Automatische Tekstanalyse voor Cyberpestdetectie

Gilles Jacobs, Cynthia Van Hee, Bart Desmet, Els Lefever & Véronique Hoste (LT3)
Chris Emmery, Ben Verhoeven, Guy De Pauw & Walter Daelemans (CLIPS)

INLEIDING & MOTIVATIE

- Sociale media houden risico's in voor jongeren.
- Door de gigantische hoeveelheid informatie op het web is het voor moderatoren van sociale netwerken onmogelijk om alle berichten manueel na te kijken.
- *Parent control* systemen bestaan, maar werken op basis van woordenlijsten → lage recall.

⇩

Nood aan intelligente systemen voor automatische detectie van cyberpesten om de annotatietaak voor moderatoren op sociale netwerken te verlichten.

STATS

» In 2012 was 11% van de Vlaamse jongeren recent slachtoffer van cyberpesten (Van Cleemput et al., 2013)

» Volgens het EU Kids Online Report werd in 2014 20% van de 11 tot 16-jarigen in Europa blootgesteld aan haatberichten.

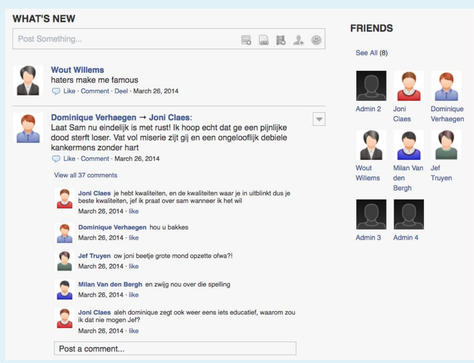
» Wereldwijd was 20 tot 40% van de jongeren ooit slachtoffer van cyberpesten (Tokunaga, 2010).

DATAVERZAMELING

Dataverzameling
- ASKfm
- Intern verzameld
- Donaties na mediacampagne

- Simulatie-experimenten

SIMULATIE-EXPERIMENTEN



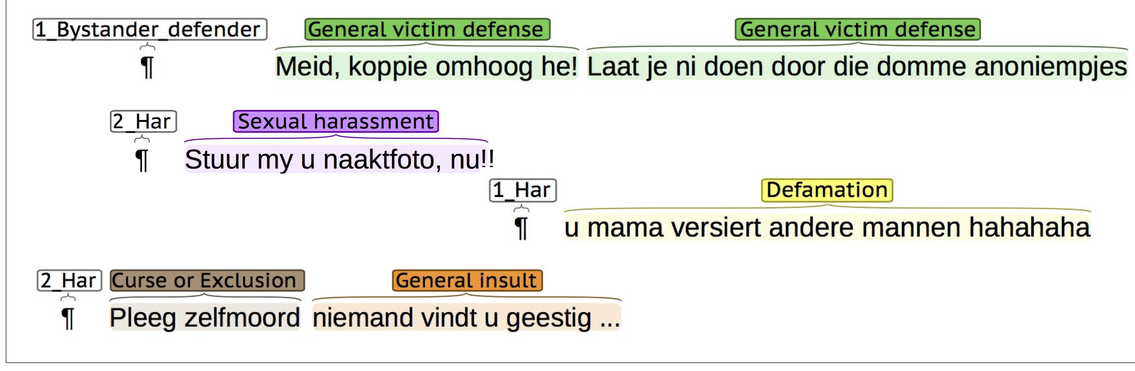
- 10 scholen
- Enquête over cyberbullying
- 2 rollenspellen in mock-up van sociaal netwerk
  - low severity
  - high severity
  - inlevingsoefening volgens karakterfiches
- afsluitende klasdiscussie
- Dataverzameling voor MiOS en CLiPS+LT3 (!)

	EN	NL
ASKfm	190.400	106.000
Simulaties	-	3.500
Intern verzameld	-	2.000
Donaties	-	368

ANNOTATIE VAN DATA

Nieuw annotatieschema
- Verschillende vormen van cyberpesten volgens 4 rollen in een cyberpestsituatie:
  → *Harasser*: *threat/blackmail, insult, curse/exclusion, defamation, sexual harassment*
  → *Victim*: *self defense*
  → *Bystander assistant: bully encouragement*
  → *Bystander defender: victim defense*
- Ernst van het pestbericht: 0 *(safe)*, 1 of 2

ANNOTATIEVOORBEELDEN



EXPERIMENTEN

ASKfm

- Hoe goed kunnen we pestberichten detecteren?
- ASKfm data (NL + EN)
- Zeer weinig gevallen van pestberichten t.o.v. niet-pestberichten
- Algoritme: Linear SVM (LIBLINEAR)
- Preprocessing: tokenisatie, stemming
- Features: woord- en lettersequenties, termenlijsten, syntactische features, polariteitsinformatie, topic models
- Optimalisatie: hyperparameters en features
- Resultaten:

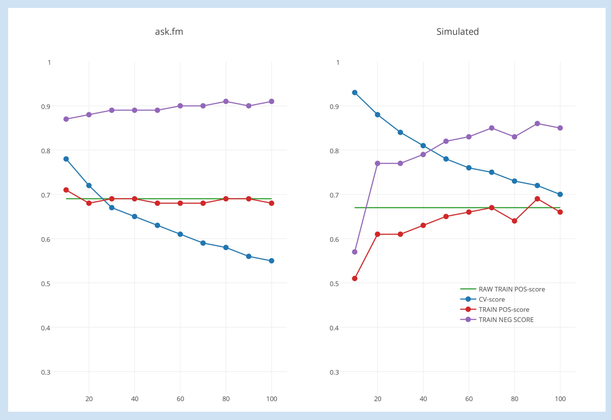
	10-fold crossvalidatie					Onafhankelijke test set				
	Accuracy	F1-score	Precision	Recall	AUC	Accuracy	F1-score	Precision	Recall	AUC
NL	96%	61%	72%	53%	76%	96%	64%	74%	57%	78%
EN	97%	64%	74%	56%	78%	97%	62%	74%	54%	76%

Simulaties

- levert gesimuleerde data werkbaar materiaal op voor de classificatie van echte cyberbully-events?
- hoeveel (gesimuleerde) data hebben we nodig?

learning-curve experimenten met OMESA

- SVM-linear trainen op verschillende databronnen, groottes en proporties pos|neg
- evaluatie dataset-intern (CV-score) en op gedoneerde data (Raw Train POS score)
- F1-score op gedoneerde data:
  - ask.fm = 69%
  - gesimuleerde data = 66%
  - ask.fm + gesimuleerde = 73%



CONCLUSIES & VERDER ONDERZOEK

- Detectie van cyberbullying is mogelijk maar het is ook een uitdagende taak: weinig data (zeker voor fijnmazige categorieën), impliciet taalgebruik (bv. bedreigingen), vaak context/wereldkennis nodig, inzicht in eerdere interacties, …
- Gebrek aan data kan opgevangen worden door middel van simulaties en rollenspellen
- Toekomstig onderzoek:
  → Optimalisatie-experimenten fijnmazige **categorieën**
  → Systeem testen op gedoneerde testdata
  → Reconstructie van een cyberpestsituatie m.b.v. **roldetectie**

REFERENTIES
Tokunaga R.S. 2010. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Computers in Human Behavior.* 26(3):277–287.
Van Cleemput K, Bastiaensens S, Vandebosch H, Poels K, Deboutte G, DeSmet, A, et al. Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen (White Paper). Universiteit Antwerpen & Universiteit Gent, 2013.