Text Analysis for Social Media Cybersecurity: the AMiCA Project

Els Lefever

Language and Translation Technology Team (LT^3)
Ghent University, Belgium

UNIVERSITEIT GENT

language and
translation
technology
team

LT^3, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

LT^3

- Dpt of Translation, Interpreting and Communication, Faculty of Arts and Philosophy, Ghent University

- fundamental and applied research in **language and translation technology**

- expertise in using **machine learning** for language technology problems (PoS-tagging and lemmatization, anaphora resolution, WSD, NER)

- Headed by Prof. Véronique Hoste

3 main research lines:

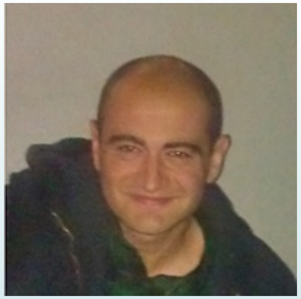
- Terminology & computational semantics
- Translation Technology
- Sentiment analysis and subjectivity detection

Terminology / computational semantics

- Lead: Prof. Els Lefever
- Automatic terminology extraction from monolingual, bilingual and comparable corpora (Ayla Rigouts Terryn)
- Automatic hypernym and synonym detection (Els Lefever)
- Term ambiguity in interdisciplinary research (Julie Mennes)
- Use of term extraction for translating documentaries (Sabien Hanoulle)

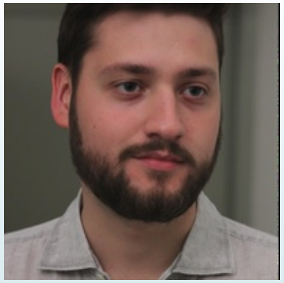
Translation Technology

- Lead: Prof. Lieve Macken
- comparison of different methods of translation: human vs. post-editing, human vs. CAT (Joke Daems)
- translation quality assessment and confidence estimation for machine translation (Arda Tezcan)



Sentiment Analysis and Subjectivity detection

- Lead: Prof. Véronique Hoste
- automatic detection of cyberbullying (Cynthia Van Hee)
- suicide detection (Bart Desmet)
- Aspect-based sentiment Analysis (Orphée De Clercq)
- detection of subjectivity in annual reports (Nils Smeuninx)
- Irony detection (Cynthia Van Hee)
- Sentiment Analysis for economic events (Gilles Jacobs)



AMICA

Outline

- The context and goals of the AMiCA project
- Text normalization
- 3 Use cases:

1. Detecting cyberbullying
2. Suicide detection
3. Age and gender profiling for detecting grooming

www.amicaproject.be

- IWT-SBO project, coordinated by CLiPS (UA)
- Partners:
  - CLiPS (text mining, UA)

    CLiPS
    Computational Linguistics & Psycholinguistics
    University of Antwerp
  - MIOS (sociology, UA)

    MIOS
    Media & ICT in Organisations & Society
    University of Antwerp
  - LT3 (text mining, UGent)

    language and
    translation
    technology
    team
  - IBCN (software development, UGent)

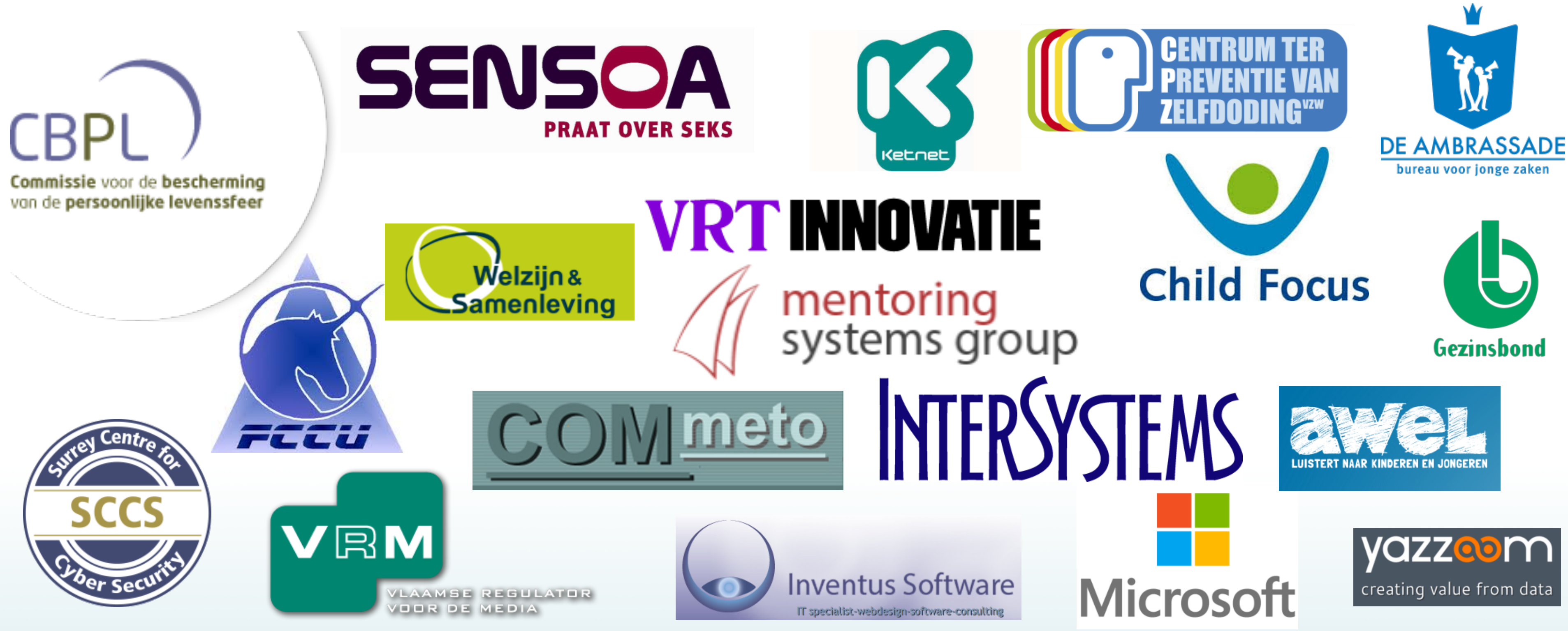
    IBCN
  - VISICS (image processing, KUL)

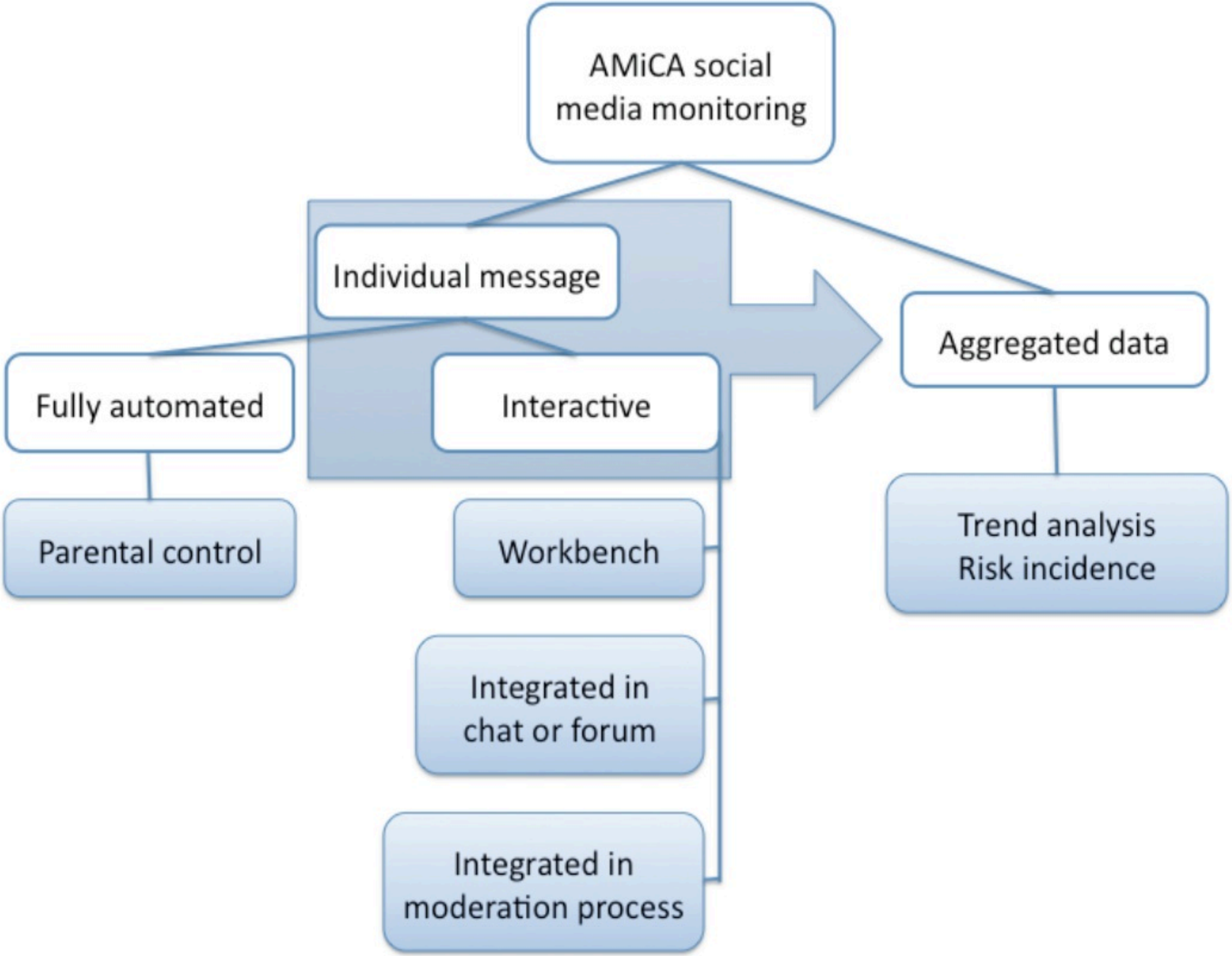
    PSI
    VISICS
- Combine text analytics, image and video analysis, and data mining

Goals

- Detect situations that are harmful or threatening to young people in social networks
  - Cyberbullying
  - Sexually transgressive behaviour (for example grooming by paedophiles)
  - Depression and suicide announcement
- Facilitate efficient action by moderators, police, parents, peer group, social services, ...
- Objective measurement, monitoring, trend analysis, …

User Committee


CBPL
Commissie voor de bescherming van de persoonlijke levenssfeer
SENSOA
PRAAT OVER SEKS
Ketnet
CENTRUM TER PREVENTIE VAN ZELFDODING VZW
DE AMBRASSADE
bureau voor jonge zaken
VRT INNOVATIE
Welzijn & Samenleving
mentoring systems group
Child Focus
Gezinsbond
FCCU
COMmeto
INTERSYSTEMS
awel
LUISTERT NAAR KINDEREN EN JONGEREN
Surrey Centre for Cyber Security
SCCS
VRM
VLAAMSE REGULATOR VOOR DE MEDIA
Inventus Software
IT specialist-webdesign-software-consulting
Microsoft
yazzoom
creating value from data


AMiCA social media monitoring
Individual message
Aggregated data
Fully automated
Interactive
Parental control
Workbench
Integrated in chat or forum
Integrated in moderation process
Trend analysis
Risk incidence

How urgent is the problem?

- European “Kids online” study (EU, 2011)
  - Motivation for the project
  - Age 9-16 in 25 European countries
  - Results
    - Children are 90 minutes per day online
    - Half of them in their bedroom
    - 33% added strangers as friends
    - 15% shared personal information with strangers (Including photographs)
    - 12% felt they experienced harm

    www.eukidsonline.net

How urgent is the problem?

- European “Kids online” study: update in 2014
  - Age 9-16 in 25 European countries
  - Results since 2010 study, 9 to 16 year olds
    - Significant rise of use of social media
    - Rise of 23% to 43% of having contact with someone not met IRL before
    - Rise of 10% to 23% of having seen sexual images
    - Rise of 9% to 20% of having received sexual images
    - Rise of 13% to 17% are upset by something seen online
    - Rise of 13% to 20% of being exposed to hate messages
    - Rise of 7% to 11% of being exposed to self-harm sites
    - Rise of 7% to 12% of being exposed to cyberbullying

www.eukidsonline.net

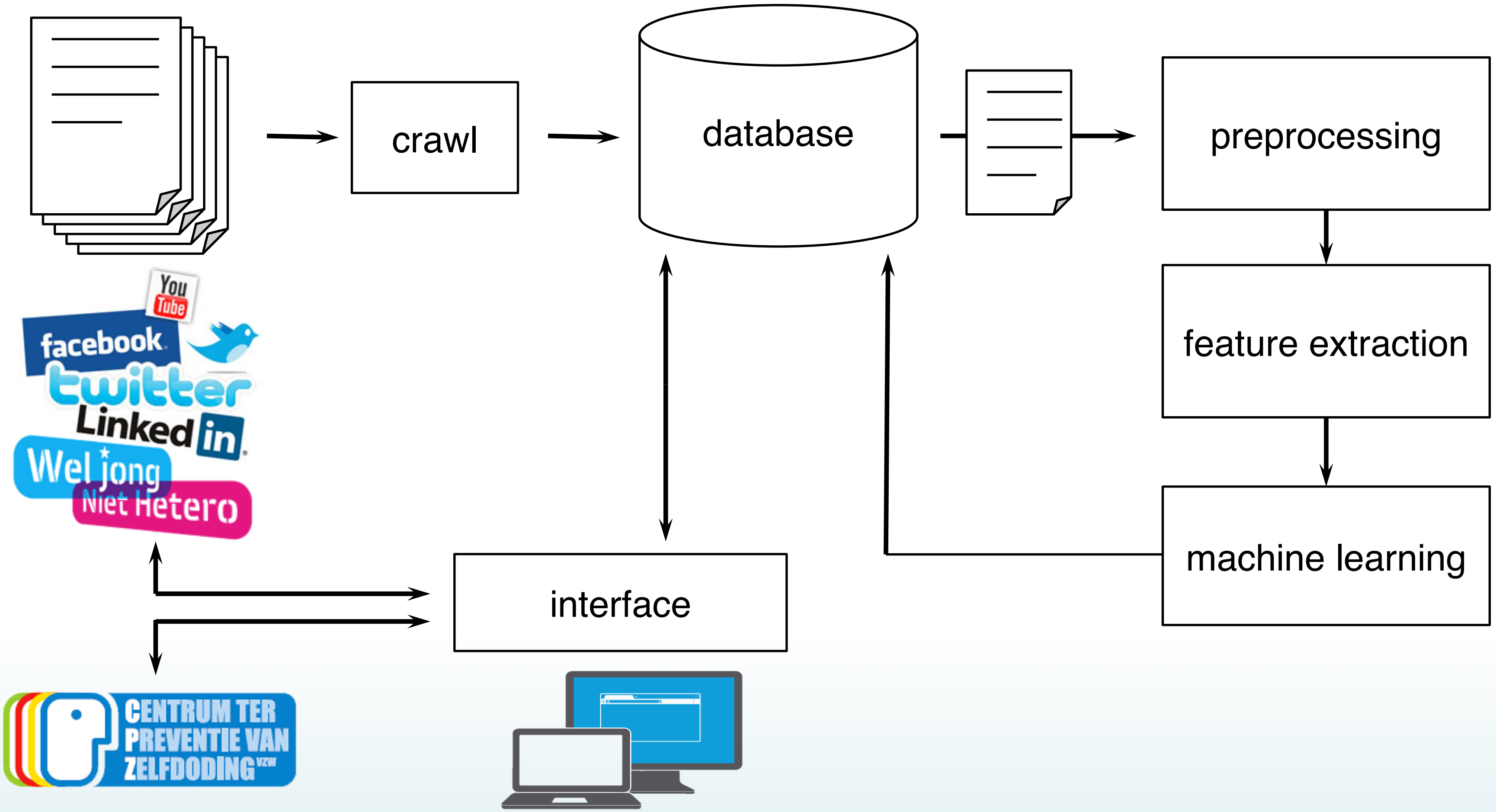
Quick poll

- Who is in favor of software monitoring automatically your interactions in social media for risks and threats?

Should we do something about it?

- Majority of experts and adolescents is in favor of automatic monitoring
  - but only for situations they perceive as uncontrollable
  - with respect for privacy and with suitable follow-up, not involving too many parties, and giving control to the victim
- Mixed opinions with the parents depending on (negative) previous experience and level of trust in their children

Workflow

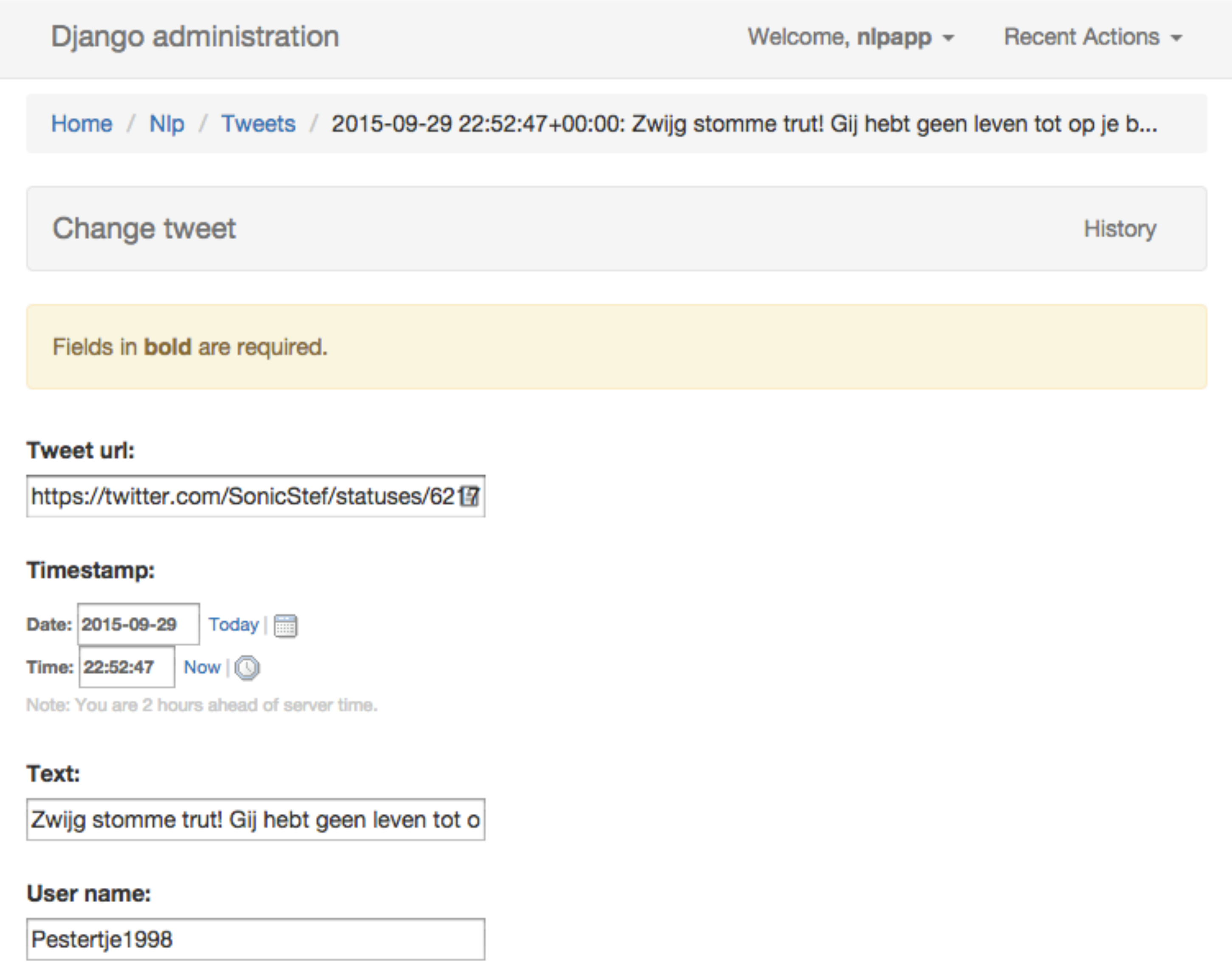

crawl
database
preprocessing
feature extraction
machine learning
interface
You Tube
facebook
twitter
Linked in
Wel jong Niet Hetero
CENTRUM TER PREVENTIE VAN ZELFDODING vzw

Crawl: example

Zwijg stomme trut! Gij hebt geen leven tot op je begravenis!!!

(English: Shut up stupid cow! You don’t have a life see you at your funeral!!!)

Crawl: example



Django administration

Home / Nlp / Tweets / 2015-09-29 22:52:47+00:00: Zwijg stomme trut! Gij hebt geen leven tot op je b...

Change tweet

Fields in **bold** are required.

Tweet url:

https://twitter.com/SonicStef/statuses/6217

Timestamp:

Date: 2015-09-29 Today |

Time: 22:52:47 Now |

Note: You are 2 hours ahead of server time.

Text:

Zwijg stomme trut! Gij hebt geen leven tot o

User name:

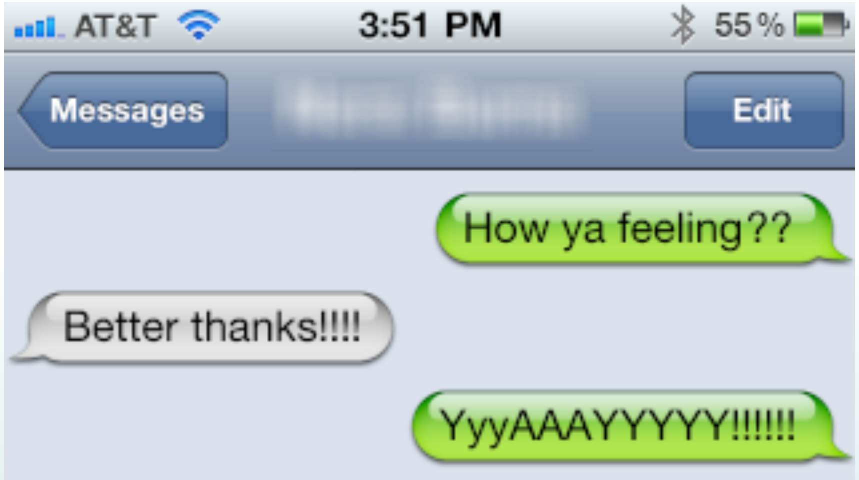
Pestertje1998

PREPROCESSING / NORMALISATION OF USER-GENERATED TEXT

User Generated Content

Social media: blogs and microblogs (Twitter: 190 million tweets/day), wikis, podcasts, social networks (Facebook: 70 billion shares/month)

⇒Enormous amount of UGC



How ya feeling??

Better thanks!!!!

YyyAAAYYYYY!!!!!!

Properties of chat language

- Omission of words / characters (spoke – spoken)
- Abbreviations, acronyms (LOL – laughing out loud)
- Deviations from standard spelling (luv – love, you iz – you are)
- Expression of emotion:
  - Flooding (loooooooove)
  - Emoticons (:p)
  - Capitalized letters (STUPID)
- Dutch-specific:
  - Concatenation of tokens (khou – ik hou)
  - Elimination of clitics and pronouns (edde – heb je)
  - Lot of dialects!

Example

	Example of Dutch SMS language
Original	Oguz ! Edde me Jana gesproke ? En ze flipt lyk omdak ghsmoord heb .. !
Normalized	Oh gods ! Heb je met Jana gesproken ? En ze flipt gelijk omdat ik gesmoord heb .. !
Translated	Oh god ! Did you speak to Jana ? And she’s flipping because I smoked … !

Problem for Text Analysis Tools

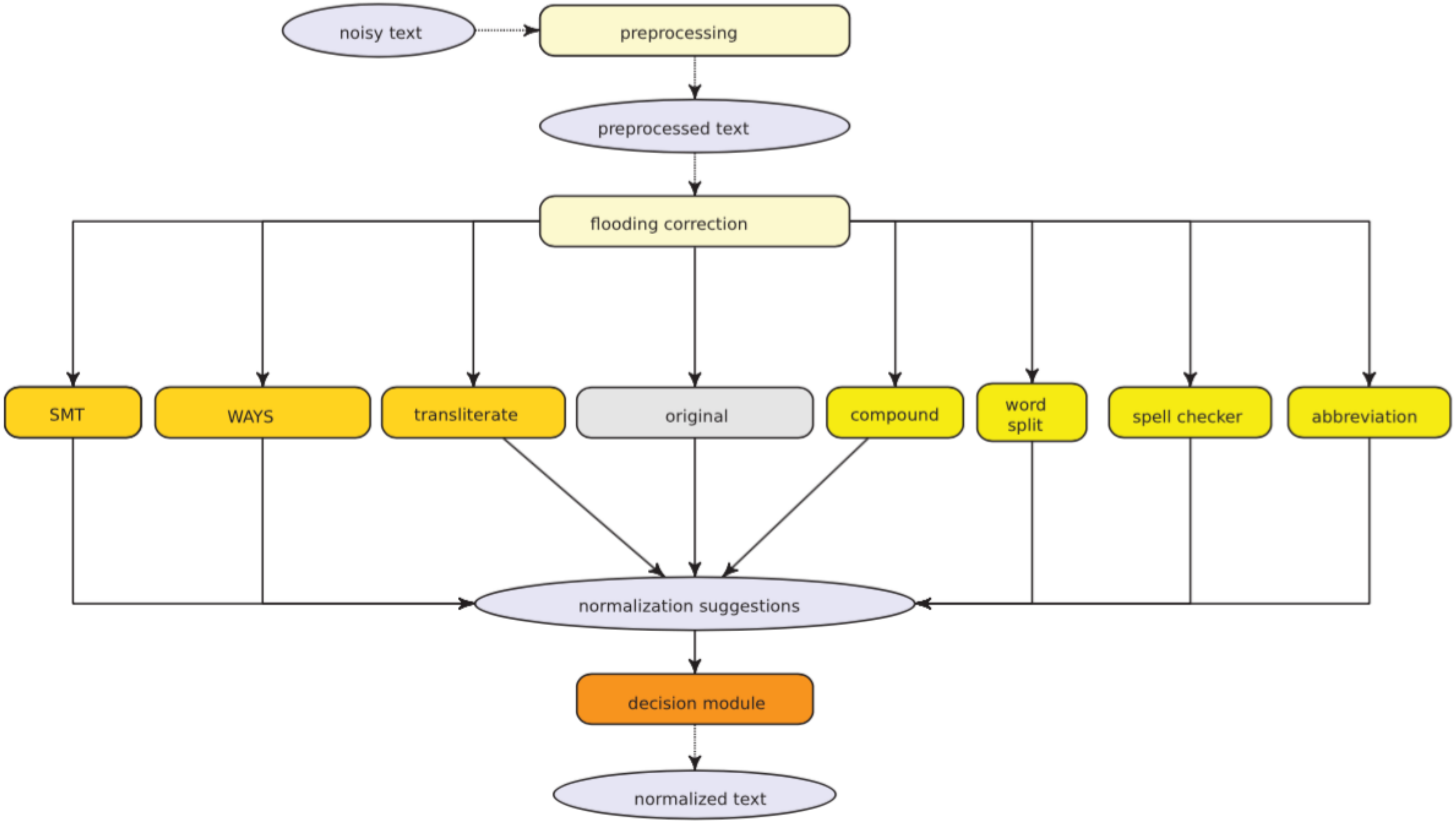
- Most NLP tools are developed for or trained on standard language
- They fail miserably on UGC
- Solutions
  - Develop new tools
    - E.g. Tweet NLP (CMU): http://www.cs.cmu.edu/~ark/TweetNLP/
  - Normalize the ‘non-standard’ language
- On the positive side, non-standard language makes some analytics tasks easier!

Normalization Approaches

- Three dominant approaches
  - Machine Translation: Source Language = non-standard and Target Language = standard
  - Spell Checking: Correct the incorrect words (statistical or dictionary-based)
  - Speech Recognition: Non-standard language = speech that has to be converted to text (HMMs)

=> We choose to follow an SMT approach and also go to the character-level

Ensemble Approach



Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. 2016. Multimodular text normalization of Dutch user-generated content. *ACM Trans. Intell. Syst. Technol.* 7, 4, (July 2016), 22 pages. DOI: http://dx.doi.org/10.1145/2850422

Modules

- Preprocessing
  - Tokenization and sentence splitting
    - includes emoticons, emojis etc.
  - Character flooooooooding
- Token-based modules
  - Abbreviations
    - Expansion dictionary (~ 350 abbrevs)
  - Spell checker
    - Levenshtein on dictionary (~ 2.3 million words)
  - Compound Module
    - Checks if a pair of words is actually one word
  - Word Splitter
    - ‘misje’ = ‘mis je’ (miss you)

Modules

- Context-based modules
  - SMT
    - Token-unigram, character unigram, character-bigram and combinations
  - Transliteration (supervised ML)
    - supervised ML, memory-based learning style
      - +da+_n **i** ++_ged -> **iet**
  - WAYS (Write As You Speak): G2P + P2G (memory-based learning)
    - ni (niet, *not*)
    - kem (ik heb, *I have*)
- "Original" Module
  - Many words are correct

Modules

- Decision Module
  - Moses decoder (SMT), dynamic search among the suggestions of the component modules
  - Uses (5-gram) language model and phrase table (dev. Set)

Evaluation

- Three types of UGC
  - Chat (Netlog)
  - SMS (Sonar corpus)
  - Microblog (Twitter)
- Train (60%) - Development (20%) - Test (20%)
- Total: 70,000 tokens, manually annotated
  - insertions, deletions, substitutions, transpositions
  - near-perfect annotator agreement
- Background corpora for language modeling

CGN (Spoken Dutch Corpus)	6,765,336
SoNaR (Balanced text corpus)	3,581,182
Open Subtitles Dutch (OSD)	90,147,315
Training set (TS)	56,523

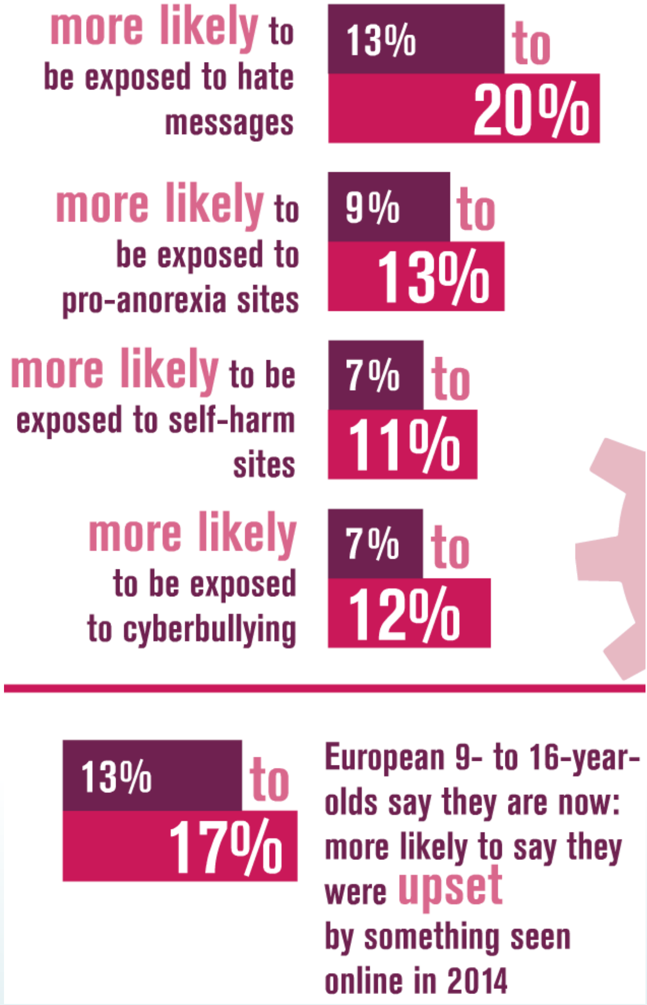
Results

- Module level evaluation:
  - SMT and Transliterate modules perform best
    - Especially compounding and splitting problems remain
- Ensemble evaluation:
  - Best ensemble system: 92.9
- Extrinsic and Portability Evaluation
  - Tested on Ask FM for NLP tasks (with and without normalizing)
    - POS (+12%), LEM (+13%), NER (+8%)
- Problems remain especially in tokens with multiple normalization problems

USE CASE 1: CYBERBULLYING DETECTION

Research Motivation

- ± 20-40% of all youth have been victimized online (Tokunaga, 2010)
- Anonymity, lack of supervision and impact make social media a convenient way for cyberbullies to target their victim (Hinduja & Patchin, 2006)
- Information overload on the Web has made manual monitoring unfeasible



more likely to be exposed to hate messages: 13% to 20%

more likely to be exposed to pro-anorexia sites: 9% to 13%

more likely to be exposed to self-harm sites: 7% to 11%

more likely to be exposed to cyberbullying: 7% to 12%

13% to 17%: European 9- to 16-year-olds say they are now: more likely to say they were upset by something seen online in 2014

Source: the EU Kids Online report (2015)
http://www.lse.ac.uk/media@lse/research/EUKidsOnline

Research Motivation

- Automatic detection systems allow for large-scale social media monitoring
- Goal => reduce manual monitoring efforts on social media

Related Research

- NLP applications for automatic cyberbullying prevention and detection
  - Cyberbullying detection (Yin et al., 2009; Reynolds et al., 2011; Nahar et al., 2013)
  - Sensitive topic identification (sexuality, race) (Dinakar et al., 2012)
  - Detection of bully profiles on social networks (Dadvar et al., 2013)

BUT:

- Focus on posts from harassers
- No distinction between different types of cyberbullying
- Datasets do not always follow a real-world distribution

Data set construction

- We need large data sets to train machine learning systems
- Data collection for Dutch and English
  - Data from relevant social media
  - BUT: few / private data
  - Media campaign for donating examples of cyberbullying messages
  - BUT: sensitive data!
  - Cyberbullying simulations

Data set construction: media campaign

HET NIEUWSBLAD WOENSDAG 27 NOVEMBER 2013 11

VLAAMSE ONDERZOEKERS ONTWIKKELEN SYSTEEM TEGEN CYBERPESTEN

'Stuur ons uw haatmail'

Waarom draagt gij altij make up aan, Verberg u lelijkheid nie helpt tog nie

gij zijt egt ne mega nerd hoe gij durft buitenkomen??? ik snap het

AMiCA

Gezocht: cyberpestberichten

Beste,

In het kader van ons onderzoek naar cyberpesten aan de universiteiten van Antwerpen, Leuven en Gent hebben we een oproep gelanceerd om zoveel mogelijk

radio 2 altijd dicht bij jou

De Universiteit van Gent zoekt cyberpestberichten

Tags: *cyberpesten*, *gent*, *universiteit*

Een aantal onderzoeksgroepen van de **universiteiten van Antwerpen, Gent en Leuven** zijn een systeem aan het

RESULT: ± 30 reactions

± 368 messages (FB messages, hate pages, Netlog, mail, chat, etc.)

Het doel van AMiCA (Automatic Monitoring for Cyberspace Applications) is een systeem te ontwikkelen dat automatisch cyberpestgedrag herkent op sociaalnetwerksites. 'Omdat te kunnen doen, hebben we enorm veel gegevens nodig', zegt Véronique Hoste van UGent. 'E-mails, sms'en, Facebook- en chatberichten ...'

Vandaar de opvallende oproep die zij en haar collega's gisteren lanceerden. 'Of je nu het slachtoffer bent van cyberpesten of een getuige, leerkracht of ouder, bezorg ons de berichten die je beschouwt als cyberpesten.' De onderzoekers benadrukken dat die volledig anoniem behandeld worden.

Het systeem dat ze met die gegevens zullen ontwikkelen, moet weldra elk bericht scannen dat binnenkomt op Twitter of eender

VÉRONIQUE HOSTE Universiteit Gent

Er bestaan nu al knoppen waarmee je kwetsende berichten kunt melden. Wij willen dat systeem automatiseren

Het project van de Vlaamse universiteiten is door zijn omvang nu al uniek. Maar voorlopig focussen ze zich enkel op het Nederlands. 'Al kan het systeem in een later stadium ook voor het Engels gebruikt worden. Maar dan zullen we opnieuw manueel gegevens moeten verzamelen.'

INFO **Bezorg je cyberpestberichten via data@amicaproject.be**

vindt u op de website:
http://www.amicaproject.be/

Voor meer informatie over deze oproep of ons project kan u steeds terecht bij:

Cynthia Van Hee	Universiteit Gent	cynthia.vanhee@ugent.be
Ben Verhoeven	Universiteit Antwerpen	ben.verhoeven@uantwerpen.be

UNIVERSITEIT GENT Universiteit Antwerpen iWT

Nieuw systeem spoort online pestgedrag automatisch op

SAM FEYS
BRUSSEL

Belgische onderzoekers willen cyberpesten tegengaan door pestgedrag op Facebook en Twitter automatisch op te sporen. Het systeem moet op termijn zelfs ingrijpen bij zelfmoordgedachten.

Cyberpesten is vandaag nog erg moeilijk op te sporen, zeker omdat slachtoffers er zelden voor uitkomen. Verschillende onderzoekers van de universiteiten van Gent, Leuven en Antwerpen slaan daarom de handen in elkaar voor AMiCA, een systeem dat automatisch pestgedrag op het internet moet herkennen. Zo kunnen zowel sitebeheerders als politie ingrijpen voor het te laat is.

"We willen zowel tekst als beeld op sociale media analyseren om zo volautomatisch alarmsignalen te laten afgaan bij verontrustende situaties", legt Véronique Hoste van UGent het project uit. "Zo kunnen we zowel gevallen van online pestgedrag en pedofilie opsporen, maar ook ingrijpen als jongeren met suïcidale gedachten kampen."

Het onderzoeksteam is daarom op zoek naar slachtoffers van cyberpesters die ooit beledigende sms'jes, e-mails of berichten kregen, ook via socialenetwerksites als Twitter, Facebook en Ask.fm. "Die zullen we uitgebreid analyseren. Op basis daarvan kunnen we accurate computermodellen bouwen die kunnen afleiden wanneer het om cyberpesten gaat. Hoe meer bruikbare voorbeelden we binnenkrijgen, hoe beter we dat model kunnen voeden."

De bedoeling is om uiteindelijk elk nieuw bericht op Facebook of Twitter te toetsen aan dat model. "Elk bericht zal een label meekrijgen", verduidelijkt Hoste. "Als bij meerdere berichten van een gebruiker rode lampjes gaan branden, zal een moderator of sitebeheerder automatisch op de hoogte worden gebracht. Die kan dan de ernst van de situatie inschatten en eventueel de politie inschakelen."

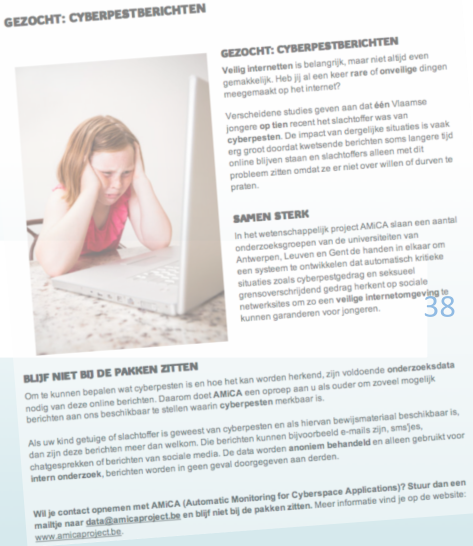
'Elk bericht op Facebook zal een label meekrijgen'
VÉRONIQUE HOSTE
ONDERZOEKER UGENT

De onderzoekers werken samen met onder meer Netlog, Child Focus en met de VRT, dat dagelijks tienduizenden kinderen naar ketnet.be ziet komen. "Het is voor ons heel belangrijk dat een site als Ketnet een veilige omgeving vormt voor kinderen", beseft VRT-woordvoerster Anneke Ernon. "We zetten sterk in op sensibilisering en de webredactie leest ook nog eens alle berichten na die de kinderen online plaatsen."

"Op sites als Facebook is volledige manuele controle echter onmogelijk", weet Hoste. "Er is gewoon te veel input."

Het Belgische onderzoeksteam speelt internationaal een voortrekkersrol in het automatisch opsporen van pestgedrag. Of het project uiteindelijk zal slagen, zal grotendeels afhangen van de medewerking van Twitter en vooral Facebook, waar veel berichten niet openbaar zijn.

Cyberpestberichten kunnen via data@amicaproject.be worden doorgestuurd aan het onderzoeksteam.



GEZOCHT: CYBERPESTBERICHTEN

GEZOCHT: CYBERPESTBERICHTEN

Veilig internetten is belangrijk, maar niet altijd even gemakkelijk. Heb jij al een keer **rare** of **onveilige** dingen meegemaakt op het internet?

Verscheidene studies geven aan dat **één** Vlaamse jongere **op tien** recent het slachtoffer was van **cyberpesten**. De impact van dergelijke situaties is vaak erg groot doordat kwetsende berichten soms langere tijd online blijven staan en slachtoffers alleen met dit probleem zitten omdat ze er niet over willen of durven te praten.

SAMEN STERK

In het wetenschappelijk project AMiCA slaan een aantal onderzoeksgroepen van de universiteiten van Antwerpen, Leuven en Gent de handen in elkaar om een systeem te ontwikkelen dat automatisch kritieke situaties zoals cyberpestgedrag en seksueel grensoverschrijdend gedrag herkent op sociale netwerksites om zo een **veilige internetomgeving** te kunnen garanderen voor jongeren.

BLIJF NIET BIJ DE PAKKEN ZITTEN

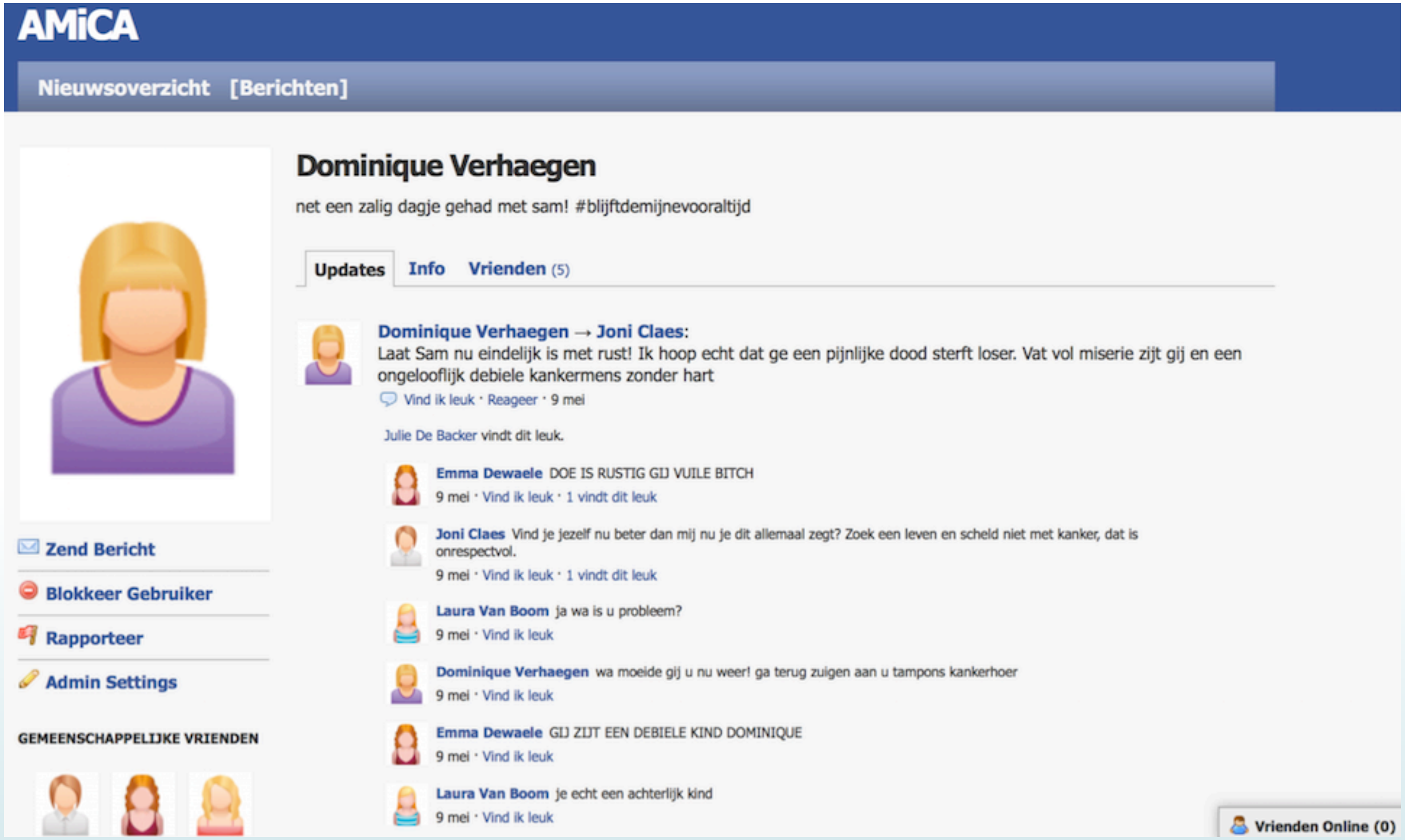
Om te kunnen bepalen wat cyberpesten is en hoe het kan worden herkend, zijn voldoende **onderzoeksdata** nodig van deze online berichten. Daarom doet **AMiCA** een oproep aan u als ouder om zoveel mogelijk berichten aan ons beschikbaar te stellen waarin **cyberpesten** merkbaar is.

Als uw kind getuige of slachtoffer is geweest van cyberpesten en als hiervan bewijsmateriaal beschikbaar is, dan zijn deze berichten meer dan welkom. Die berichten kunnen bijvoorbeeld e-mails zijn, sms'jes, chatgesprekken of berichten van sociale media. De data worden **anoniem behandeld** en alleen gebruikt voor **intern onderzoek**, berichten worden in geen geval doorgegeven aan derden.

Wil je contact opnemen met AMiCA (Automatic Monitoring for Cyberspace Applications)? Stuur dan een mailtje naar data@amicaproject.be en blijf niet bij de pakken zitten. Meer informatie vind je op de website: www.amicaproject.be.

Dataset Construction: simulation experiments

- Role playing in secondary schools on social media platform: FB-like social network, scenarios, profile cards (roles), debriefing
- Additional goal: education (prevention)



Dominique Verhaegen

net een zalig dagje gehad met sam! #blijftdemijnevooraltijd

Dominique Verhaegen → Joni Claes:
Laat Sam nu eindelijk is met rust! Ik hoop echt dat ge een pijnlijke dood sterft loser. Vat vol miserie zijt gij en een ongelooflijk debiele kankermens zonder hart
Vind ik leuk · Reageer · 9 mei

Julie De Backer vindt dit leuk.

Emma Dewaele DOE IS RUSTIG GIJ VUILE BITCH
9 mei · Vind ik leuk · 1 vindt dit leuk

Joni Claes Vind je jezelf nu beter dan mij nu je dit allemaal zegt? Zoek een leven en scheld niet met kanker, dat is onrespectvol.
9 mei · Vind ik leuk · 1 vindt dit leuk

Laura Van Boom ja wa is u probleem?
9 mei · Vind ik leuk

Dominique Verhaegen wa moeide gij u nu weer! ga terug zuigen aan u tampons kankerhoer
9 mei · Vind ik leuk

Emma Dewaele GIJ ZIJT EEN DEBIELE KIND DOMINIQUE
9 mei · Vind ik leuk

Laura Van Boom je echt een achterlijk kind
9 mei · Vind ik leuk

Data Annotation

- Brat rapid annotation tool (Stenetorp et al., 2012)
- Two annotation levels (Van Hee et al., 2015)
  - Post level
    - Cyberbullying -vs- non-cyberbullying

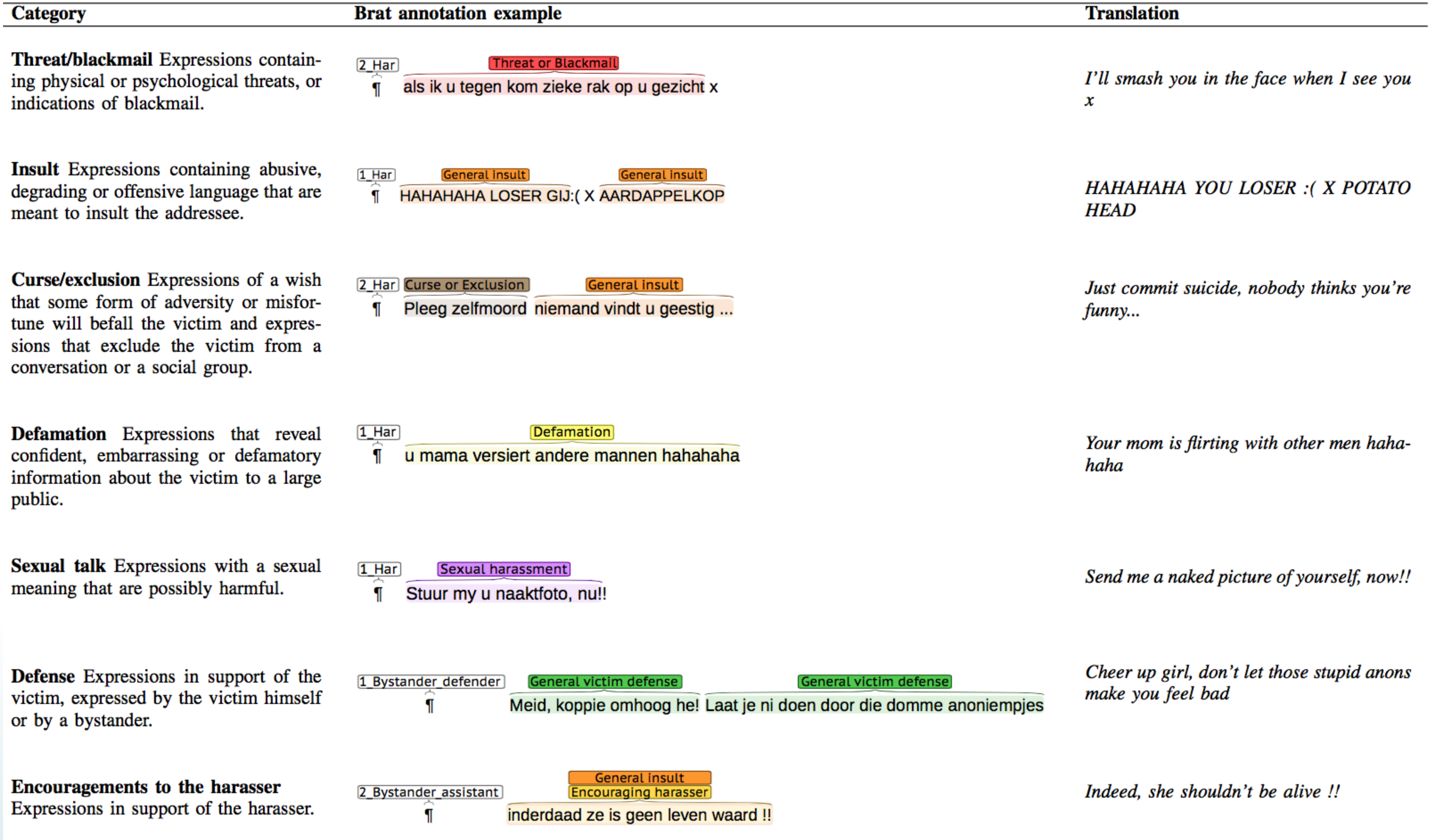
      textual content that is published online by an individual and that is aggressive or hurtful against a victim.
    - Harmfulness score
      - 0 → the post does not contain indications of cyberbullying
      - 1 → the post contains indications of cyberbullying, although they are not severe
      - 2 → the post contains serious indications of cyberbullying
    - Author's role
      - Harasser
      - Victim
      - Bystander-defender
      - Bystander-assistant

Data Annotation

- (Sub)sentence level: identification of fine-grained text categories related to cyberbullying
  - Threat/blackmail
  - Insult
  - Curse/exclusion
  - Defamation
  - Sexual talk
  - Defense
  - Encouragements (to the harasser)

Guidelines for the fine-grained analysis of cyberbullying, version 1.0 (2015)
Van Hee, C., Verhoeven, B., Lefever, E., De Pauw, G., Daelemans, W., & Hoste, V.

Data Annotation



Category	Brat annotation example	Translation
Threat/blackmail Expressions containing physical or psychological threats, or indications of blackmail.	2_Har Threat or Blackmail ¶ als ik u tegen kom zieke rak op u gezicht x	*I'll smash you in the face when I see you x*
Insult Expressions containing abusive, degrading or offensive language that are meant to insult the addressee.	1_Har General insult General insult ¶ HAHAHAHA LOSER GIJ:(X AARDAPPELKOP	*HAHAHAHA YOU LOSER :(X POTATO HEAD*
Curse/exclusion Expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.	2_Har Curse or Exclusion General insult ¶ Pleeg zelfmoord niemand vindt u geestig ...	*Just commit suicide, nobody thinks you're funny...*
Defamation Expressions that reveal confident, embarrassing or defamatory information about the victim to a large public.	1_Har Defamation ¶ u mama versiert andere mannen hahahaha	*Your mom is flirting with other men hahahaha*
Sexual talk Expressions with a sexual meaning that are possibly harmful.	1_Har Sexual harassment ¶ Stuur my u naaktfoto, nu!!	*Send me a naked picture of yourself, now!!*
Defense Expressions in support of the victim, expressed by the victim himself or by a bystander.	1_Bystander_defender General victim defense General victim defense ¶ Meid, koppie omhoog he! Laat je ni doen door die domme anoniempjes	*Cheer up girl, don't let those stupid anons make you feel bad*
Encouragements to the harasser Expressions in support of the harasser.	2_Bystander_assistant General insult Encouraging harasser ¶ inderdaad ze is geen leven waard !!	*Indeed, she shouldn't be alive !!*

Ask.fm preliminary experiments

- Class
  - Binary (bullying or non-bullying)
  - Binary (for each fine-grained class)
- Features
  - Word unigrams and bigrams
  - Character trigrams
  - Sentiment features
- Classifier: SVM (Pattern) with linear kernel
- Data: ~85,000 posts
- Annotation agreement (kappa) 60-65%
- Very skewed data, scarce positive data (~10%)

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. & Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. Proceedings of RANLP, 672–680. Hissar, Bulgaria.

Results

	Precision	recall	F1-score
NL	76%	56%	65%
EN	74%	55%	63%

BUT:

- Ambiguity

  "Hi bitches, anyone in for a movie tonight?"

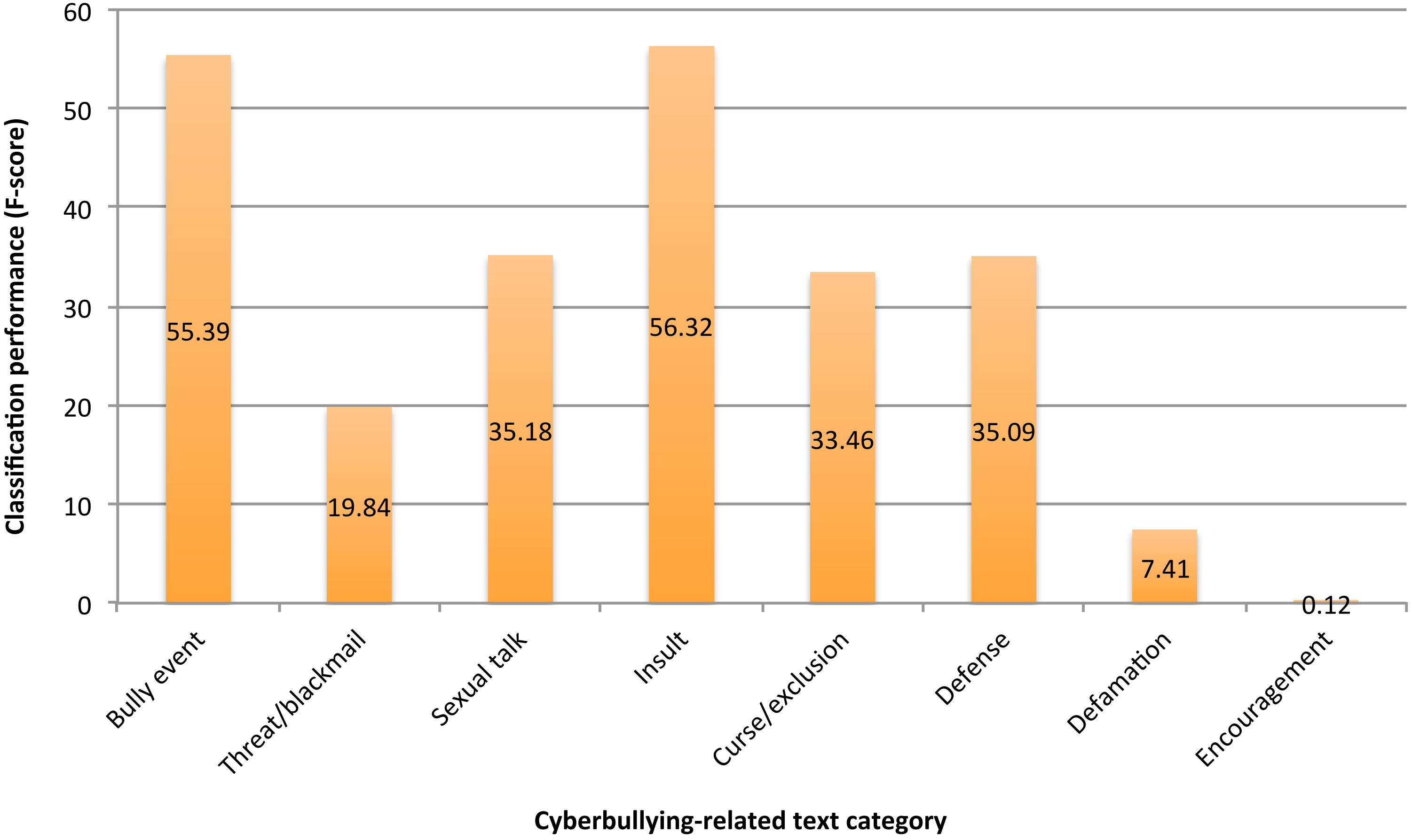
  "Shut up, you bitch!"

- Implicit realizations of cyberbullying

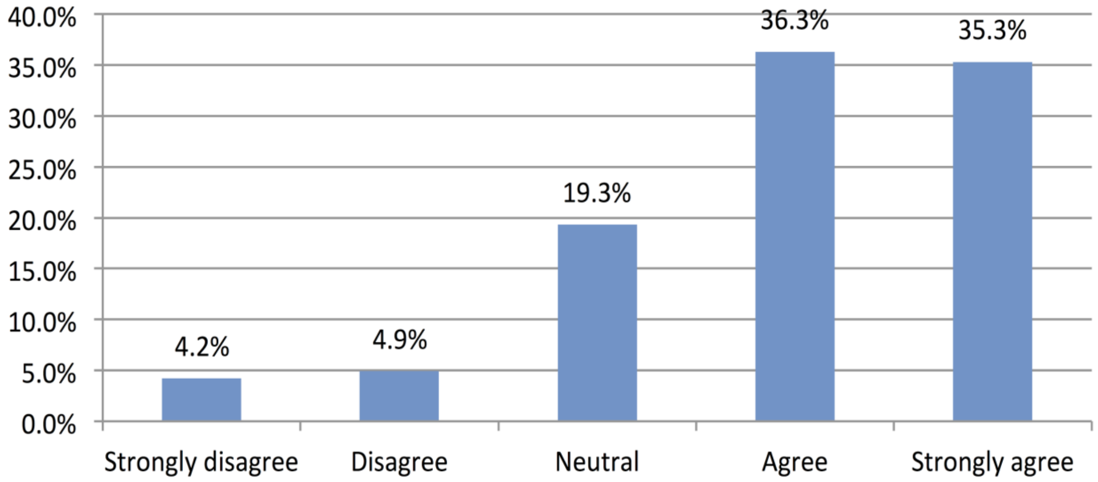
  "You make my fists itch…"

- Data sparseness

Results (Van Hee et al. 2015)



Monitoring desirable?



- Follow-up is needed
- Privacy of youngsters should be respected
- Technical feasibility?

(Van Royen et al., 2014)

More info?

Cynthia Van Hee: cynthia.vanhee@ugent.be

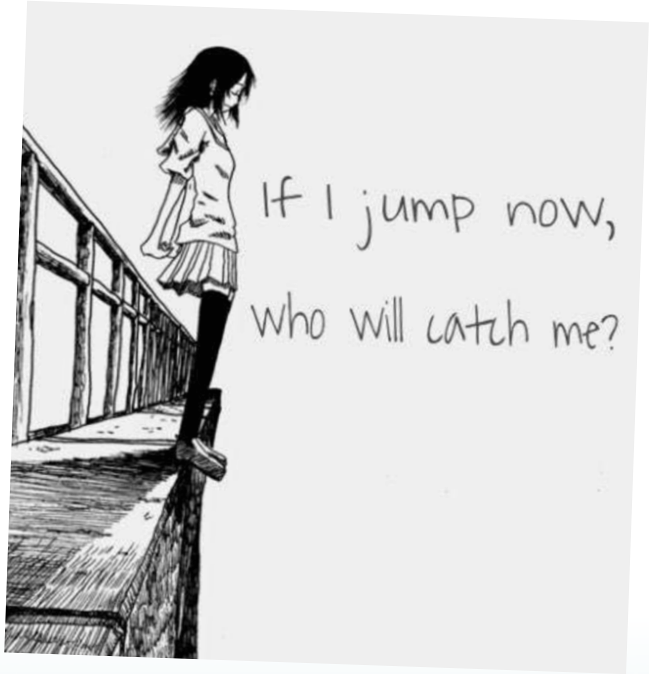
USE CASE 2: SUICIDE DETECTION

Alarming figures Flemish adolescents

- **Self-mutilation:**
  - Every year by 7% at the age of 14-17
  - 2/3 through cutting & scratching (Van Rijsselberghe et al., 2009)
- **Suicidal behaviour:**
  - 15-20% (age of 18) have thoughts of suicide (more than once) (Hublet et al., 2010)

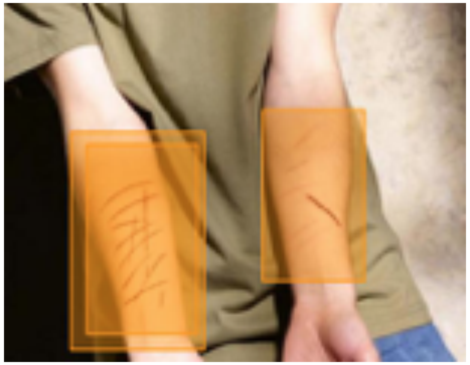
Online self-harm behaviour

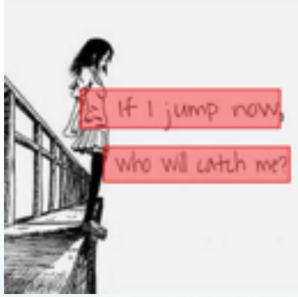
Kheb het al 3 keer geprobeerd, ma kloop ier nog altijd rond... soms zeg ik spijtig genoeg, soms ben ik ook blij dat ik nog leef.

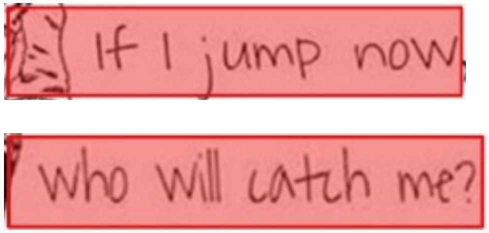


AMiCA technology: image analysis

- Automatic classification of images
- Object recognition in images
- Tekst recognition in images + OCR







```
If I jump now

who will catch me?
```

AMiCA technology: text analysis

Machine learning system **analyses** every message (word sequences, topic models, sentiment analysis, …) and **answers two questions:**

- Is the message about suicide?

✖ I never thought about cutting or suicide, because it leaves scars …

- Is there a serious suicidal threat?

✖ I already tried 3 times, but I'm still alive

✖ Sometimes I feel bad, sometimes I'm glad I'm still alive

Text analysis: results

Experiments carried out on a data set of 10,000 messages, of which 851 are relevant and 257 are serious:

- Is the message about suicide? => recall: 9/10, 3% noise
- Is there a serious suicidal threat? => recall: 2/3, 25% noise

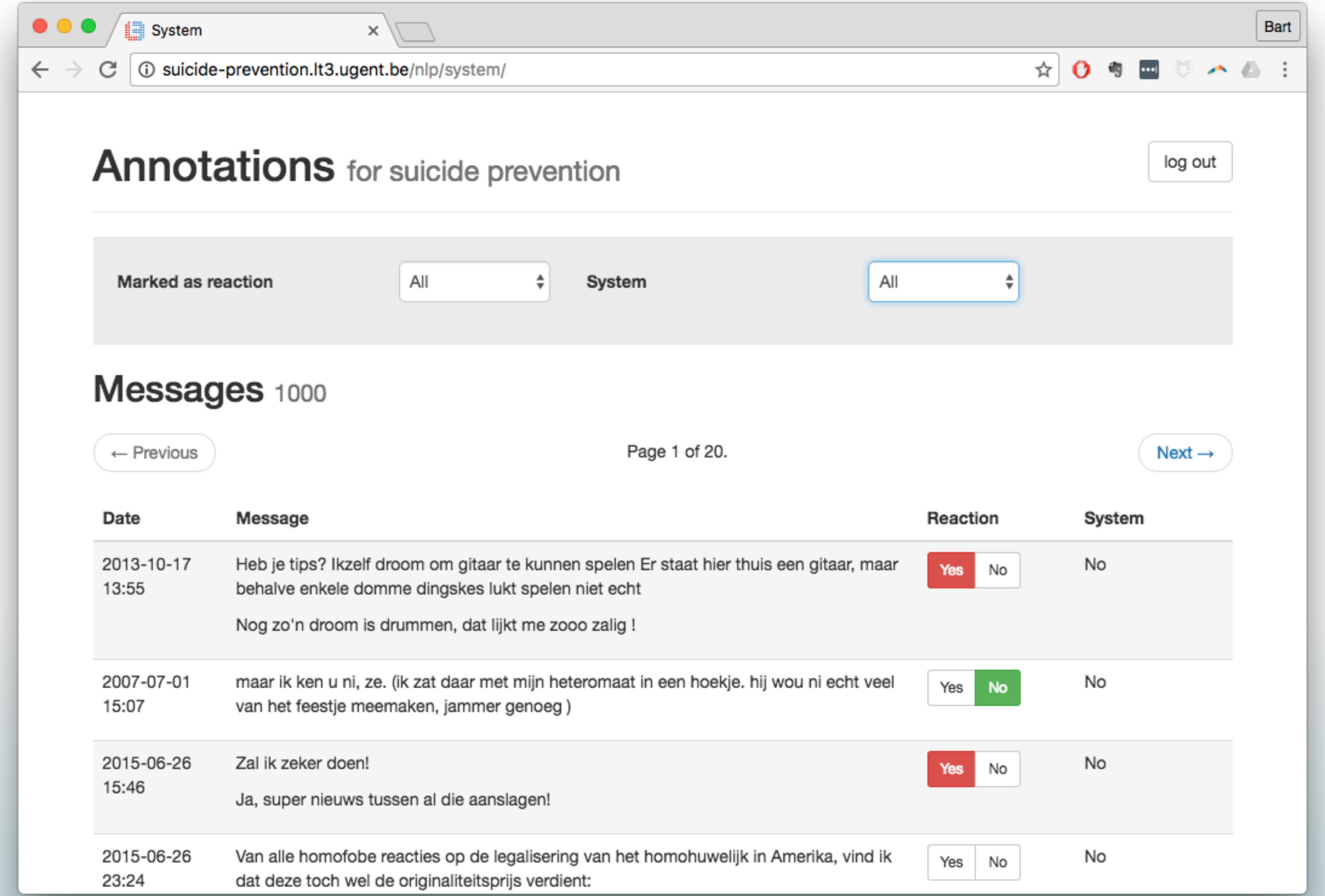
Does it work in practice?

What is the impact of the automatic detection system in a moderator setting?

Simulation of high work load of moderators:

- task: identify alarming messages that need a response (75)
- Lots of messages (1000)
- Limited moderation time (1 hour)
- Collaboration with CPZ (Flemish centre for suicide prevention) and moderators of the website “Wel Jong Niet Hetero” (LGBT web site)
- 1 group with / 1 group without system aid

Valorisation: interface



Annotations for suicide prevention

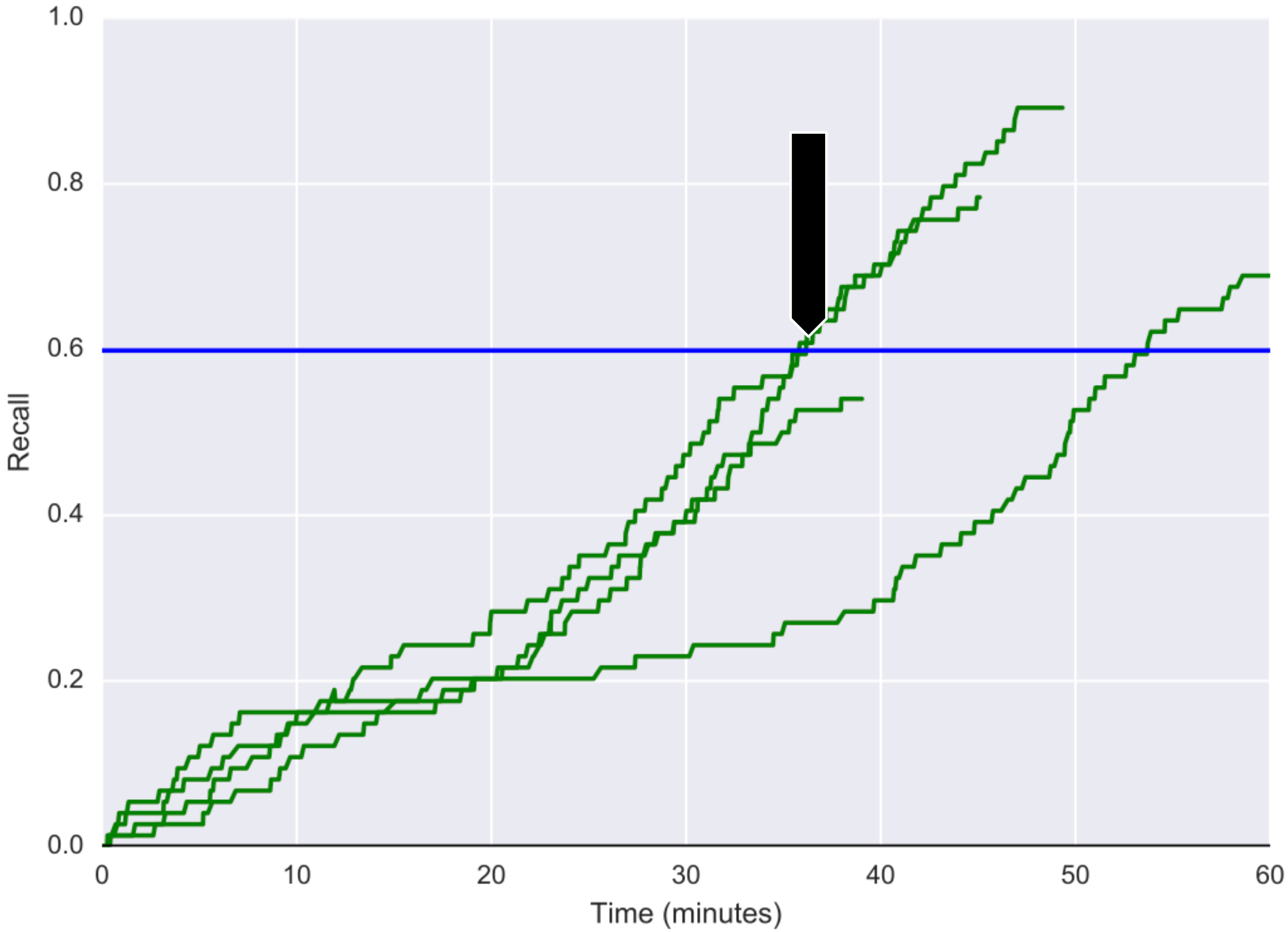
Marked as reaction All

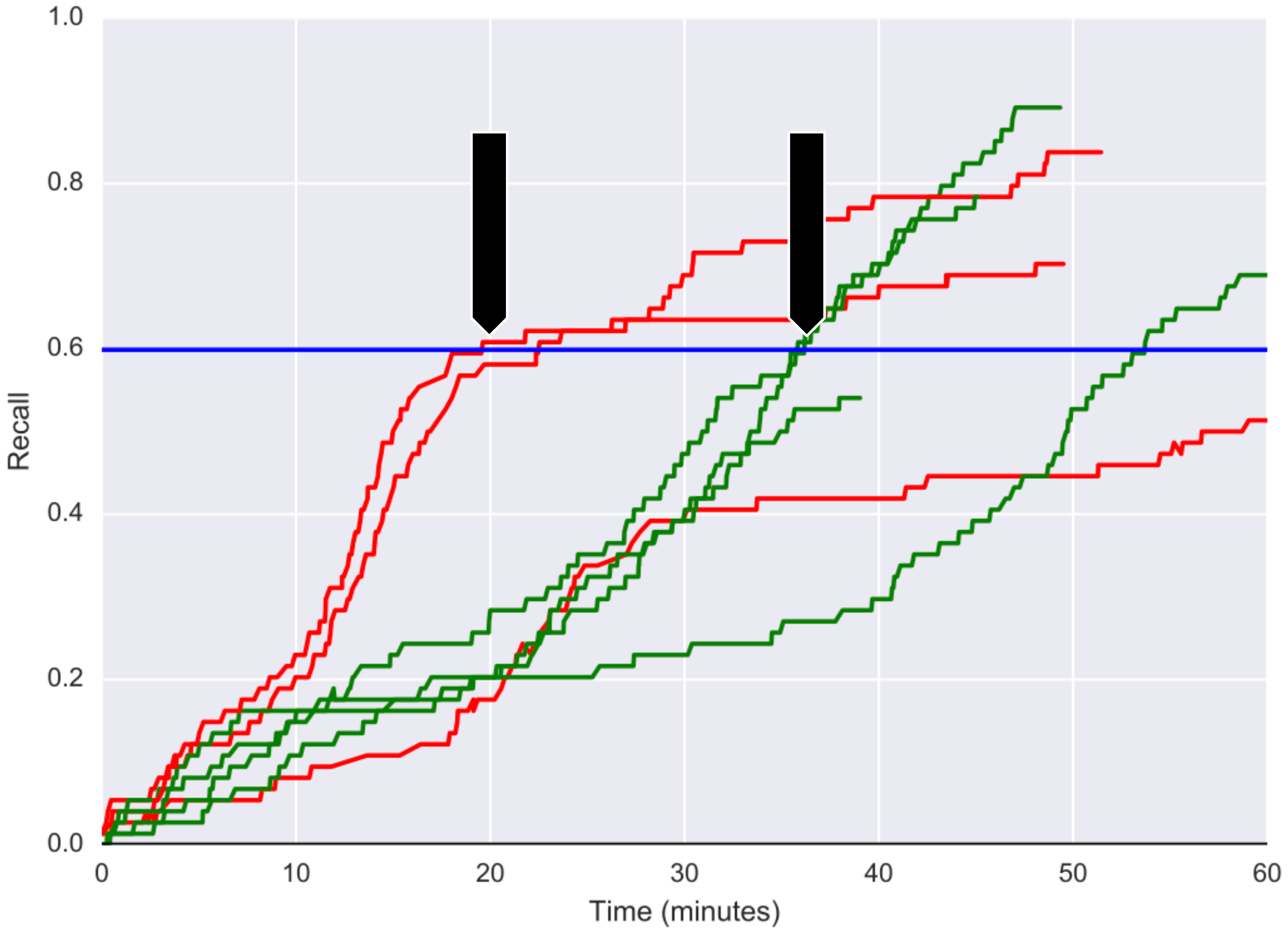
System All

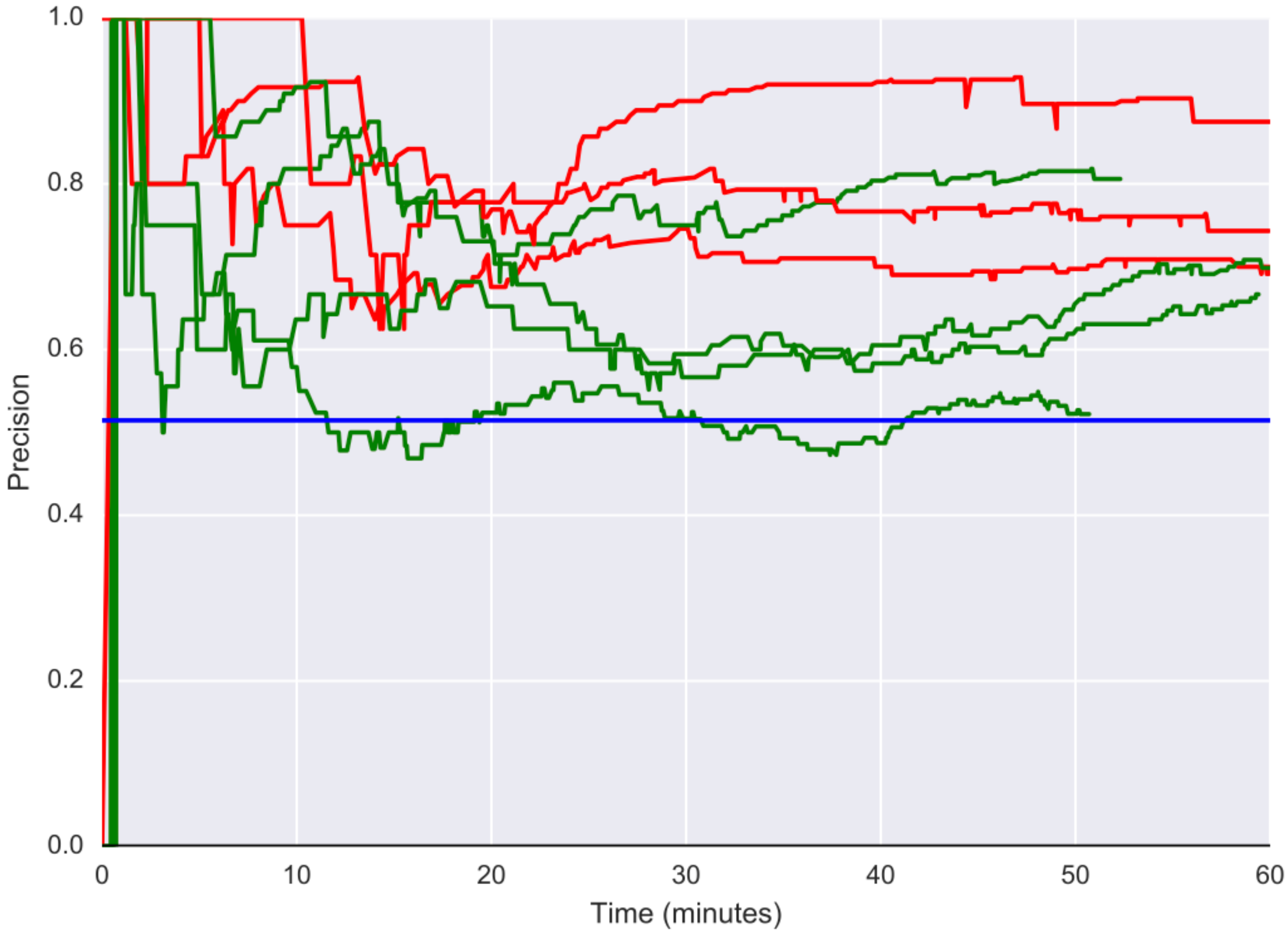
Messages 1000

Page 1 of 20.

Date	Message	Reaction	System
2013-10-17 13:55	Heb je tips? Ikzelf droom om gitaar te kunnen spelen Er staat hier thuis een gitaar, maar behalve enkele domme dingskes lukt spelen niet echt Nog zo'n droom is drummen, dat lijkt me zooo zalig !	[✓ Yes] No	No
2007-07-01 15:07	maar ik ken u ni, ze. (ik zat daar met mijn heteromaat in een hoekje. hij wou ni echt veel van het feestje meemaken, jammer genoeg)	Yes [✓ No]	No
2015-06-26 15:46	Zal ik zeker doen! Ja, super nieuws tussen al die aanslagen!	[✓ Yes] No	No
2015-06-26 23:24	Van alle homofobe reacties op de legalisering van het homohuwelijk in Amerika, vind ik dat deze toch wel de originaliteitsprijs verdient:	Yes No	No


1.0
0.8
0.6
0.4
0.2
0.0
Recall
0
10
20
30
40
50
60
Time (minutes)


1.0
0.8
0.6
0.4
0.2
0.0
Recall
0
10
20
30
40
50
60
Time (minutes)

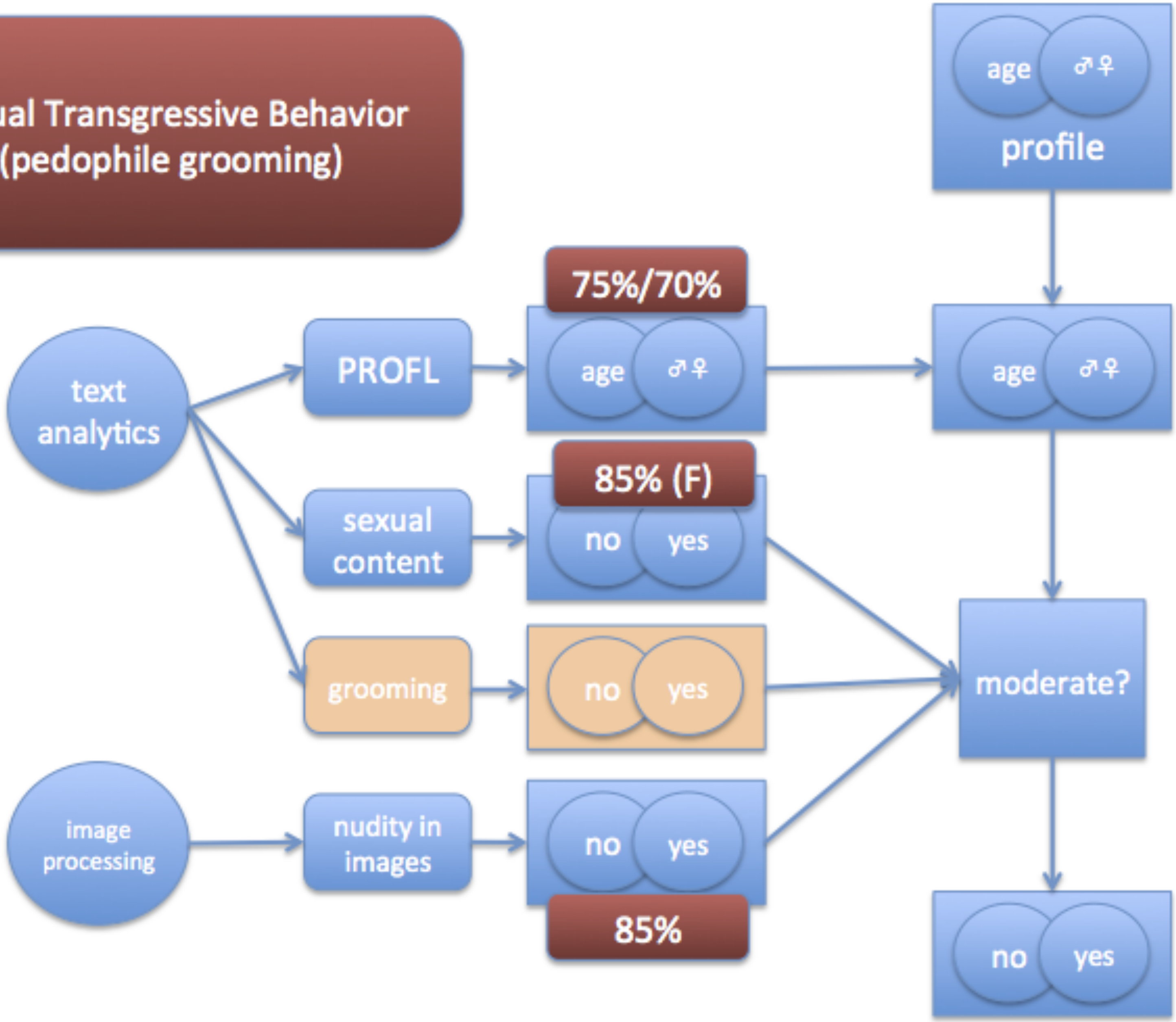

1.0
0.8
0.6
0.4
0.2
0.0
Precision
0
10
20
30
40
50
60
Time (minutes)

More info?

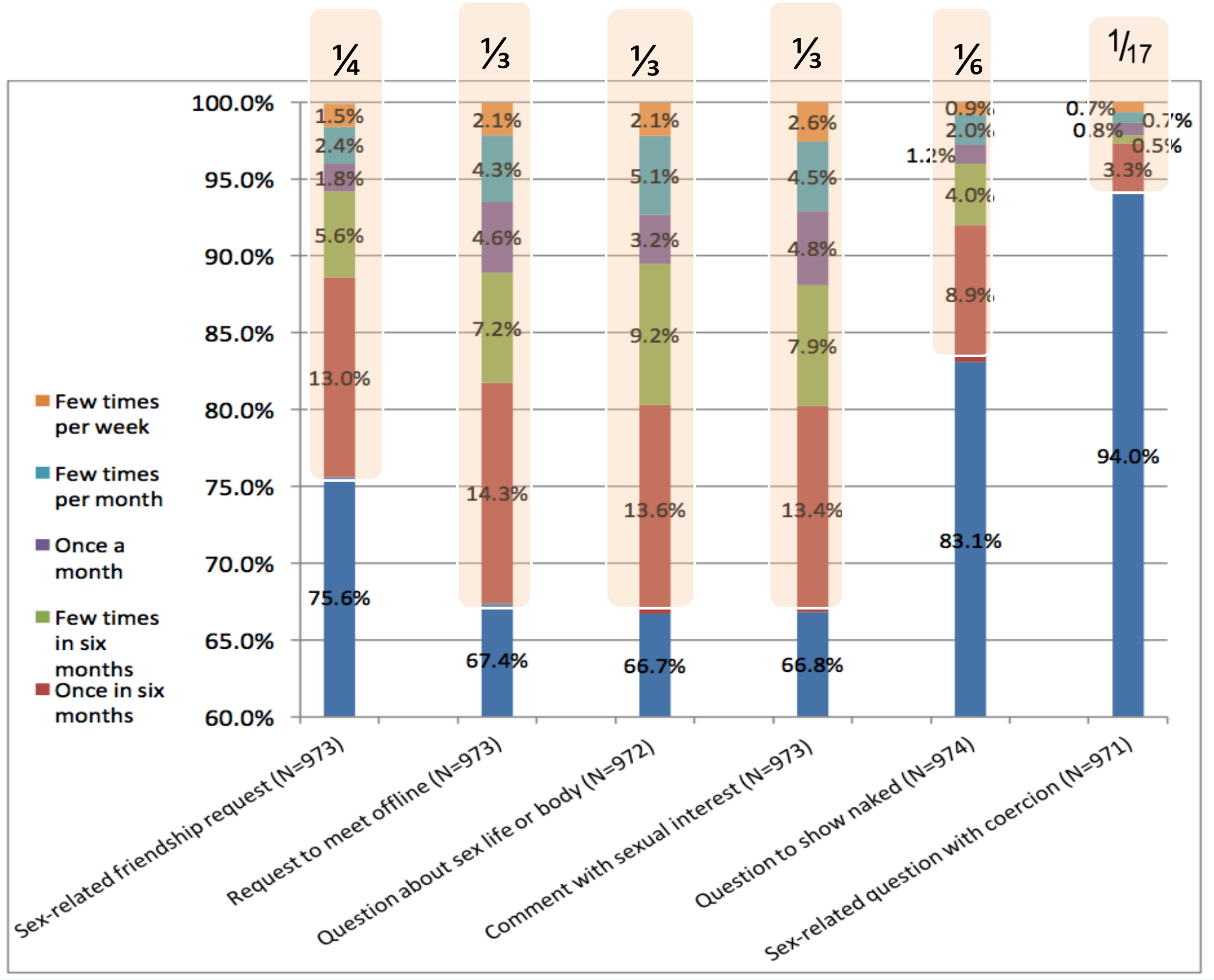
Bart Desmet: bart.desmet@ugent.be

USE CASE 3: PROFILING FOR DETECTING PEDOPHILE GROOMING

Sexual Transgressive Behavior (pedophile grooming)



Motivation



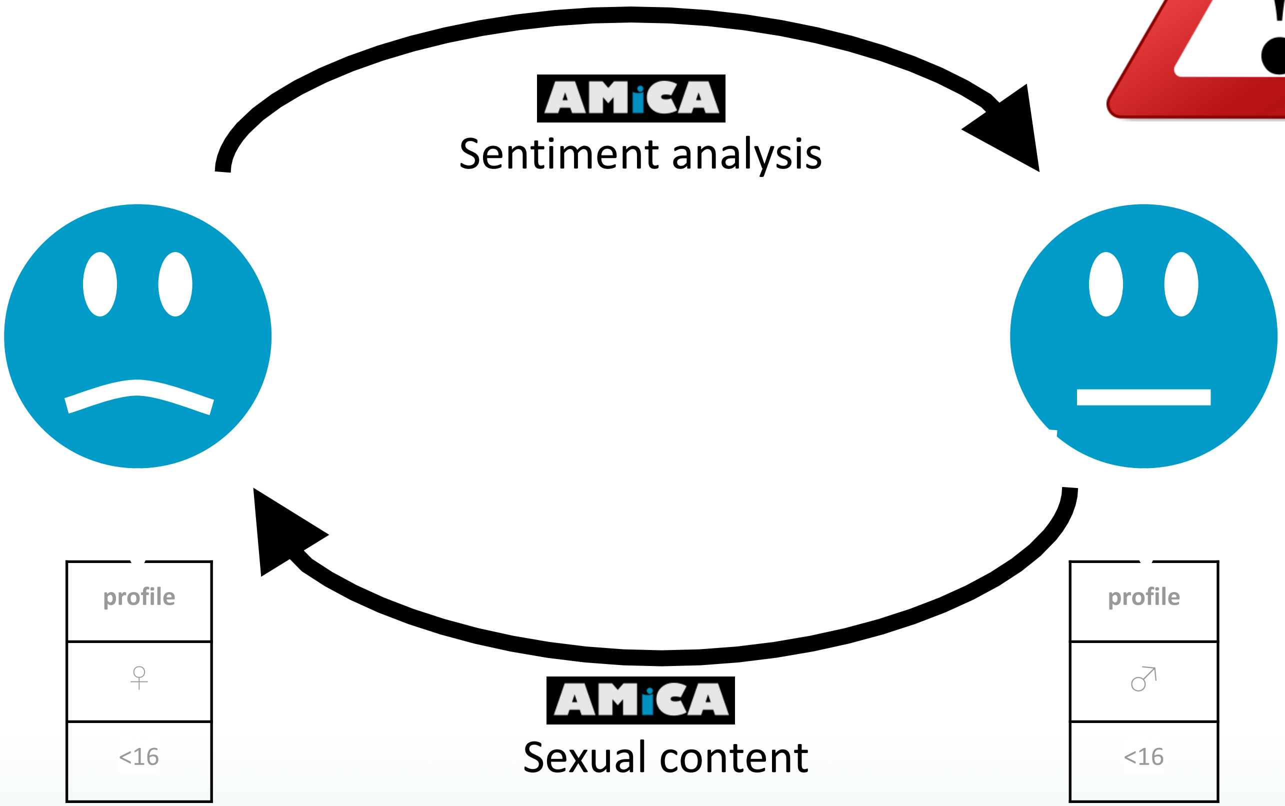
¼
⅓
⅓
⅓
⅙
1/17
100.0%
95.0%
90.0%
85.0%
80.0%
75.0%
70.0%
65.0%
60.0%
🟧 Few times per week
🟦 Few times per month
🟪 Once a month
🟩 Few times in six months
🟥 Once in six months
1.5%
2.4%
1.8%
5.6%
13.0%
75.6%
2.1%
4.3%
4.6%
7.2%
14.3%
67.4%
2.1%
5.1%
3.2%
9.2%
13.6%
66.7%
2.6%
4.5%
4.8%
7.9%
13.4%
66.8%
0.9%
2.0%
1.2%
4.0%
8.9%
83.1%
0.7%
0.8%
0.7%
0.5%
3.3%
94.0%
Sex-related friendship request (N=973)
Request to meet offline (N=973)
Question about sex life or body (N=972)
Comment with sexual interest (N=973)
Question to show naked (N=974)
Sex-related question with coercion (N=971)

Motivation

- Survey: ±1000 youngsters about the frequency, nature and appropriateness of sexual messages on social media
- Especially on Facebook
- Who?
  - 32% strangers
  - 29% friends IRL
  - 19% online friends

67% didn’t like the message + 11% reported the incident

Approach



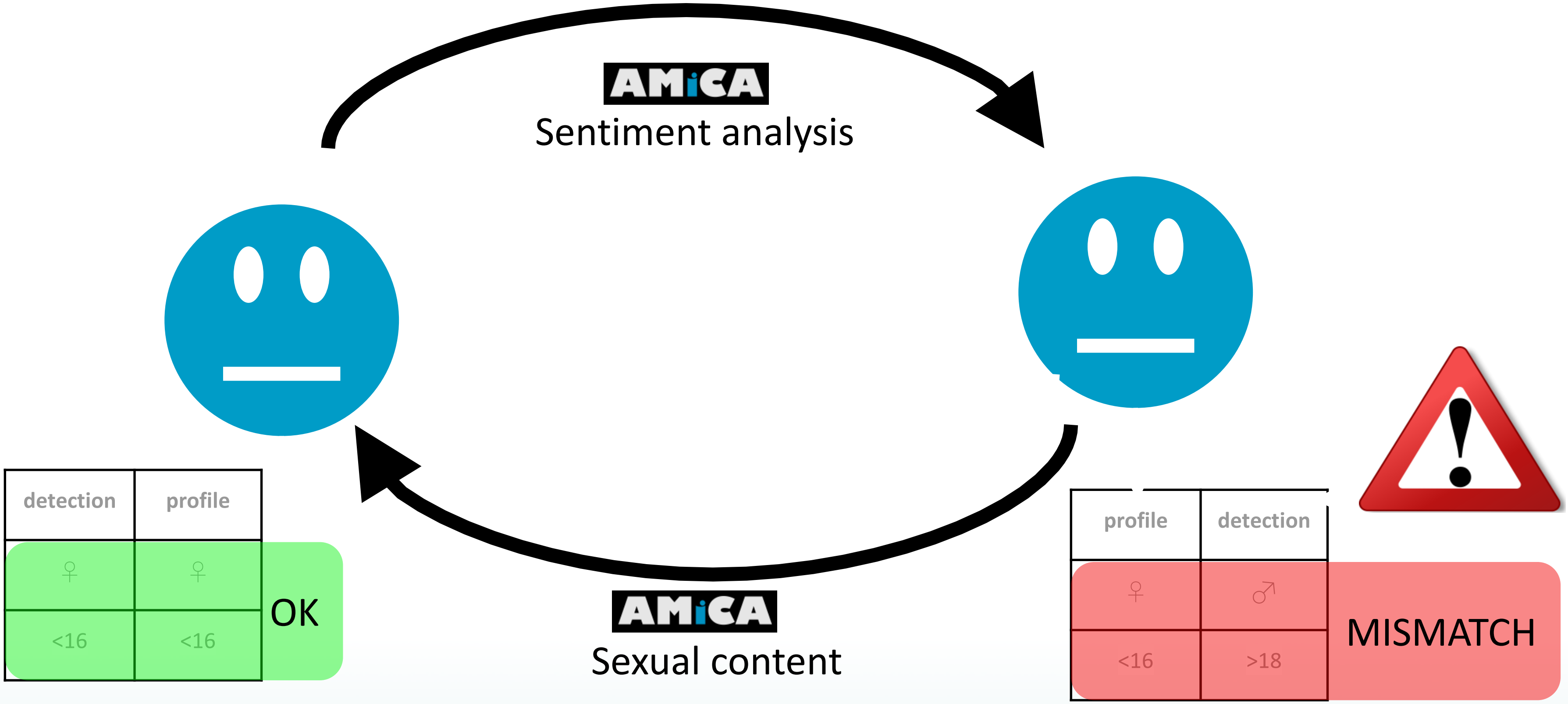
AMiCA

Sentiment analysis

AMiCA

Sexual content

Approach



AMiCA

Sentiment analysis

AMiCA

Sexual content

detection	profile
♀	♀
<16	<16

OK

profile	detection
♀	♂
<16	>18

MISMATCH

Profiling

- AMiCA profiler
  - Based on Chris Emmery's OMESA
    - https://github.com/cmry/omesa
- Age and Gender
  - Finding dubious SN profiles
    - Computed age and gender does not match given information
    - Optimizing recall (for moderator application)
    - Adapting to binary classification
      - Legally relevant age difference

Approach

- SN chat data (Netlog, 2010-2011)
  - 380k posts
  - 87k users
  - Data point = combined posts of a single user
  - Self-reported age, gender, and location
- Classes: age (binary), gender, age+gender
- 5-fold cross-validation
- SVM with linear kernel
- Features:
  - token n-grams
  - character n-grams

Results

- Gender
  - ~70%
  - Adding different types of features (LIWC, POS patterns, sentiment, etc) boosts f-scores slightly

Results

- Age:
  - Distinguish between users above and below age of consent (16 in Belgium), -16 versus +18 has priority
  - Optimize recall
    - Using cost and confidence parameters in SVMs
    - Up to 95% recall for -16; 92% recall for +18

Ref: Janneke van de Loo , Guy De Pauw, Walter Daelemans, Text-Based Age and Gender Prediction for Online Safety, International Journal of Cyber-Security and Digital Forensics (IJCSDF), 2016, 46-60.

Predator Detection

- Two classifiers
  - LiBSVM
  - Classify at the post level, aggregate at user level
  - Classify at the user level directly
    - Weighted voting of previous
  - Additional constraints
    - E.g. only one pedophile per conversation

Claudia Peersman, Frederik Vaassen, Vincent Van Asch, Walter Daelemans. Conversation Level Constraints on Pedophile Detection in Chat Rooms. CLEF 2012 (PAN), 2012.

Overall test results

- Grooming detection
  - Predator detection
    - 72 % f-score, 89% precision, 60% recall
  - Suspicious posts
    - 30% f-score, 36% precision, 26% recall

More info?

Walter Daelemans:
walter.daelemans@uantwerpen.be

Guy De Pauw:
guy.depauw@uantwerpen.be

DISCUSSION

discussion

- Is normalization and automatic detection accurate enough for applications in cybersecurity?
  - Precision - Recall trade-off
- Should we protect children and young people in social networks against their will?
  - Protection - privacy trade-off

Thank you!

Els.lefever@ugent.be

http://www.amicaproject.be/