language and translation technology team

SUÏCIDALITEITSDETECTIE IN ONLINE TEKSTBERICHTEN

BART DESMET & VÉRONIQUE HOSTE

INLEIDING & MOTIVATIE

- Online platformen blijken een **toegankelijk medium** om suïcidale gedachten te uiten[1, 2]
- Moderatoren kunnen **niet alle berichten manueel** nakijken
- Voor goede preventie (op tijd en gepast): nood aan **automatische moderatorondersteuning**
- Filteren met **zoektermen** is problematisch[3]:
  - veel irrelevante hits (lage precisie)
  - enkel expliciete berichten (lage recall)
- De aanpak in AMiCA: elk bericht analyseren en **classificeren met een zelflerend model**

AUTOMATISCH DETECTIESYSTEEM

Detectietaken

Op basis van het annotatieschema werden **2 taken** gedefinieerd:

1. **Relevantie**: detectie van berichten over zelfdoding
2. **Ernst**: detectie van berichten met een ernstige suïcidale dreiging

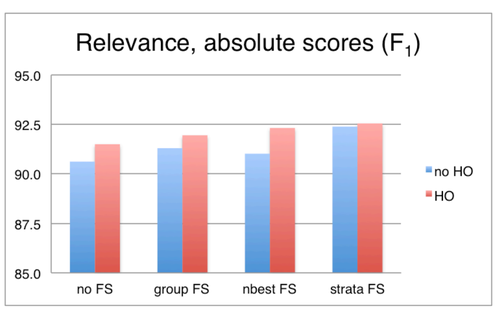
Experimentele opzet

- Model: SVM (linear, polynomial & sigmoid kernels)
- Features: woord- en lettersequenties, topic models, termenlijsten, polaritieitslexicons, capitalisatie en named entity-informatie
- Featureselectie en hyperparameteroptimalisatie met genetisch algoritme
- 10-fold crossvalidatie op trainingscorpus
- Foutenanalyse en schaalbaarheidexperimenten op achtergrondcorpus

Resultaten

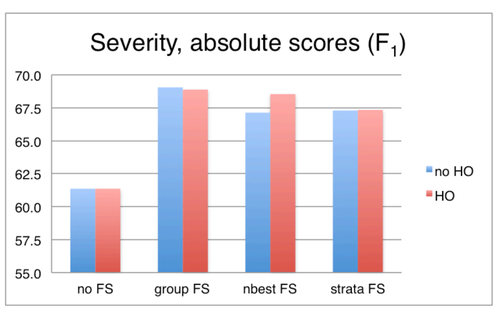
Relevantie:

- goede performantie: F1-score van 92.55%
- precisie en recall in balans



Ernst:

- moeilijkere taak: F1-score van 69.04%
- vooral sterke precisie (ca. 80%)



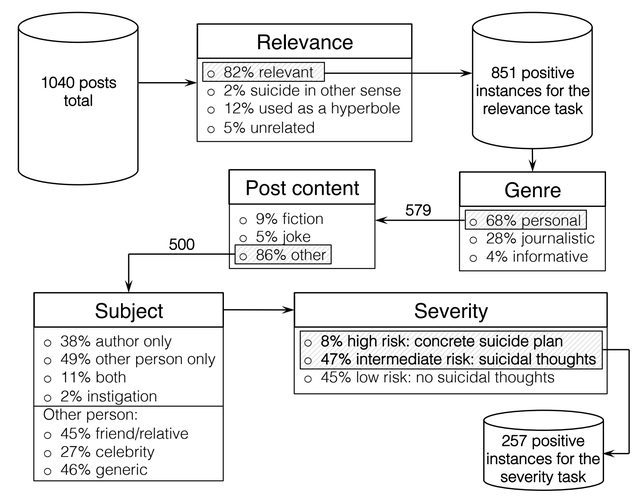
DATA

Corpus

- Forum- en blogberichten van sociale netwerksite Netlog
- 300.000 in achtergrondcorpus
- 10.000 in trainingscorpus, waarvan 1040 over suïcide

Annotatie

- Schema ontwikkeld in overleg met Centrum ter Preventie van Zelfdoding
- Corpus geannoteerd door staf en vrijwilligers van CPZ



EVALUATIE

Schaalbaarheid

Beide systemen **schalen goed** naar datasets met **grote *skew***

- nauwelijks valse positieven
- valse negatieven bij impliciet taalgebruik
- systeem bruikbaar in realistische online setting

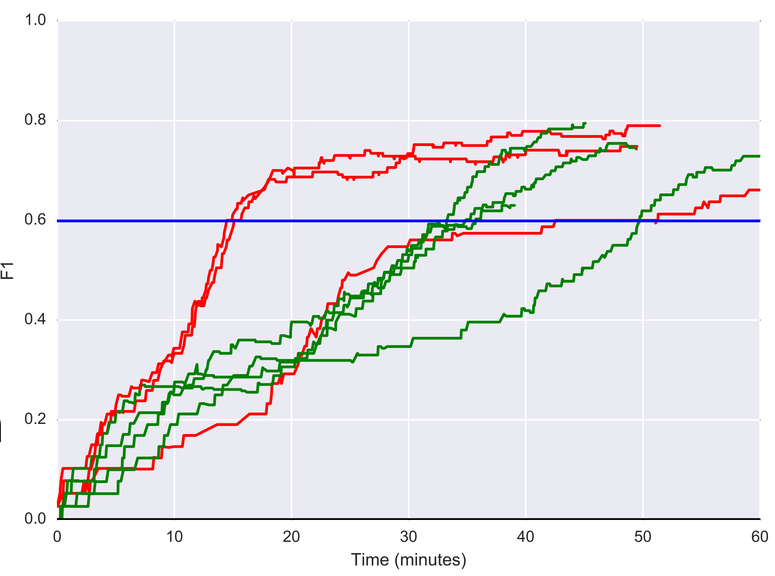
Valorisatie

Experiment: **werken moderatoren sneller en preciezer met systeemhulp?**

- 7 moderatoren (3 met, 4 zonder hulp van systeem)
- 1000 forumberichten, 75 suïcidaal
- Taak: *duid suïcidale berichten aan waarop een reactie nodig is*
- 60 minuten, evaluatie doorheen de tijd

Resultaten:

- bij beperkte tijd (20 min) zorgt systeem voor verdubbeling van F1-score
- effect zou sterker zijn bij grotere skew
- individuele verschillen tussen moderatoren



UNIVERSITEIT GENT
AGENTSCHAP INNOVEREN & ONDERNEMEN
AMiCA

1. Luxton, D. D., June, J. D., & Kinn, J. T. (2011). Technology-based suicide prevention: current applications and future directions. *Telemedicine and E-Health*, 17(1), 50–54.
2. Peter, J., Valkenburg, P. M., & Schouten, A. P. (2005). Developing a model of adolescent friendship formation on the internet. *Cyberpsychology & Behavior*, 8(5), 423–30.
3. Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking Suicide Risk Factors Through Twitter in the US. *Crisis*, 35(1), 51–59.