COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER / CLiPS

# Computational Stylometry

*Walter Daelemans*
CLiPS, Linguistics Department
University of Antwerp
walter.daelemans@ua.ac.be

## Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies

## Who, what, where, when, …

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

+ links to ontologies, e.g. Wikipedia

## Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies
- Subjective
  - Sentiment, opinion, emotion
  - Modality, (un)certainty

# Subjectivity

- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

# Three levels of knowledge from text

- Objective (Machine Reading)
  - Events, concepts, attributes, relations
  - Space, time, causality, discourse
  - Linking to ontologies
- Subjective
  - Sentiment, opinion, emotion
  - Modality, (un)certainty
- Metaknowledge
  - Authorship, author attributes (educational level, age and gender, personality, region, illness), text attributes (date of writing, …)

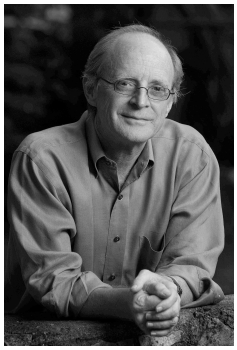
# Meta knowledge

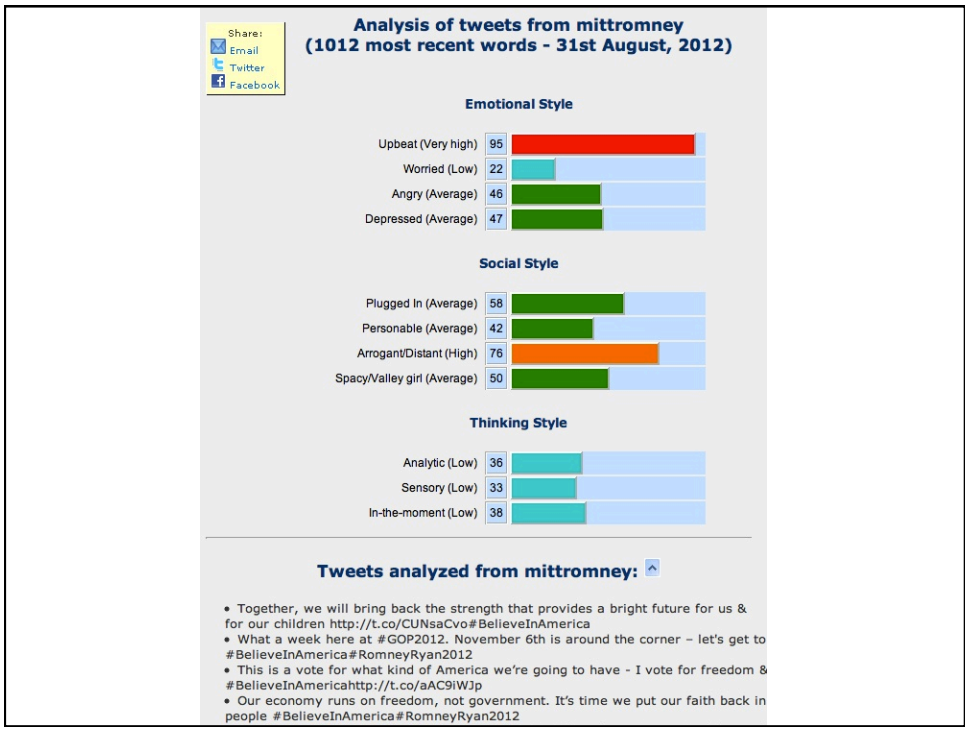
- The former Liechtenstein and later Diestrichstein chateau on the rock has been a unique dominant of the Mikulov skyline for centuries. The original governor's castle was donated by Přemysl Otakar II in 1249 to the Liechtenstein family as the fief. In late 16th century the new owners of the seat, the Dietrichstein family, had the chateau reconstructed to the present appearance after the fire in 1719. The chateau burned to the ground in 1945 while retreat of the German army but thanks to the care of The Association for recovery of the chateau Mikulov the difficult repair was done in the 1950's. Chateau library along with the Hall of Ancestors belong to the most interesting sections of the chateau.

Male, adult, non-native author?

# Personality from Twitter

- http://www.analyzewords.com
- James Pennebaker et al. (Univ. of Texas)
- Linguistic Inquiry and Word Count (LIWC)
  - "the secret life of pronouns" (2011)
  - Pronoun use reveals personality and mental states
    - When depressed, we use more pronouns
    - After 9/11 "we" used more than "I"
    - …






# Landmark: gender from text

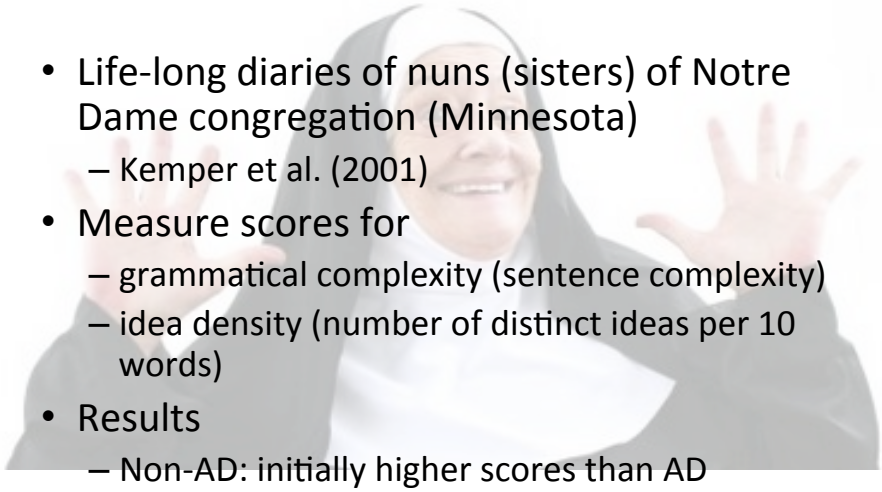
- Shlomo Argamon, Moshe Koppel et al. (from 2002)
- Documents: British National Corpus (fiction and non-fiction)
- Class: gender of author
- Feature construction:
  - lexical (Function Words)
  - POS (Function Words)
- Supervised learning: linear separator
- Results: gender ~ 80% predictable from text

# Gender Differences

- Use of pronouns (more by women) and some types of noun modification (more by men)
  - *"Male" words: a, the, that, these, one, two, more, some*
  - *"Female" words: I, you, she, her, their, myself, yourself, herself*
- More "relational" language use by women and more "informative" (descriptive) language use by men
- Even in formal language use!
- Strong correlation between male language use and non-fiction, and female language use and fiction

- LIWC categories (in Blogs):
  - Men talk more about jobs, money, sports, tv
  - Women talk more about sex, family, eating, friends, sleep, emotions

# The nun study



- Life-long diaries of nuns (sisters) of Notre Dame congregation (Minnesota)
  - Kemper et al. (2001)
- Measure scores for
  - grammatical complexity (sentence complexity)
  - idea density (number of distinct ideas per 10 words)
- Results
  - Non-AD: initially higher scores than AD
  - Non-AD declines at a faster rate

# Explanation

- "The findings could also mean that language abilities in the early 20s can predict the risk of developing dementia several decades later"
- Low linguistic ability in early life may reflect suboptimal neurological and cognitive development which might increase susceptibility to the development of Alzheimer's disease pathology in late life

# Contents of this talk

- Does a human stylome exist and if so, how can it be measured?
- How do we overcome current problems with the dominating approach (text categorization)?
  - Making it work in realistic contexts
    - Scalability
    - Verification instead of attribution
    - Cross-genre
    - On social network language
- Recent CLiPS case studies addressing these problems
  - Cross-genre authorship verification
  - Identifying pedophiles in social networks
  - Stylome-wide association studies

Mike Kestemont

Tom De Smedt

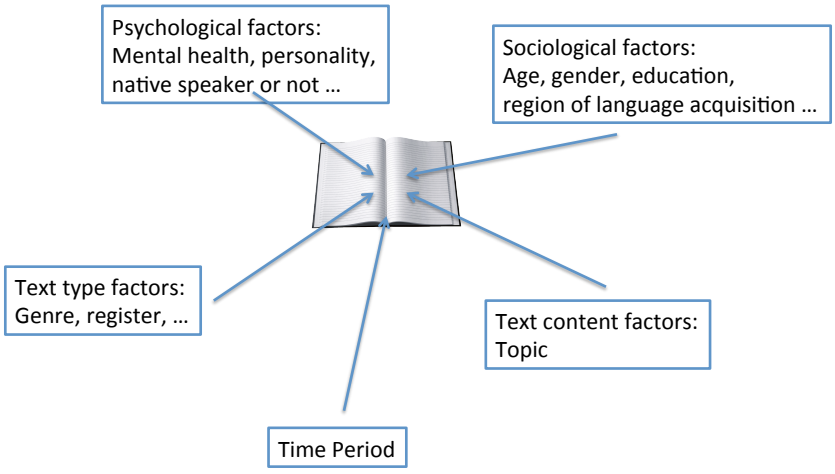
Guy De Pauw

Kim Luyckx

Claudia Peersman

Frederik Vaassen

# Sources of language variation



# Definition of Style

- Variation due to "style"
  - A *combination* of specific, *invariant* and *unconscious* decisions in language generation at all linguistic levels (discourse, syntactic structures, lexical choice, …) associated with specific authors or author traits
- Computational stylometry is
  - The attribution of texts to individual authors or author traits using computational models that recognize writing style
- Human Stylome hypothesis
  - If style is unique (like fingerprint or genome):

  "(…) authors can be distinguished by measuring specific properties of their writings, their stylome as it were" (Van Halteren et al., JQL, 2005)

# The method of choice: Text Categorization = Supervised ML

$$F : D \rightarrow C$$

$$Max_i P(c_i \mid d), \forall c_i \in C$$

- Document representation
  - Linguistic analysis
  - Feature selection
  - Dimensionality reduction
- Machine Learning
  - Supervised machine learning (cross-validation)
    - Svm, NB, MBL, …

# Stylometry features

- Letter frequency, punctuation, spelling errors …
- Character n-grams
- Word n-grams
- Distributions of function words, content words, frequent words, pronouns, …
- Morphology: prefixes and suffixes
- Syntax: POS tag (distributions), rewrite rule frequencies, chunk distributions
- Semantics: semantic subclasses (wordnet), case frame distributions, …
- Special subclasses, e.g. LICW (positive emotions, negative emotions, self-references, family, causality, …)
- Stable words (stay the same if translated and translated back, will probably survive editing)
- Complexity measures (readability, average word length, average sentence length, …)
- Vocabulary richness: type token ratio, hapax (dis)legomena ratio
- Discourse level features (connectors)
- …

# Applications

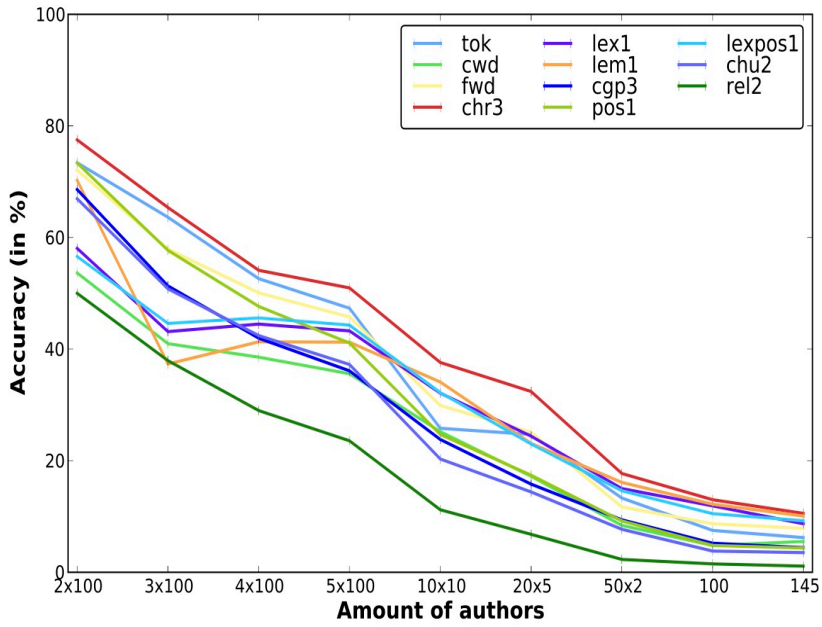
- Authorship attribution
- Profiling
  - Personality, illness, depression, gender, age, education level, region of language acquisition, non-native speakers, …
- In the service of: literary science, forensic sciences, sociolinguistics, language psychology, social psychology, medical diagnosis, …

# Problems with the text categorization approach

## Authorship Attribution

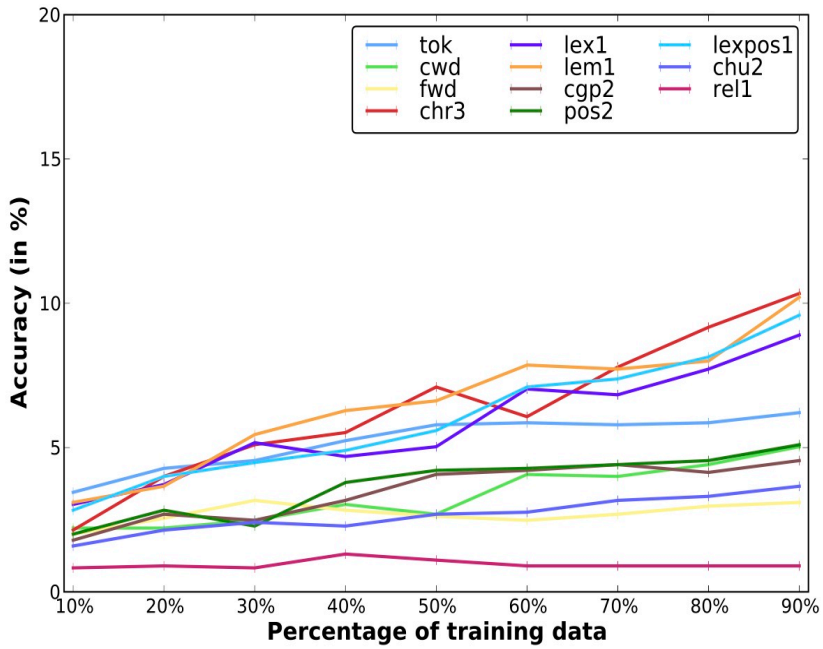
- Scalability issues
  - Many authors
  - Short texts
- Does the Text Categorization approach generalize to many authors and short texts?
  - Personae corpus

Kim Luyckx and Walter Daelemans, The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing*, 26(1), 2011.




# Observations

- Significant decrease of performance with more authors (80% to 20% for the best feature set)
- Character n-grams are most robust
- Goodness of features is robust (few crossing curves as number of authors increases)




# Observations

- Significant increase of performance with more data (3% to 10% for the best features)
- Character n-grams and lemmas are most robust
- Goodness of features is robust (few crossing curves as size of data increases)

## Why do char n-grams often work so well?

- Good trade-off between sparseness and information content
- Implicit punctuation, morphology, semantics, lexical items (function words are often short words), context
- Tolerant to errors (two spelling variants still share many character n-grams)

## Cross-Genre Authorship Verification

# Authorship Attribution versus Verification

- Attribution is easy
  - Given texts of author A, B, C, … and an unknown text X from one of these, decide by which of the authors it was written (*closed* case)
- Verification is difficult
  - Given an unknown text X and texts of candidate authors, decide which of the authors has written X *but noone is also possible* (*open* case)
- No negative information
  - "One-class learning"
  - Problematic in machine learning
    - How do we find a representative sample of the negative class?

# Authorship Verification

- Did *A* write *X*?
- Naïve method:
  - Divide A and X into chunks
  - Train a classifier to discriminate between X and A chunks
  - Evaluate generalization of the classifier
    - High accuracy then different authors
    - Low accuracy then same author
- Doesn't work
  - Small set of features can wrongfully maximize differences

Koppel & Schler, 'Authorship verification as a one-class classification problem.' ICML 2004, 489-495, 2004.

# Authorship Verification

- Did *A* write *X*?
- Unmasking
  - A = text by author of interest, B = text by other author(s) (impostors)
  - Divide A, B into equal-sized chunks
  - For each combination of two texts (e.g., A1, B2)
    - Generate a degradation curve
      - Train svm to discriminate between the two texts (cross-validation)
      - Iteratively remove k best features and re-train and test
  - Use a metalearner on the degradation curves to predict same / different author on test curves and run X through the system
  - Visually:
    - Slow and gradual deterioration: not same author
    - Sudden and dramatic deterioration: same author
      - If written by same author then number of differences will be relatively low
    - Measures the *depth* of the difference between the texts

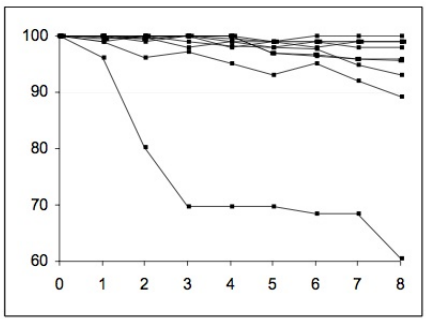



Figure 2. Unmasking *An Ideal Husband* against each of the ten authors (n=250, k=3). The curve below all the authors is that of Oscar Wilde, the actual author. (Several curves are indistinguishable.)

## Can we use unmasking for cross-genre?

- Cross-genre authorship attribution is difficult and hardly addressed in the literature
  - Train on genre A, test on genre B
  - *Can we assume that we can determine authorship of a suicide letter when all we have to train are scientific papers, or blogs, or student essays?*
- Combining two difficult tasks: cross-genre + authorship verification
- Hypothesis: the most discriminating features will be genre-specific

Kestemont, Luyckx, Daelemans & Crombez, 'Cross-Genre Authorship Verification Using Unmasking'. *English Studies*, 93 (3), 340-356, 2012.

## Data

- 5 contemporary authors
  - Edward Bond, David Mamet, Harold Pinter, Sam Shepard, Arnold Wesker
  - Theatre and prose (11 works)
  - Minimal text length 20 chunks of 500 words
- Sanity check: single genre
  - Features = most frequent words
  - Prose: Accuracy 96%, F-score 95%
  - Theatre: Accuracy 84%, F-score 62%
    - Explanation: unmasking is known to become less effective with shorter texts

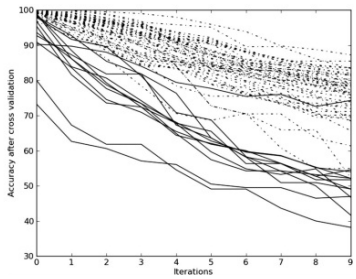



**Figure 1** Unmasking Applied to Prose Texts by Five A

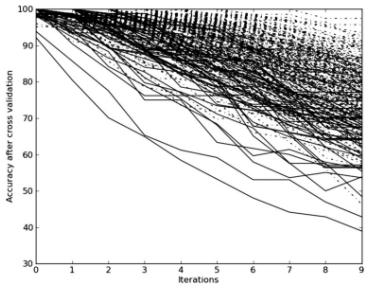
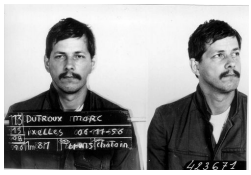



**Figure 2** Unmasking Applied to Theatre Plays by Five Authors.

# Cross-genre

- Restrict degradation curves to different genre texts
- Doesn't work very well
  - Accuracy 77%, F-score 56%
  - Explanation
    - Sensitive to many parameters that should be optimized (e.g. number of features to be removed)
  - Removed features make sense
    - Names of principal characters
    - Stage directions & colloquialisms in theatre
    - Descriptive and introspective words in prose

# Detecting pedophiles in social networks

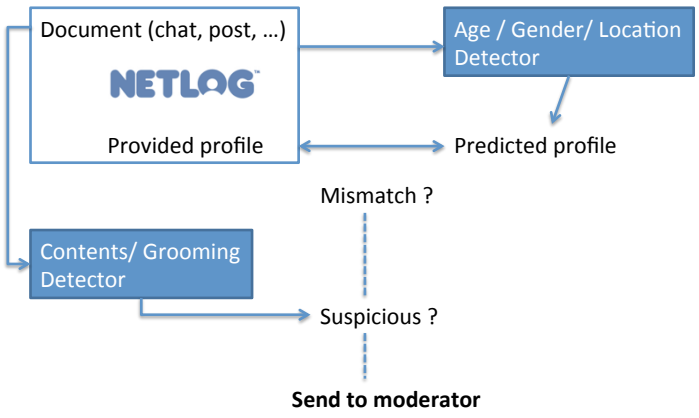



# DAPHNE project

- Goal: Automatically detect pedophiles in social networks
- Data extracted from Dutch part of Netlog
  - http://nl.netlog.com
  - Text and metadata about gender, age, location
- Compare information provided in profile with profile predicted by stylometry
- Include features on the basis of properties of "grooming" language of pedophiles

# DAPHNE project

Defending Against Paedophiles in Heterogeneous Network Environments



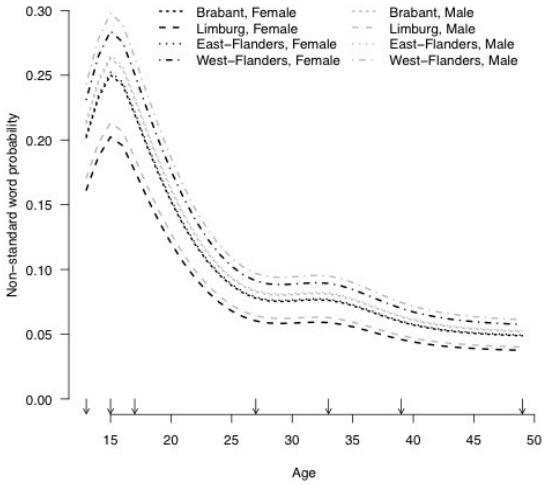

# Maxims of (Dutch) chat language

- Write as fast as you can to ensure a fluent interaction
- Write the way you speak to ensure the informal character of the conversation
  - (Vandekerckhove, 2010)
- Huge problem for automatic text analysis (even POS tagging is impossible)
  - Normalization
  - Special purpose tools
  - But: blessing in disguise for accurate prediction!

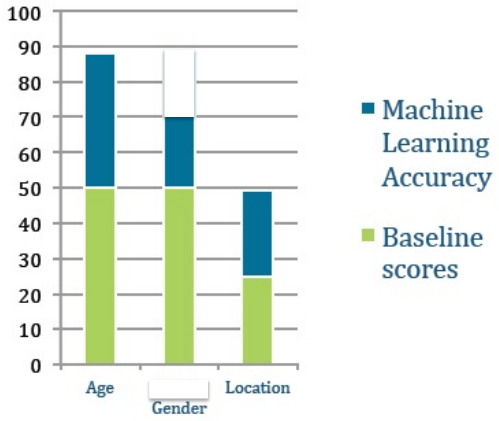
# Properties of chat language

| Variation type | Netlog example | Standard Dutch | English |
|---|---|---|---|
| Omission of letters or words | kbda nimr | Ik heb dat niet meer. | I don't have that anymore. |
| Abbreviations | wrm<br>W8 | waarom<br>wacht | why<br>wait |
| Acronyms | hjg | hou je goed | take care |
| Character flooding | keiii mooiii | heel mooi | very beautiful |
| Concatenation | IkKanOokNiiiZonderU! | Ik kan ook niet zonder jou! | I can't live without you either! |

# Language Variation in Netlog chat



# ASL detector



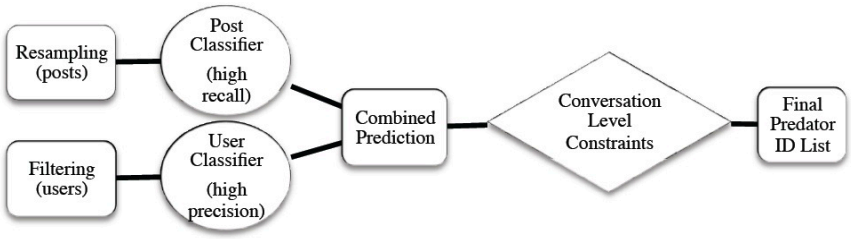

# Results

- Given the difficulty of text analysis on these data, it seems logical that bags of words perform best
  - Different age groups and genders
    - use different intensifiers, emoticons, words in general …
- Character n-grams underperforming!

# PAN 2012 competition (CLEF) "Sexual predator identification in chat"

- Data:
  - Transcripts of www.perverted-justice.com
    - Volunteers try to trick pedophiles
    - Controversial and artificial data!
- Tasks
  - Identify pedophiles
  - Identify most distinctive utterances

# Task 1 Architecture



# Handling skewed data

- Post classifier
  - Training data balanced for + / - written by predator
  - Aggregated to individual user
  - High recall, low precision
- User classifier
  - Training data only users with at least one 'suspicious posts (better training data)
  - High precision, low recall
- Combination
  - Weighted Voting
- Postprocessing
  - Reduce false positives by looking at complete conversation

# Results (cross-validation)

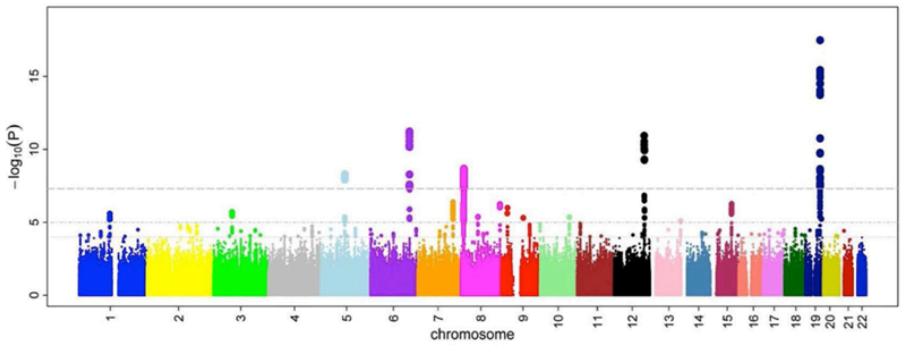
| Results | True Positives | False Negatives | False Positives | Predator Precision | Predator Recall | Predator F-score |
|---|---|---|---|---|---|---|
| Post Classifier | **56** | 4 | 93 | 0.38 | **0.93** | 0.54 |
| User Classifier | 49 | 11 | **7** | 0.88 | 0.82 | 0.84 |
| Post + User | 51 | 9 | 10 | 0.84 | 0.85 | 0.84 |
| After Post-processing | 51 | **9** | 3 | **0.94** | 0.85 | **0.90** |

# Task 2: Suspicious Utterances

- Behavioral analysis
  - Different stages in online grooming (Lanning, 2010)
  - Analysis of positive training data
- Dictionary-based filter
  - Terminology related to:
    - Sexual topic, reframing, approaching, requesting data, isolating from supervision, age-related references
- Results in the PAN competition
  - Highest f-score task 2, Sixth place task 1 (overfitting: f-score = .72)

# Stylome-wide association studies

- Genome-wide association studies (from 2005)



## Stylome-wide association studies

- Associate linguistic properties with author traits
  - Compare subcorpora with and without the trait
  - Compare individual subcorpus to reference corpus
- Large reference corpus needed with authors with various traits of interest
  - Sociological, psychological, …
  - **Sample stratification**
- Reliable lexical, syntactic, semantic, … features needed
- This research only provides associations, no *explanations*

## Conclusion

- Mapping linguistic features to author identity or author traits for prediction has met with some success
- But:
  - We want explanation rather than prediction
    - E.g. what does a bunch of predictive character n-grams mean?
  - We want authorship verification rather than attribution
    - More metalearning work like Unmasking
  - We want scalable methods
  - Interaction of style variation with topic and genre variation is unresolved
  - Text analysis state of the art may not be good enough for semantic / discourse features and for social media language
- We need large scale systematic studies on large balanced corpora

## Acknowledgements:
www.clips.ua.ac.be

- Dr. Guy De Pauw (alzheimer)
- Dr. Kim Luyckx (authorship, personality)
  Now at University Hospital Antwerp
- Dr. Mike Kestemont (medieval authorship)
- Claudia Peersman (age, gender, pedophiles)
- Frederik Vaassen (emotion, communication attitudes)
- Tom De Smedt (communicative stance, emotion)
- MA and AMA students
- Other themes:
  - Ideology, deception, suicide letters, cyberbullying, …


## www.clips.ua.ac.be