Can We Hide in the Web?

Age and Gender Author Profiling in Social Media

DIPF Bildungsforschung und Bildungsinformation

TECHNISCHE UNIVERSITÄT DARMSTADT

...most of this sugar comes from high fructose corn syrup which is the chief ingredient in chips, cereals or breads. And just because it is "all natural", it does not mean it's good for you. To the body, it's all sugar!

Was this text written by a...

Flekova, Lucie +^

Gurevych, Iryna +^

+ Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research and Educational Information

^ Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universität Darmstadt

System Setup – DKPro Lab Framework

PAN Challenge Data

- from web environment (chats, blogs, fora)
- 236 600 documents in English
- 75 900 documents in Spanish
- 3 age groups (13-17, 23-27, 33-47 years)
- male and female authors
- over 200 mil. words

Text Processing
DKPro Core

- XML parsing
- Tokenization
- POS Tagging - TreeTagger
- Lemmatizing - TreeTagger
- Chunking - Stanford Parser
- Named Entity Recognition - Stanford NER

Extracting Features
DKPro Text Classification

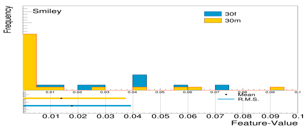
- **Surface:** Long/short words, words per sentence, number of hyperlinks, number of smileys, type-token ratio, text length...
- **Readability:** Flesch, Kincaid, Coleman-Liau, SMOG, FOG, LIX
- **Content:** Emotion words (e.g. anger), topic words (e.g. school)
- **Syntax:** POS ratios, Contextuality measure, plurals, modals
- **Punctuation:** Inner punctuation, questions, exclamations
- **Lexical:** Emotional endings (e.g. –ous, -ly...)

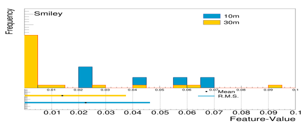
Classification
DKPro Text Classification & WEKA

- Multiclass classification for six classes (one-against-all approach)
- Logistic regression with ridge estimates (WEKA)
- Information Gain filter

Age and gender differences are related

Style becomes more "male" with age - we get more descriptive while showing less emotions. For relevant features, such as frequency of smileys, the difference between adult men and women (upper plot) is similar to the one between teenage and adult men (lower plot).





Content-based features outperform style

Features based on word lists (mainly teenage slang and emotions) contributed to the overall performance more than stylistic features. However, they were more successful in determining age than gender.

Word list	Words	Example
Teenage words	117	Bro, geez, tonite, lol
People words	134	Relative, team-mate, friend
Work words	287	Employee, bonus, recruiter
Positive words	297	Cheerful, amused, joyful
Negative words	507	Miserable, scared,stressed

Performance *(classifier accuracy)*

Subsystem	English	Spanish
Maj.class baseline	0.17	0.17
Human evaluation	0.25	-
Surface features only	0.20	0.21
Syntactic & punct.features	0.23	0.30
Content & lexical features	0.27	0.33
Syntax.& punct.& content & lex.	0.29	0.38
All features combined	**0.29**	**0.38**

System	EN gen.	EN age	EN all	ES gen	ES age	ES all
Baseline	0.5	0.33	0.17	0.5	0.33	0.17
Humans	0.5	0.55	0.25	-	-	-
Our system	**0.58**	**0.53**	**0.29**	**0.65**	**0.57**	**0.38**

User study

- 20 participants,
- 20 random texts from the PAN challenge
- Age accuracy 0.55, no teenagers identified
- Gender accuracy 0.5 = random decisions
- Human prediction based on stereotypes, fails on neutral topics

Conclusions - Gender classification *(a = .62)*

Men
use longer words, more articles and hyperlinks, and talk more often about computers.

Women
use more emotional words, smileys, exclamations and "love" words

Conclusions - Age classification *(a = .55)*

Older authors
write less readable longer posts, use longer words, commas, links, talk more about work and god.

Teenagers
use more pronouns and smileys, less nouns and articles, speak with more emotional words, neologisms and slang, talk more about people and computers and often violate the spelling rules.

References

Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval
Richard Eckart de Castilho and Iryna Gurevych, In: Maristella Agosti and Nicola Ferro and Costantino Thanos: Proceedings of the 2011 workshop on Data infrastructurEs for supporting information retrieval evaluation, vol. DESIRE '11, p. 7-10, ACM, October 2011. ISBN 978-1-4503-0952-3.

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, In: CLEF 2013 Labs and Workshops - Notebook Papers, p. (to appear), September 2013.

Lucie Flekova, flekova@dipf.de, Ubiquitous Knowledge Processing Lab, http://www.ukp.tu-darmstadt.de