DOMAIN ADAPTATION OF SIMULATED DATA FOR CYBERBULLYING RESEARCH

16 October, 2015

Chris Emmery, Ben Verhoeven,
Guy De Pauw, Walter Daelemans

Universiteit Antwerpen

CLiPS

INTRODUCTION

- Cyberbullying detection & AMiCA.
- Public data is scarce.
- Social application; contents are sensitive.

CYBERBULLYING

- Increase of access and mobility amongst youngsters.
- More intrusive, bigger platform.
- Classic scenario, novel forms.

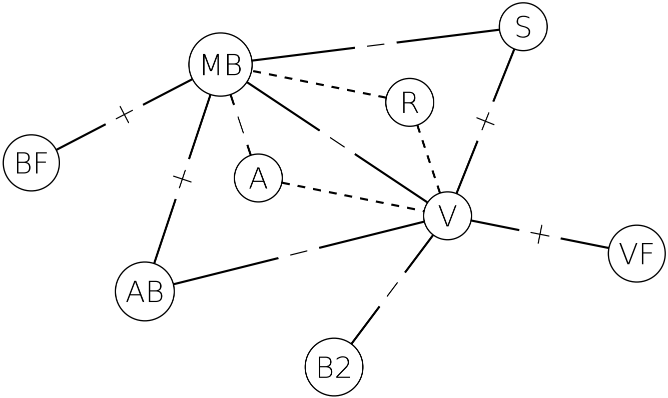


Figure: Bullying Role Graph.

PREVIOUS RESEARCH

Three categories:

- Binary: is this bullying Y/N?
- Fine-grained: role labels, different types of messages.
- Macro: meta-data (profile), network, and image information.

PREVIOUS DATA

name	platform	pos	neg
CAW 2.0	Kongregate	42	4802
CAW 2.0	Slashdot	60	4303
CAW 2.0	Myspace	65	1946
DiYT	YouTube	2277	4500
SanTwi	Twitter	300	160
XuTREC	Twitter	684	1762
KForm	Formspring	369	3915
DadvI	Myspace	311	8938
DadvY	YouTube	449	4177
BretT	Twitter	220	5162
BretTS	Twitter	194	2599

Table: Available cyberbullying datasets.

PREVIOUS PERFORMANCE

paper	data	scores
Yin2000	CAW 2.0	$F = .442$
Ptaszynski2010	OPJSSS	$F = .882$
Dinakar2011	DiYT	$Acc = .667$
Reynolds2011	KForm	$Acc = .674$
Sanchez2011	SanTwi	$Acc = .673$
Xu2012	XuTREC	$F = .770$
Kontostathis2013	KForm	$F = .570$
Dadvar2013	DadvI	$F = .350$
Nahar2013	CAW 2.0	$F = .920$
Dadvar2013	DavY	$F = .640$
Bretschneider2014	BretT	$F = \mathbf{.726}$
VanHee2015	AMiCA	$F = .554$

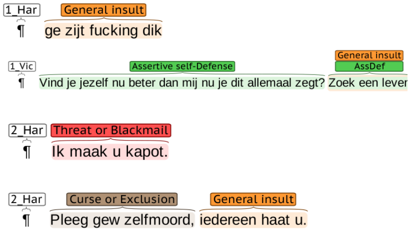
Table: Overview of scores per publication.

AMICA DATA

- Ask.fm - Q&A only, anonymity.
- Simulated - role-play on SocialEngine by 200 adolescents (14-18).

ANNOTATION

- Detection and Fine-Grained Classification of Cyberbullying Events - Van Hee, Lefever, Verhoeven, Mennes, Desmet, De Pauw, Daelemans & Hoste (2015).



ANNOTATION II

- Types: Threat/blackmail, insult, curse/exclusion, defamation, sexual talk, defense, encouragements.
- Roles: harasser, victim, bystander-defender, bystander-assistant.

OUR DATA

name	platform	pos	neg
AMiCA	Ask.fm	3988	88276
AMiCA	Simulated	1180	4612

DOMAIN ADAPTATION

- Compare simulated and real data.
- Not necessarily focussed on classifier performance for *testing* on different domains.
- Identify performance across different *training* sets.

RESEARCH QUESTIONS

- How does human-generated, simulated data relate to real-life instances of cyberbullying in terms of both content and classification performance?
- To what degree does simulated data offer a plausible alternative for real-life data and therefore solve the need for sensitive data?
- How can simulated data help the classification of cyberbullying content through enriching existing data?

LABEL INFORMATION

Text Category	Average Positive	
	Ask.fm	Simulated
insult	49.13	49.91
curse_exclusion	12.80	10.06
defense	25.64	29.69
sexual_post	5.70	0.27
threat_blackmail	2.35	2.86
defamation	1.87	2.36
encouragment	0.48	2.77
sarcasm	2.03	2.09
other	0.00	0.00

Table: Percentage (%) of categorical labels for each AMiCA corpus platform respectively.

EXPERIMENTAL SET-UP

- X = Ask.fm
- X_s = Simulated

	X	X_s	$X + X_s$
X	CV	tt	x
X_s	tt	CV	x
$X + X_s$	x	x	CV

- test with equal instances
- test with different POS/NEG ratios (1:1, 1:3, 1:10)

RESULTS: CV(SET)

		POS	NEG	NB	SVM
full	X	4000	88000	.195	.621
full	X_s	1000	4000	.367	.396
$1:10$	X	1000	10000	.364	.613
$1:10$	X_s	400	4000	.093	.213
$1:3$	X	4000	12000	.560	.756
$1:3$	X_s	1000	3000	.417	.449
$1:1$	X	4000	4000	.739	.834
$1:1$	X_s	1000	1000	.702	.673

RESULTS: $X \to X_S$

full	$X \to X_s$	.276	.139
$1:10$	$X \to X_s$	.115	.168
$1:3$	$X \to X_s$	.341	.341
$1:1$	$X \to X_s$	.530	.593

RESULTS: $X_S \to X$

full	$X_s \to X$	.148	.115
$1:10$	$X_s \to X$	.076	.132
$1:3$	$X_s \to X$	.448	.319
$1:1$	$X_s \to X$	.697	.694

RESULTS: $X + X_S$

full	$X + X_s$	.224	.526
$1:10$	$X + X_s$	.252	.358
$1:3$	$X + X_s$	.490	.579
$1:1$	$X + X_s$	.697	.736

RESULTS: EQUAL FREQUENCIES ACROSS SETS

$1:10$	X	1000	1000	.364	.550
$1:10$	X_s	1000	1000	.093	.213
$1:3$	X	1000	1000	.539	.698
$1:3$	X_s	1000	1000	.417	.438
$1:1$	X	1000	1000	.739	.786
$1:1$	X_s	1000	1000	.701	.681

PRELIMINARY CONCLUSION

- Results are overall not that surprising!
- $X_s < X + X_s < X$.
- Some hints of over and under-fitting.

$1:1$	X_s	$.673$
$1:1$	$X_s \rightarrow X$	$.694$

FUTURE WORK

- Try to use $X + X_s$ to predict X and X_s respectively.
- Balance the fit across sets.
- Qualitative analysis of errors, differences in data.