Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media

Notebook for PAN at CLEF 2013

Lucie Flekova[†‡] and Iryna Gurevych[†‡]

† Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt
‡ Ubiquitous Knowledge Processing Lab
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de

Abstract Would you target your audience differently, knowing the real age and gender of the text authors on your website forum? This paper examines hundreds of thousands of online documents, e.g. chat lines or blog posts, showing that computers are capable to address this task better than humans, without relying on content stereotypes. Pointing out that age and gender profiling are not independent problems, we approach the task as a multiclass classification problem, combining the age and gender information to define six classes. Utilizing a wide range of stylistic and content features and a large number of readability measures we demonstrate the high predictive abilities of the parts of speech, the punctuation and the amount of emotions and slang used in the text, independently of the topic discussed.

1 Introduction

The author profiling task aims at revealing certain categorical information about the author, rather than reveal his/her exact identity. Such categories can be his/her age, his/her gender, but also the native country, degree of education or other socio-demographic information. Beside its obvious applications in marketing, author profiling can be beneficial also in the educational domain, e.g. in large scale screenings of pupils, where it can help to reveal the exceptional talents. It can also help to estimate the appropriate knowledge level of the audience in an educational forum.

The PAN challenge task targeted the prediction of age and gender of a document author. Training corpora were provided for the English and Spanish language. They consisted of XML documents containing blog posts or chat messages (HTML format) grouped into one document per author and labelled with his/her language, gender and age group. The final software had to be ran on an assigned virtual machine, having a single CPU with 4 GB of RAM.

This paper presents our classification approach and implemented features, after which we discuss our experimental results for age and gender separately and combined.

2 Related Work

Studying gender differences and comparing them to social stereotypes has been a popular task in many psychological studies of 20th century [10] [17] [20]. Traditional studies worked on small datasets, which often led to contradictory results (see e.g. [22] v. [25]). The majority of the studies agree, that there are two main feature groups to distinguish gender - stylistic and content-based. The first detailed gender study in a larger scale was performed by Newman, Pennebaker et al. [23] on 14,324 samples from 70 different studies (conversation, exams, fiction etc.). According to them, women are more likely to include pronouns, verbs, negations, references to home, family, friends and to various emotions. Men tend to use longer words, more articles, prepositions and numbers. Men also swear more often and discuss current concerns (e.g. money, leisure or sports). Schler, Koppel et al. [26] apply machine learning techniques to a corpus of 37,478 blogs from blogger.com. Using classes of content words from the LIWC Framework [24] extended by blog slang (words and abbreviations such as *LOL* or *OMG*) and style-related features such as part-of-speech (POS) and function words, they were able to obtain an accuracy of 80% for gender and 76% for age, based on the Multi-Class Real Winnow classification algorithm. They found differences in topics which men and women discuss, as well as the increasing number of prepositions and determiners with age, together with the decreasing number of pronouns and negations. They report that the usage of hyperlinks increases with age. In another publication [2] they reach 72% accuracy for gender and 67% for age on the same corpus, using only stylistic features - POS tags, function words and contracted words without apostrophe (*im, dont*...). Koppel et al. [16] also analyse gender differences based on 566 fiction and non-fiction documents from the British National Corpus [5], using POS n-grams and function words. They reach an accuracy of 77%, however training on fiction and testing on non-fiction does not beat the 50% random baseline. Heylighen and Dewaele [14] introduce a contextuality measure based on the proportion of formal parts of speech (nouns, verbs...) to informal ones (adverbs, interjections...). Corney et al. [7] introduce emotionally intense adjective and adverb endings, based on the assumption that women use more emotional words such as *fabulous* or *awfully*.

3 Corpus properties

Table 1 shows the distribution of documents in the corpus. Beside blogs and chats, snippets from authors who pretend to be minors have been included (e.g., documents composed of chat posts of sexual predators).

The online origin of the corpus brought new challenges into the task. First of all, the age and gender were given by the bloggers themselves. Hence e.g. a male teenage author surprises us by talking about the benefits of trade fairs: *"...presently there have been far more trade fairs performed and media people will be certainly accessible at these locations so this enhances your expo to be observed by the people..."*. The second problem with the online data is the plagiarism. For example, the only post of one of the authors is a text about the city of Bhopal, an article which can be found on over hundred

sites of various travel agencies[1]. The third problem relates to the spam in the data. In more than 20,000 English documents, words from the WordPress Codex spamlist[2] constitute at least 0.1% of all document words - meaning that if any of these document texts appeared in the comments under a WordPress blog with an active spamlist, it would be quarantined for manual spam moderation by the blog administrator.

4 Our Approach

We combine age and gender information to create six separate document classes and perform multiclass classification, using the one-against-all training approach. Certain stylistic features can be highly predictive both of gender and of age, which makes it necessary to determine both gender and age at the same time in the classification. For example smileys, as illustrated on Figures 1(a)-1(d). This correlation was previously observed also by Schler and Koppel [1].

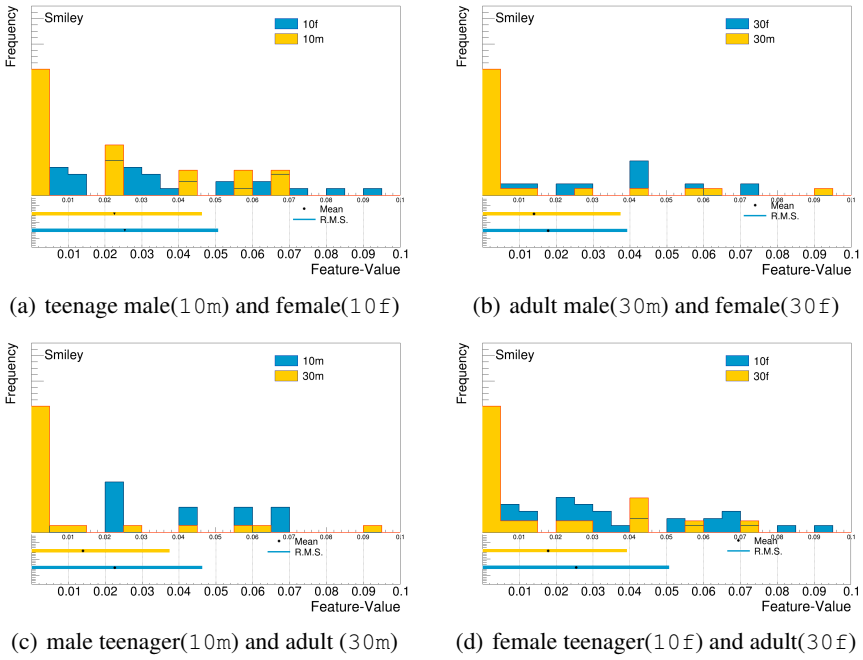
Our system builds upon the Darmstadt Knowledge Processing Software Repository (DKPro Core)[3] [12], an open-source Natural Language Processing framework based on Apache Unstructured Information Management Architecture (UIMA)[4]. The system uses the DKPro Lab [4] framework to combine NLP components into pipelines. We preprocess the data using the TreeTagger [27] for POS tagging and lemmatization for both languages. For English we additionally use the Stanford Named Entity Recognizer [8]. Having experimented with the SVM with the polynomial (1,2,3) and RBF kernel, and with the Updateable Naive Bayes classifier, we trained the final system using logistic regression with an unlimited number of iterations and with the ridge estimators [18] in their default configuration in the Weka [13] machine learning framework. While SVM performed the best on a small training set (6,000 documents), the computational complexity of the training was growing too fast. Since the system should deal with unknown test sets, we preferred to sacrifice some performance for scalability.

To select the training subsets, we first eliminate the documents whose text consists of more than 0.1% spam words. From the remaining data, we randomly select 5000 documents for each age and gender combination, thus obtaining 30,000 training doc-

[1] e.g. http://www.ganesha-holidays.com/madhya.html
[2] http://codex.wordpress.org/Spam_Words
[3] http://code.google.com/p/dkpro-core-asl/
[4] http://uima.apache.org/

Age group	Gender	No. of Authors in the English Corpus (180,809,187 words)	No. of Authors in the Spanish Corpus (21,824,198 words)
13-17 (10s)	Male	8,600	1,250
	Female	8,600	1,250
23-27 (20s)	Male	42,900	21,300
	Female	42,900	21,300
33-47 (30s)	Male	66,800	15,400
	Female	66,800	15,400

Table 1. Corpus characteristics



(a) teenage male(`10m`) and female(`10f`)

(b) adult male(`30m`) and female(`30f`)

(c) male teenager(`10m`) and adult (`30m`)

(d) female teenager(`10f`) and adult(`30f`)

Figure 1. Proportion of smileys for different categories

uments for English and 22,500 training documents for Spanish (the corpus contained 2,500 Spanish teenage authors only).

4.1 Experiments

We compare our results to the majority class baseline - accuracy 0.5 for gender and 0.33 for each of the three age classes (0.17 combined). Although the provided data contained 41%, resp.57%, of authors over 30 years, our training sets have more equally distributed instances, as finding the outliers in author profiles is often more useful in practice.

To measure an approximative human performance on the corpus, we conducted a user study. Twenty randomly selected English documents, containing about 500 words each, were evaluated by 15 participants. We measured the accuracy based on the majority vote. The confusion matrix is displayed in Table 2. Human participants reached an overall accuracy of 25% (50% on determining gender and 55% on determining age), while simple majority class baseline would result in 55% accuracy on gender and 50% on age. They assigned majority of the texts to authors in their 30's. This could possibly suggest that teenage authors may copy their blog content from other sources, or that they do not give correct age information about themselves.

We divide the features to five classes described below. We use the Information Gain feature selection approach [29] to rank and prune the feature space, using the top 1500 features.

Surface features To capture the surface properties of text, we measure the length of documents, sentences and words and their proportions to each other. We also count the ratio of words longer than five letters and words shorter than three letters compared to all words, and we count the occurrence of web links and smileys. Several further features are extracted using regular expressions, such as words with repetitive letters (e.g. *cooool*, *wooow*), words with numbers (e.g.*w8*, *ton8*) and number patterns such as phone numbers.

Syntactic features and punctuation Syntactic features constitute the majority of all features, as they proved helpful in previous work and at the same time are conveniently robust to be used across corpora and languages. We extract POS unigrams, bigrams, trigrams and quadrigrams as well as the ratio of each POS type separately. We implement the contextuality measure (Heylighen and Deweale, 2002), comparing impliciteness and expliciteness of the text based on POS tags used: *F = (noun frequency + adjective freq. + preposition freq. + article freq. - pronoun freq. - verb freq. - adverb freq. -0 interjection freq. + 100) * 0.5* .
We measure the proportion of singular and plural nouns, proper nouns and pronouns (both together and separately), as well as the ratio of personal pronouns for each grammatical form separately (*I, me* v. *he, him*...). From measuring the ratios of comparative and superlative adjectives and adverbs, and question mark and exclamation mark patterns, we expect clearer distinction of the teenage style. We retrieve the proportions of inner punctuations, end punctuations and commas, as labelled by the POS tagger. We further extract the proportion of modal verbs, which we granulate in English on modals expressing certainty (*shall, will*...) and uncertainty (*could, may*..). We also measure the ratio of future and past verb tenses. Some of the features have not been adapted to Spanish due to the different POS tagset used.

Readability measures We implemented the most prominent readability measures, such as the Flesch-Kincaid Grade Level [15], the Automatic Readability Index [28], the LIX Index [3], the Coleman-Liau Index [6] and the Flesch Reading Ease [9]. The majority of those is computed using the average word and sentence lengths and number of syllables per sentence, combined with manually determined weights. The SMOG grade [19] and the Gunning-Fog Index [11] also consider the number of complex words defined as words with three or more syllables. We did not adapt these readability measures to the Spanish corpus.

actual/pred.	Female 10s	Male 10s	Female 20s	Male 20s	Female 30s	Male 30s	Total
Female 10s	0	0	1	0	1	1	3
Male 10s	0	0	0	0	1	2	3
Female 20s	0	0	1	1	1	1	3
Male 20s	0	0	0	1	0	0	1
Female 30s	0	0	0	2	1	2	5
Male 30s	0	0	0	0	3	2	5

Table 2. Confusion matrix for the user study. The prediction is based on the majority vote.

Semantic features We experimented with retrieving the most frequent semantic triples, which are popular mainly in question answering tasks. A semantic sentence triple consists of a discourse entity, a semantic relation and a governing word to which the entity relates, e.g. *i-want-you*, *you-think-this*. We suspected men to refer more to actions (*you-should-X*) and women to feelings (*i-love-X*). However, the rank of these features decreased with the dataset growth, such as word n-grams did. We performed WordNet lookup using Java WordNet Library[5] to extract the number of senses for nouns and verbs in the text. Unfortunately, we had to exclude these semantic features from the final configuration in favour of processing time on the given machine.

Content features, lexical features and stopwords We use word unigrams and bigrams and stopword unigrams and bigrams based on the Snowball stopword list [6]. The Named Entity Features capture the number of named entities in the article, using the Stanford Named Entity Recognizer, in particular the 3-class model with distributional similarity features for tagging all entities of the types Person, Organization and Location. We use both the overall named entity counts and the average number of named entities per sentence as features. We also composed 23 word lists inspired by web resources[7][8] and previous work [23] - their full overview can be found in Table 3. As our main goal was to create a robust, dataset independent system, we focused mainly on lists expressing various emotions (anger, fear...) or language styles (teenage neologisms, web slang words, swear words...) rather than discussion topic areas.

List name	Size in words	Example	List name	Size in words	Example
Teenage words	117	bro, geez, tonite, lol	Certainty words	16	convinced, certain, clearly, always
Spam words	85	viagra, casino, shoes, -online	Politics words	309	voter, slogan, campaign
People words	134	relative, sister, team-mate	Clothing words	279	skirt, trousers, earrings
Emotion words	297	angry, calm, crazy, bored	Clarification words	17	pardon, repeat, example, clarify
Family words	166	family, grandpa, husband, wife	Uncertainty words	13	perhaps, maybe, unsure
Swear words	102	shit, fuck	Car words	207	engine, diesel, gearbox, chrome
Computer words	270	gigabyte, CPU, network	Work words	287	employee, bonus, recruiter, boss
Positive emotions	297	cheerful, amused, gracious, joyful	School words	69	homework, math, teacher
Positive feelings	89	delighted, proud, pleased	Sadness words	34	sorrowful, hopeless, broken, sad
Negative words	507	miserable, scared, stressed, angry	Anger words	52	mad, aggressive, outraged
Sensation words	141	sore, tight, cold, sharp	Fear words	45	nervous, worried, panicked

Table 3. Word lists

5 Evaluation

By the time of writing this paper, the challenge results are not yet finalized. We randomly split our selected training sets to 80% training and 20% test data for the evaluation. Performance of our systems was compared to the majority class baseline. Results are shown in Table 5.

[5] http://jwordnet.sourceforge.net

[6] http://anoncvs.postgresql.org/cvsweb.cgi/pgsql/src/backend/snowball/stopwords/

[7] http://www.enchantedlearning.com/

[8] http://eqi.org/fw_neg.htm

5.1 Gender profiling

When trained only for gender profiling, our system reaches an accuracy of 0.58 on English and 0.65 on Spanish dataset. As we observed much noise in the English corpus, we tested our system also on the English corpus from Mukherjee et al. [21] (3227 authors), on which we reach an accuracy of 0.65, comparable to previous experiments with similar classification setup [30].

Features that appeared in the 50 best performing ones in at least two datasets are listed in table 6. Men tend to use more articles, longer words and articles, in accordance with [23], and talk more about computers. Women are likely to use more emotional words, smileys and exclamations. They are also more likely to talk about love. Longer word ngrams have no impact in any of the datasets. In the English dataset we observed also higher usage of hyperlinks by men, as previously noted by [26], and highly ranked readability measures which are based on word length (ARI, LIX). However it is not the case for readability measures based on number of syllables (Flesch, SMOG, FOG). Hence the usage of hyperlinks and long words may only suggest, that long words could be names of specific websites and male blogs in our corpus are simply more likely to contain spam.

5.2 Age profiling

Training our system for age profiling only, we reach an accuracy of 0.53 on the English and 0.57 on the Spanish dataset.

Top ranked features shared by both datasets are listed in table 6. The older authors tend to write longer posts using longer words. They pay more attention to commas, although their sentences are not necessarily longer. Younger authors also use more pronouns and less nouns and articles - similar features distinguish male and female authors, as pointed out also by [1]. The highest ranked features in the Spanish dataset were the smileys, commas, and different writing of words using the letter *q*, instead of which teenagers use *k*, such as *ke, kiero*. Adults also talk more about work and god. On the English dataset, we observed a higher usage of hyperlinks by older authors, lower readability and more frequent punctuation. Topic word lists played an important role - younger people use more emotional words, neologisms and slang, talk more about other people (classmates, parents...) and about computers.

The English dataset suffered from different errors than the Spanish one. While the major issue in the Spanish dataset was distinguishing teenage authors from 20's , in case of English dataset it was to distinguish teenagers from mature authors (30's). If we assume that all authors reported their correct age, this might be caused by the plagiarism and by the fact that the corpus contained also chat conversations of sexual predators from PAN 2012 [9], which we did not particularly address.

5.3 Final system

We reach an overall accuracy of 0.29 on the English dataset and 0.38 on the Spanish one, with the majority baseline being 0.17. English and Spanish confusion matrices

[9] http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html

Feature Class	Feature	InfoGain English	InfoGain Spanish	Feature Class	Feature	InfoGain English	InfoGain Spanish
Surface	Word length	.024	.012	Syntactic	Noun rate	.025	n/a
	Words >5 letters	.023	.021		Pronoun rate	.012	.002
	Words <3 letters	.009	.007		Adverb rate	.011	.003
	Document length	.020	.017		Preposition rate	.010	.003
	Number of sentences	.007	.018		Verb rate	.008	.006
	Number of tokens	.018	.015		Contextuality measure	.022	n/a
	Tokens per sentence	.012	.006		Plural ratio	.016	n/a
	Number of smileys	.025	.033		Pronoun singular	.014	n/a
	Number of web links	.046	.0		Pronoun I	.012	n/a
	Type-token ratio	.009	.008				
Readability	ARI	.031	n/a	Punctuation	End punctuation	.030	.010
	Flesch	.017	n/a		Inner punctuation	.025	.024
	Kincaid	.024	n/a		Punctuation rate	.010	.010
	SMOG	.021	n/a		Comma	.006	.024
	LIX	.029	n/a		Exclamation rate	.005	.017
	FOG	.022	n/a				
	Coleman-Liau	.027	n/a				
Content	Teenage words	.018	.017	Lexical	Ending -ly	.011	n/a
	Emotion words	.011	.007				
	Certainty words	.004	.010				
	Work words	.008	.012				

Table 4. Information Gain rankings for selected features in the final English and Spanish run (multiclass age+gender classification). N-gram-based features are omitted for space reasons. Features marked 'n/a' were not adapted for the Spanish system.

System	Gender EN	Age EN	Combined EN	Gender ES	Age ES	Combined ES
Major.class baseline	0.5	0.33	0.17	0.5	0.33	0.17
Human evaluation	0.5	0.55	0.25	-	-	-
Our system	0.58	0.53	0.29	0.65	0.57	0.38

Table 5. Classification accuracy on English and Spanish data

Dataset	PAN English	PAN Spanish	Mukherjee English	Dataset	PAN English	PAN Spanish
Smileys	.001	.012	-	No.of characters	.021	.013
Word length	.001	.005	.026	Words > 5 letters	.018	.018
Type-token ratio	.002	.005	-	No. of sentences	.010	.012
amor,love	0	.005	.018	Usage of comma	.010	.023
HTML links	.002	.004	-	Pronoun ratio	0.005	.010
Words > 5 letters	.001	.002	.022	Noun ratio	.009	.010
Exclamation ratio	0	.006	.023	Article ratio	0.005	.010
Sent.-length in char.	.001	.002	.017	Modal verbs	0.005	.011
Number of words	.003	.005	0	*love, kiero*	0.006	.008

Table 6. Gender (left) and Age (right) features selected by the Information Gain Ranking Filter in more than one data set

act./pred.	10s F	10s M	20s F	20s M	30s F	30s M	10s F	10s M	20s F	20s M	30s F	30s M
10s F	.43	.05	.20	.16	.08	.08	.26	.17	.13	.13	.18	.13
10s M	.06	.31	.15	.24	.10	.15	.16	.25	.10	.14	.21	.13
20s F	.06	.04	.36	.18	. 20	.15	.14	.13	.26	.21	. 16	.09
20s M	.05	.05	.20	.33	.14	.23	.12	.09	.13	.43	.12	.11
30s F	.03	.03	.18	.09	.39	.27	.17	.15	.13	.11	.30	.13
30s M	.02	.03	.10	.15	.24	.46	.16	.14	.11	.19	.16	.25

Table 7. Confusion matrices for six classes on the Spanish (left) and English (right) test data using all features

System	English dataset	Spanish dataset
Maj.class equal distr. baseline	0.17	0.17
Human evaluation	0.25	-
Surface features	0.20	0.21
Syntactic & punct. features	0.23	0.30
Content & lex. features	0.27	0.33
Synt. & punct.& cont. & lex.	0.29	0.38
All features combined	0.29	0.38

Table 8. Performance of individual feature classes

for the final system are displayed in Table 7. For the English corpus, we achieve the best recall for 20's men (43%) and the lowest for teenage men (25%), who are often misclassified as 30's women (21%). In Spanish we obtain the best recall for 30's men (46%) and the worst also for teenage men (31%), but these are often misclassified as 20's men(24%). The feature ranking in the final system is listed in Table 4. The highest ranked feature is the number of hyperlinks for English and the number of smileys for Spanish, and in both cases the ratio of words longer than 5 letters and the punctuation features. Surface features are ranked surprisingly high as well, followed by readability measures. While in English the ratio of individual part-of-speech tags plays an important role, in Spanish POS trigrams and quadrigrams are preferred. From the vocabulary lists, teenage words, emotion words and work words (see Table 3 are the most dominant, followed by the expressions of positive feelings and uncertainty. From all the word n-grams, only the unigrams *love* and *ur* were selected for the English corpus and the unigrams distinguishing letters, such as *k, q, ke, que*, for the Spanish corpus.

We compare the performance of the classifiers trained on each feature class separately and all of them together. The results are shown in Table 8. Neither of the datasets is sufficiently separable by surface features alone, reaching the accuracy of 0.20, resp. 0.21 only.

Syntactic features performed well on the Spanish dataset. Errors occured mainly for teenagers being incorrectly classified as 20's men (18%), some 20's men classified as 30's women (21%), and some 30's women classified as 30's men (25%) and vice versa (20%). On the English dataset syntactic features show lower accuracy. Many women in their 20's were classified as 20's men (22%), some 30's men were misclassified as one decade younger (18%), and both genders of teenagers were in 20% of the cases incorrectly classified as adult women in their 30's, potentially due to plagiarism.

Content features alone were the best performing of all individual feature classes. They were suitable to distinguish age groups, but had problem with recognizing gender - on Spanish dataset 21% of 20's women classified as 20's men and vice versa, 29% of 30's women as 30's men and 23% of 30's men as 30's women. The English dataset suffered from similar errors as with syntactic features. Teenage slang and emotion words were the most helpful word lists, hence removing the topic bias (work words, family words etc. did not impact the performance.

6 Conclusions

To our knowledge, our system was the first to approach the age and gender problem as a single multiclass classification problem, which helped us to observe both tasks in context and confirm, that the age and gender profiling are not independent problems. We have shown that both of them can be determined by the same features (young men are more emotional than older ones, and so are women, which is visible through stylistic features). We were the first to employ readability measures in this task and we show, that these are ranked high in both age and gender classification. This is in accordance with the high ranks of words longer than five letters, which are used more by men and mature authors. While we observe, with regards to syntactic and content features, similar findings to previous work [23] [26] [2] [1] mainly on the Spanish corpus, syntactic features were not dominant on the English corpus, probably due to strong noise potentially caused by the presence of spammers, plagiators and sexual predators. When we run our system on a cleaner English corpus [21], we drew similar conclusions to state of the art literature. We have demonstrated that humans perform worse than computers in this task (close to random), as they cannot capture patterns in the data well and rely on content stereotypes. While content features perform overall better than syntactic features, the accuracy of the latter is satisfactory and can be easier adapted for a multilingual system, while e.g. translation of teenage slang is challenging without very good knowledge of the target language.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (Feb 2009)
3. Björnsson, C.: Läsbarhet: Lesbarkeit durch Lix. (Aus dem Schwedischen). (Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6.), Liber (1968)
4. Eckart de Castilho, R., Gurevych, I.: A lightweight framework for reproducible parameter sweeping in information retrieval. In: Proceedings of the 2011 workshop on Data infrastructurEs for supporting information retrieval evaluation. pp. 7–10. ACM (2011)
5. Clear, J.H.: The digital word. chap. The British national corpus, pp. 163–187. MIT Press, Cambridge, MA, USA (1993)
6. Coleman, M., Liau, T.: A computer readability formula designed for machine scoring. Journal of Applied Psychology 60(2), 283 (1975)
7. Corney, M., de Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: Computer Security Applications Conference, 2002. Proceedings. 18th Annual. pp. 282–289 (2002)
8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics (2005)
9. Flesch, R.: A new readability yardstick. The Journal of applied psychology 32(3), 221 (1948)
10. Gleser, G.C., Gottschalk, L.A., John, W.: The relationship of sex and intelligence to choice of words: A normative study of verbal behavior. Journal of Clinical Psychology 15(2), 182–191 (1959)

11. Gunning, R.: The fog index after twenty years. Journal of Business Communication 6(2), 3–13 (1969)
12. Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., Zesch, T.: Darmstadt knowledge processing repository based on uima. In: Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the GSCL (2007)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
14. Heylighen, F., Dewaele, J.M.: Variation in the contextuality of language: An empirical measure. Foundations of Science 7(3), 293–340 (2002)
15. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., DTIC Document (1975)
16. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
17. Lakoff, R.T.: Language and woman's place, vol. 56. Cambridge Univ Press (1975)
18. Le Cessie, S., Van Houwelingen, J.: Ridge estimators in logistic regression. Applied statistics pp. 191–201 (1992)
19. McLaughlin, G.H.: Smog grading: A new readability formula. Journal of reading 12(8), 639–646 (1969)
20. McMillan, J.R., Clifton, A.K., McGrath, D., Gale, W.S.: Women's language: Uncertainty or interpersonal sensitivity and emotionality? Sex Roles 3(6), 545–559 (1977)
21. Mukherjee, A., Liu, B.: Improving Gender Classification of Blog Authors. In: EMNLP'10. pp. 207–217 (2010)
22. Mulac, A., Studley, L., Blau, S.: The gender-linked language effect in primary and secondary students' impromptu essays. Sex Roles 23(9-10), 439–470 (1990)
23. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker: Gender Differences in Language Use: An Analysis of 14,000 Text Samples. Discourse Processes pp. 211–236 (2008)
24. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates (2001)
25. Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology 54(1), 547–577 (2003)
26. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. pp. 199–205 (2006)
27. Schmid, H.: Treetagger. TC project at the Institute for Computational Linguistics of the University of Stuttgart (1994)
28. Smith, E., Senter, R., (U.S.), A.F.A.M.R.L.: Automated Readability Index. AMRL-TR-66-220, Aerospace Medical Research Laboratories (1967)
29. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97). pp. 412–420. Morgan Kaufmann Publishers, Inc. (1997)
30. Zhang, C., Zhang, P.: Predicting gender from blog posts. Tech. rep., Technical Report. University of Massachusetts Amherst, USA (2010)