinborn et al., 2012), and netspeak aspects such as the proportion of elongations (*wooow*) and words with numbers (*good n8*). We quantify the number of hedges (Hyland, 2005) and abstract words[1] used, and the ratio of standalone numbers stated per user as these are indicators of specificity (Pennebaker et al., 2003; Pitler and Nenkova, 2008). We also capture the ratio of hapax legomena, and of superlatives and plurals using Stanford POS Tagger (Toutanova et al., 2003) using the Twitter model.

4 Temporal Patterns in Style

Social media data offers the opportunity to interpret the features in a richer context, including time or space. In our income data set, a timestamp is available for each message. Golder and Macy (2011) showed user-level diurnal and seasonal patterns of mood across the world using Twitter data, suggesting that individuals awaken in a good mood that deteriorates as the day progresses. In this work we explore user-level daily temporal trends in style for the 1500 highest- and 1500 lowest-income users (mean income $\geq$ £35,000 vs mean income $\leq$ £25,000). In Figure 1 we present normalized temporal patterns for a selected set of features.

While the difference between groups is most striking, we also observe some consistent daily patterns. These display an increase in readability (Figure 1a) starting in the early hours of the morning, peaking at 10AM and then decreasing constantly throughout the day, which is in accordance with the mood swings reported by Golder and Macy (2011). The proportion of pronouns (Figure 1b) and interjections (Figure 1c) follows the

[1] www.englishbanana.com

exact opposite pattern, with a peak in frequency during nights. This suggests that the language gets more contextual (Heylighen and Dewaele, 2002) towards the end of the day. Finally, named entities (Figure 1d) display a very distinctive pattern, with a constant increase starting mornings, which increases throughout the day. While the first three patterns mirror the active parts of the day, coinciding with regular working hours, the latter pattern is possibly associated with mentions of venues or news. An increase in usage of named entities in the evening is steeper for low-income users - we hypothesize that this phenomenon could be reasoned by a stronger association of named entities with leisure in this user group. Overall, we notice a similarity between income groups, which, despite strongly separated, follow similar – perhaps universal – patterns.

Features	**Income** ($\mathcal{D}_1$)		**Age** ($\mathcal{D}_2$)		**Income-Age** ($\mathcal{D}_1$)	
Readability	Lin	RSVM	Lin	RSVM	Lin	RSVM
ARI	.282	.311	.269	**.318**	.230	.263
Flesch-Kincaid	.285	.319	.263	.310	.234	.284
Coleman-Liau	.230	.197	.203	.265	.202	.289
Flesch RE	.277	**.345**	.186	.295	.239	**.318**
FOG	.291	.309	.222	.270	.238	.267
SMOG	.288	.339	.240	.263	.234	.301
LIX	.208	.286	.215	.268	.177	.245
ALL	**.301**	**.380**	**.278**	**.329**	**.249**	**.354**
Syntax	Lin	RSVM	Lin	RSVM	Lin	RSVM
Nouns	.155	.200	.278	**.302**	.078	**.150**
Verbs	.044	.071	(.046)	.104	.093	.114
Pronouns	.264	**.297**	.148	.180	.114	.127
Adverbs	.115	.110	.077	.111	.135	.131
Adjectives	(.030)	.149	.162	.200	(.046)	.139
Determiners	(.040)	.070	.135	.154	.103	.121
Interjections	.123	.188	.084	.122	.059	.139
ALL	**.323**	**.258**	**.319**	**.229**	**.299**	**.267**
Style	Lin	RSVM	Lin	RSVM	Lin	RSVM
Named entities	.241	**.288**	.282	.293	.255	**.281**
Contextuality	(.044)	.204	.287	**.310**	(.030)	.134
Abstract words	.108	.120	.141	.183	.125	.139
Hedging	(.019)	.079	(.015)	.000	.(000)	.083
Specific (num)	.093	.011	.072	.176	.059	.124
Elongations	.097	.160	.072	.073	.056	.114
Hapax legom.	.056	.066	.160	.219	.064	.067
ALL	**.279**	**.347**	**.306**	**.134**	**.296**	**.312**
Surface	Lin	RSVM	Lin	RSVM	Lin	RSVM
# char. / token	.085	.144	.104	.148	.051	.101
# tokens / tweet	.158	.159	.228	.237	.115	.116
# char. / tweet	.214	**.261**	.262	**.278**	.153	**.169**
# words >5 char.	.139	.191	(.009)	.087	.112	.163
Type/token ratio	.099	.132	.090	.180	.100	.126
Punctuation	.218	.123	.093	.086	.057	.084
Smileys	.064	.113	.146	.144	(.030)	.090
URLs	.084	.128	.187	.194	(.040)	.077
ALL	**.379**	**.330**	**.294**	**.307**	**.352**	**.126**

Table 1: Predictive performance (Pearson correlation) for Income, Age and Income controlled for predicted age using linear (Lin) and non-linear (RSVM) learning methods. The last line of each sub-table shows the results for all features from that block together, while individual rows display individual performance for the predictive features. Numbers in bold represent the highest correlations from the specific block of features and data set. All correlations are significant on $p < 0.001$ level except for those in brackets.

5 Analysis

We view age and income as continuous variables and model them in a regression setup. This is in contrast to most previous studies on age as a categorical variable (Rangel et al., 2014) to allow for finer grained predictions useful for downstream applications which use exact values of user traits, as opposed to being limited to broad classes such as young vs. old. We apply linear regression with Elastic Net regularization (Zou and Hastie, 2005) and support vector regression with an RBF kernel (as a non-linear counterpart) for comparison (Vapnik, 1998). We report Pearson correlation results on 10-fold cross-validation. We also study if our features are predictive of income above age, by controlling for age assigned by a state-of-the-art model trained on social media data (Sap et al., 2014). Similar results have been obtained with log-scaling the income variable. Table 1 presents our prediction results. The strength of the correlation to the income and age, together with the sign of the correlation coefficient, are visually displayed in Figure 2.

As expected, all features correlate with age and income in the same direction. However, some features and groups are more predictive of one or the other (depicted above or below the principal diagonal in Figure 2). Most individual surface features correlate with age stronger than with income, with the exception of punctuation and, especially, words longer than 5 characters. The correlation of each readability measure is remarkably stronger with high income than with age, despite the fact these are to a large extent based on the surface features. Notably, Flesch Reading Ease – previously reported to correlate with education levels at a community level (Davenport and DeLine, 2014) and with the usage of pronouns (Štajner et al., 2012) – is highly indicative for income. On the syntactic level we observe that increased use of nouns, determiners and adjectives is correlated higher with age as opposed to income, while a high ratio of pronouns and interjections is a good predictor of lower income but, only to a lesser extent, younger age, with which it is traditionally associated (Schler et al., 2006). From the stylistic features, the contextuality measure stands out as being correlated with increase in age, in line with Heylighen and De-

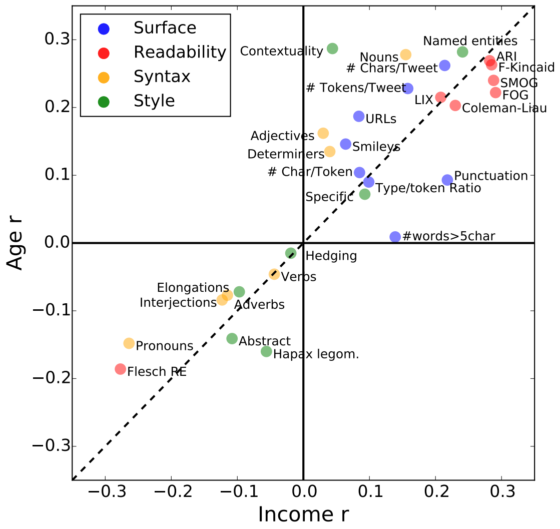


Figure 2: Predictive performance (Pearson correlation) for Income and Age. Individual points display univariate correlations (including sign) of the most predictive features.

waele (2002), but is almost orthogonal to income. Similarly, the frequency of named entities is correlated with higher income, while elongations have stronger association with younger age. Our results show, that based on the desired application, one can exploit these differences to tailor the style of a document without altering the topic to suit either age or income individually.

6 Conclusions and Future Work

Using two large data sets from thousands of users, annotated with their age and income, we presented the first study which analyzes these variables jointly, in relation to writing style. We have shown that the stylistic measures not only obtain significant correlations with both age and income, but are predictive of income beyond age. Moreover, we explored temporal patterns in user behavior on Twitter, discovering intriguing trends in writing style. While the discovery of these patterns provides useful psychosocial insight, it additionally hints to future research and applications that piggyback on author profiling in social media e.g., taking the message timestamp into account for stylistic features may yield improved results in user sociodemographic predictions. Likewise, utilizing additional proxies to control for income and education may lead to improvements in user age prediction.

Acknowledgments

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048. We also wish to thank Prof. Iryna Gurevych for supporting the collaboration.

References

Jonathan Anderson. 1983. LIX and RIX: Variations on a Little-Known Readability Index. *Journal of Reading*, pages 490–496.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, 52(2).

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2012. Towards Fine-Grained Readability Measures for Self-Directed Language Learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL.*

Maité Boyé, Thi Mai Tran, and Natalia Grabar. 2014. Nlp-oriented contrastive study of linguistic productions of alzheimer and control people. In LNCS 8686 Springer, Advances in Natural Language Processing, editor, *POLTAL*, pages 412–424.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM.

Meri Coleman and TL Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2).

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my Words!: Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web*, WWW.

James RA Davenport and Robert DeLine. 2014. The Readability of Tweets and their Geographic Correlation with Education. *arXiv preprint arXiv:1401.6058*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Jacob Eisenstein, Noah A Smith, and Eric P Xing. 2011. Discovering Sociolinguistic Associations with Structured Sparsity. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, NAACL.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-Local Information into Information extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL.

Rudolf Flesch. 1948. A New Readability Yardstick. *The Journal of Applied Psychology*, 32(3).

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics*, ACL.

Scott A. Golder and Michael W. Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333.

Robert Gunning. 1969. The Fog index after Twenty Years. *Journal of Business Communication*, 6(2).

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7(3).

Dirk Hovy and Anders Søgaard. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL.

Ken Hyland. 2005. Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies*, 7(2):173–192.

Anders Johannsen, Dirk Hovy, and Anders Sogaard. 2015. Cross-lingual syntactic variation over age and gender. In *CONNL*.

Timothy A. Judge, Chad A. Higgins, Carl J. Thoresen, and Murray R. Barrick. 1999. The big five personality traits, general mental ability, and carreer success across the life span. *Personnel Psychology*, 52.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

William Labov. 2006. *The Social Stratification of English in New York City*. Cambridge University Press.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of Three British Novelists. *Literary and Linguistic Computing*, 26(4).

Annie Louis and Ani Nenkova. 2013. What makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics*.

Annie Louis and Ani Nenkova. 2014. Verbose, Laconic or Just Right: A Simple Computational Model of Content Appropriateness under Length Constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-Shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL.

G Harry McLaughlin. 1969. SMOG Grading: A New Readability Formula. *Journal of Reading*, 12(8).

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. 'How Old do you Think I am?'; A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM.

Thomas Oakland and Holly B Lane. 2004. Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. *International Journal of Testing*, 4(3):239–252.

J.W. Pennebaker, Matthias R. Mehl, and K.G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1).

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Daniel Preoţiuc-Pietro, Sina Samangooei, Trevor Cohn, Nick Gibbins, and Mahesan Niranjan. 2012. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *Workshop on Real-Time Analysis and Mining of Social Streams*, ICWSM, pages 38–42.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PLoS ONE*.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering User Attribute Stylistic Differences via Paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI.

Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle H Ungar. 2015. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd Author Profiling Task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, CLEF.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC.

Sara Rosenthal and Kathleen McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre-and Post-Social Media Generations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*.

R.J. Senter and E.A. Smith. 1967. *Automated Readability Index*. Aerospace Medical Research Laboratories.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *NLP for Improving Textual Accessibility workshop*, pages 14–22.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL.

Vladimir N Vapnik. 1998. *Statistical learning theory*. Wiley.

Svitlana Volkova and Yoram Bachrach. 2015. On predicting socio-demographic traits and emotions in social networks and implications to online self-disclosure. *Cyberpsychology, behavior and social networking*, 18(12):726–736.

Hui Zou and Trevor Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67.