Data Set Construction and Exploratory Experiments for Cyberbullying Detection

Cynthia Van Hee, Ben Verhoeven, Els Lefever,
Guy De Pauw, Véronique Hoste, Walter Daelemans

LT3, Ghent University

CLiPS, University of Antwerp

ATILA Research Meeting
Ghent, 20 November 2014

Introduction

Youngsters online

- Mostly safe
- Risks!

Potentially harmful situations on social networks

- Suicidal behaviour
- Sexual harassment (e.g. grooming by paedophiles)
- Cyberbullying

Introduction

Protection

Several initiatives to protect children by prevention, follow-up and curation

- Regional: CPZ, Friendly Attac, Pest@pen,...
- National: FCCU,...
- European: iCOP, Child Focus,...

But!

Massive information overload $\rightarrow$ need for automatic systems

Cyberbullying

An aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly or over time against a victim who cannot easily defend him or herself. [1]

Under investigation

- Flaming
- Harassment
- Defamation
- ...

Other forms

- Masquerading
- Hate pages
- ...

[1] Slonje and Smith (2008)

Cyberbullying

Typical roles

- Victim
- Harasser or perpetrator
- Bystander
  - Bystander-defender
  - Bystander-assisant

In Belgium

- 43.6% of youngsters were victim once
- 55.1% of victims did not report it

Dataset Construction

Data collection

- Dutch corpus
- 110,615 messages

Different ways of collection

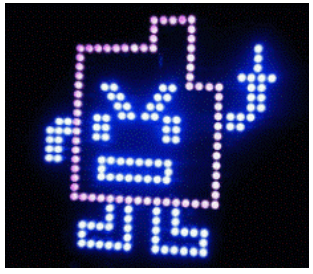
1. Ask.fm crawling
2. Manually found Netlog and Facebook posts
3. Data donation media campaign
4. Cyberbullying simulations

Press Attention

Universiteiten willen software die waarschuwt voor cyberpestgedrag

26/11/2013 - 15:32

Medewerkers van het onderzoeksproject AMiCA van de universiteiten van Antwerpen, Gent en Leuven doen een oproep om digitale pestberichten te verzamelen voor wetenschappelijk onderzoek. Op termijn wil men software ontwikkelen die waarschuwt voor pest of zelfbeschadigend gedrag.



Cyberpesten

Wie e-mails, sms'jes, chatgesprekken of berichten via socialenetwerksites in zijn bezit heeft waaruit pestgedrag blijkt, kan deze doorsturen naar AMiCA (Automatic Monitoring for Cyberspace Applications), een onderzoeksproject van de universiteiten van Antwerpen, Gent en Leuven. AMiCA wil deze data gebruiken voor wetenschappelijk onderzoek.

Taaltechnologie

"Met het project willen we algoritmes ontwikkelen om in tekst en beeld op zoek te gaan naar seksueel grensoverschrijdend gedrag zoals pedofilie, automutilatie of zelfmoordgedrag, alsook cyberpestgedrag," zegt hoogleraar taaltechnologie Véronique Hoste (UGent). "Over cyberpestgedrag hebben we nog veel te weinig materiaal. Hoe meer data, hoe fijnmaziger we onze systemen kunnen maken. We willen op sleutelwoorden zoeken, maar ook op uitingen van sarcasme, die via taaltechnologie vooralsnog moeilijker te ontdekken zijn."

Met medewerking van Microsoft en Netlog

De bedoeling is om over een viertal jaar te komen tot prototype-software, die gebruikt kan worden binnen socialenetwerksites of zogenaamde parental controlsoftware. Er wordt in dat verband samengewerkt met een bedrijvenconsortium, waaronder softwareleverancier Microsoft en de socialenetwerksite Netlog. Ook onder meer de Federal Computer Crime Unit, Sensoa en Child Focus zijn betrokken.

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 ni

> dikzak ik ga u morge kapot slage en u zus gooi ik derbove ik maak u kappot

Enkele haatcommentaren op Ask.fm. De onderzoekers zullen het taalgebruik in pestberichten analyseren. Foto: rr

OPROEP

Wetenschappers verzamelen pestberichten

Medewerkers van het onderzoeksproject AMiCA (Automatic Monitoring for Cyberspace Applications) van de universiteiten van Antwerpen, Gent en Leuven verzamelen digitale pestberichten voor wetenschappelijk onderzoek. Op termijn wil men software ontwikkelen die waarschuwt voor pest- of zelfbeschadigend gedrag. Wie sms'jes, chatgesprekken, e-mails of berichten in zijn bezit heeft waaruit pestgedrag blijkt, kan die doorsturen naar het AMiCA. «Met het project willen we algoritmes ontwikkelen om in teksten en beelden op zoek te gaan naar seksueel grensoverschrijdend gedrag, zoals pedofilie, maar ook naar automutilatie, zelfmoordgedrag of pesterijen», klinkt het. *(JM)*

Meer informatie: www.amicaproject.be

Universiteit Gent

Universiteit Antwerpen

AMiCA

Cyberbullying Simulations

Dataset Construction

Source	**Number of posts**
Ask.fm, Netlog, Facebook	106,418
Donated data	367
Simulations	3,830
Total Dutch	**110,615**

Dataset Construction

Data Annotation

Two levels of annotation

- Message
  - Harmfulness score (0-1-2)
  - Role of author:
    - Harasser
    - Victim
    - Bystander-defender
    - Bystander-assistant
- Text span
  - Fine-grained categories related to cyberbullying (e.g. threats, insults)

Data Annotation

Text span categories

- Threat or Blackmail
- Insult
- Curse or Exclusion
- Defamation
- Sexual Talk
- Defense
- Encouragement
- Sarcasm
- Other

* Most of these had a number of subcategories

Data Annotation

Annotations

- 110,615 posts were considered
- in 8,790 of them at least one categorie was annotated (7.9%)

Inter-annotator agreement

! Very skewed distribution of annotations

- Use of Gwet's AC1 score
  - Similar to Cohen's Kappa: taking class distributions into account
  - But more robust for skewed distributions

Data Annotation

Inter-annotator agreement

On all instances

- Bully event -vs- non-bully event: 96% - AC1: 0.96
- Author roles: 96% - AC1: 0.95
- Categories: $>$ 97% - AC1: 0.96

Only on instances annotated by at least one annotator

- Bully event -vs- non-bully event: 80% - AC1: 0.74
- Author roles: $>$ 79% - AC1: 0.76
- Categories: $>$ 86% - AC1: 0.79
  with most of them $>$ AC1: 0.90

Data Annotation Examples

1 ¶ 2_Har: Threat or Blackmail — Ik maak u kapot.

1 ¶ 1_Vic: Assertive self-Defense — Vind je jezelf nu beter dan mij nu je dit allemaal zegt? Curs / AssDef — Zoek een leven

1 ¶ 1_Har: General insult — ge zijt fucking dik

1 ¶ 2_Har: Curse or Exclusion — Pleeg gew zelfmoord, General insult — iedereen haat u.

1 ¶ 2_Har: Sexual harassment — Post nu gew een naaktfoto van jezelf!!

1 ¶ 1_Bystander_defender: GenIn — Ptn Amelie heeft gn konijnentanden; General victim defense — kijkt eerst naar u eigen GenIn — lelijke!! :D; General victim defense — Amelie is echt een kei toffe en lieve!; Good characteristics

Experimental Setup

Focus on:

1. Binary classification: cyberbullying event -vs- non-cyberbullying event (ratio ~1:11)
2. Five-way classification into fine-grained categories:
   - Threats (ratio ~1:371)
   - Sexual Posts (ratio ~1:180)
   - Exclusions (ratio ~1:71)
   - Defenses (ratio ~1:32)
   - Insults (ratio ~1:16)

Approach

- SVM-approach (linear kernel, c=1)
- 10-fold cross-validation
- Pattern [2]

[2] De Smedt and Daelemans (2012)

Experimental Setup

Lexical features

- **Token n-gram features:** word token uni- and bigrams
- **Character n-gram features:** character tri- and fourgrams (within tokens)
- **Token skip-n-gram features:** 2-, 3-, and 4-skip-bigrams
- **Features based on existing sentiment lexicons** [3]
  - number of positive, negative and neutral lexicon words averaged over text length
  - overall polarity (i.e. the sum of the values of identified sentiment words)

[3] Jijkoun and Hofmann (2009); De Smedt and Daelemans (2012b)

Results

Binary classification

Category	Baseline (w1gr)	Word n-grams	Filtered word n-grams	Word+char n-grams	Word+char +skip n-grams	Word+char +sentiment
Bully Event	47.27	48	34.19	53.18	50.43	54.71

Five-way classification

Category	Baseline (w1gr)	Word n-grams	Filtered word n-grams	Word+char n-grams	Word+char +skip n-grams	Word+char +sentiment
Threats	4.62	4.57	1.34	18.11	19.27	20.51
Sexual Posts	12.47	10.37	3.76	30.99	29.12	32.50
Insults	45.57	45.24	29.14	52.60	50.58	55.07
Exclusions	17	17.98	7.32	28.55	27.92	32.13
Defenses	19.77	23.29	11.89	30.65	30.46	32.64

(F-scores)

Conclusions and Future Work

Main Insights

- Binary classification obtains acceptable results (F= 54.71)
- Fine-grained classification is a harder task (F-scores between 20.51 and 55.07)
- Filtering based on PoS-tags does not improve classification performance (removal of *bitch*, *you*, *haterss*,...)

Further Research

- Feature selection
- Reducing data skewness
- Including more advanced features
- Normalization

Questions?

References

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13:2063–2067.

De Smedt, T. and Daelemans, W. (2012b). "vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for dutch adjectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3568–3572.

Jijkoun, V. and Hofmann, K. (2009). Generating a non-English subjectivity lexicon: relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–405.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Slonje, R. and Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, 49(2):147–154.