Forecasting the presence and intensity of hostility on Instagram using linguistic and social features

PING LIU*, JOSHUA GUBERMAN*, LIBBY HEMPHILL+, ARON CULOTTA*

*ILLINOIS INSTITUTE OF TECHNOLOGY – CHICAGO

+UNIVERSITY OF MICHIGAN – ANN ARBOR

ICWSM 2018

Motivation

- **Goal: hostile comment forecasting**
  - A hostile comment is the one that contains harassing, threatening, or offensive language directed toward a specific individual or group

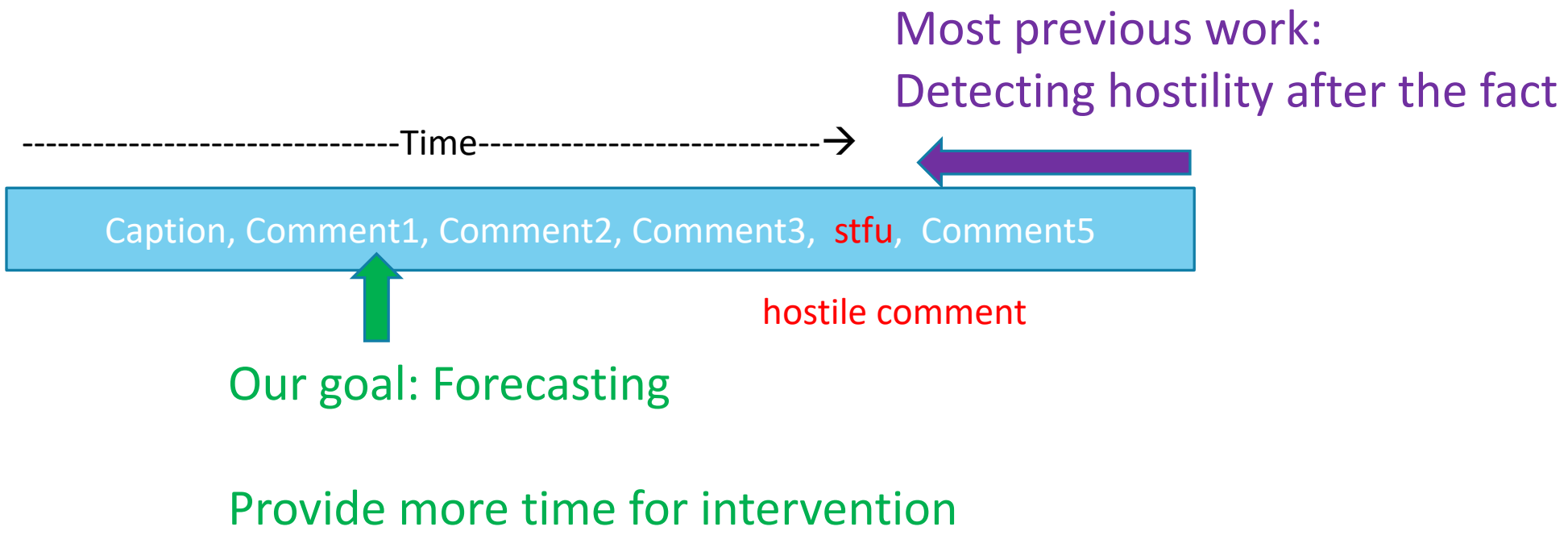
Type	Text	User	Time
Caption	Awww my Mommy Saturday😍	User1	11:30 AM
Comment1	Super cute picture 😍😍	User2	12:44 PM
Comment2	Nice 👍👍	User3	12:50 AM
Comment3	I love them two	User4	12:51 AM

Related work

- Previous related work
  - Sood, Sara Owsley, Elizabeth F. Churchill, and Judd Antin. "Automatic identification of personal insults on social news sites." *Journal of the Association for Information Science and Technology* 63.2 (2012): 270-285.
  - Geiger, R. Stuart. "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space." *Information, Communication & Society* 19.6 (2016): 787-803.
  - Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE, 2011.

- Concurrent similar work
  - Zhang, Justine, et al. "Conversations Gone Awry: Detecting Early Signs of Conversational Failure." *Proceedings of ACL 2018. To appear.*

Motivation


Most previous work:
Detecting hostility after the fact
----------Time----------→
Caption, Comment1, Comment2, Comment3, stfu, Comment5
hostile comment
Our goal: Forecasting
Provide more time for intervention

Our tasks

- **Hostility presence forecasting**
  - Given the initial sequence of non-hostile comments in a post, predict whether some future comment will be hostile
- **Hostility intensity forecasting**
  - Given the first hostile comment in a post, predict whether the post will receive more than N hostile comments in the future

Examples of task 1 - presence

	Index	Text
Observed	Caption	Pain
	Comment1	😢
	Comment2	
	Comment3	Love you guys
	Comment4	when white people think they black
Leading time T	……	……
Future hostility presence	Comment7	your a faggot honestly
	Comment8	mind your own buisness kid
	Comment9	lemme block you Latino bucket
	Comment10	lol go ahead no one stoping ya

Examples of task 2 - intensity

Observed up to 1st hostile comment

Future hostility intensity

Index	Text
Caption	Pain
Comment1	😢
Comment2	
Comment3	Love you guys
Comment4	when white people think they black
……	……
Comment7	your a faggot honestly
Comment8	mind your own buisness kid
Comment9	lemme block you Latino bucket
Comment10	lol go ahead no one stoping ya

Data processing

- **Collecting posts from Instagram - ~15M comments**
  - Post: A user-provided image and its associated comments
- **Uniform sampling**
  - Search engine
- **Annotation by AMT**

Data

	posts	comments	hostile comments
hostile posts	591	21,608	4,083
non-hostile posts	543	9,379	0
total	1,134	30,987	4,083

Hostile and non-hostile post examples

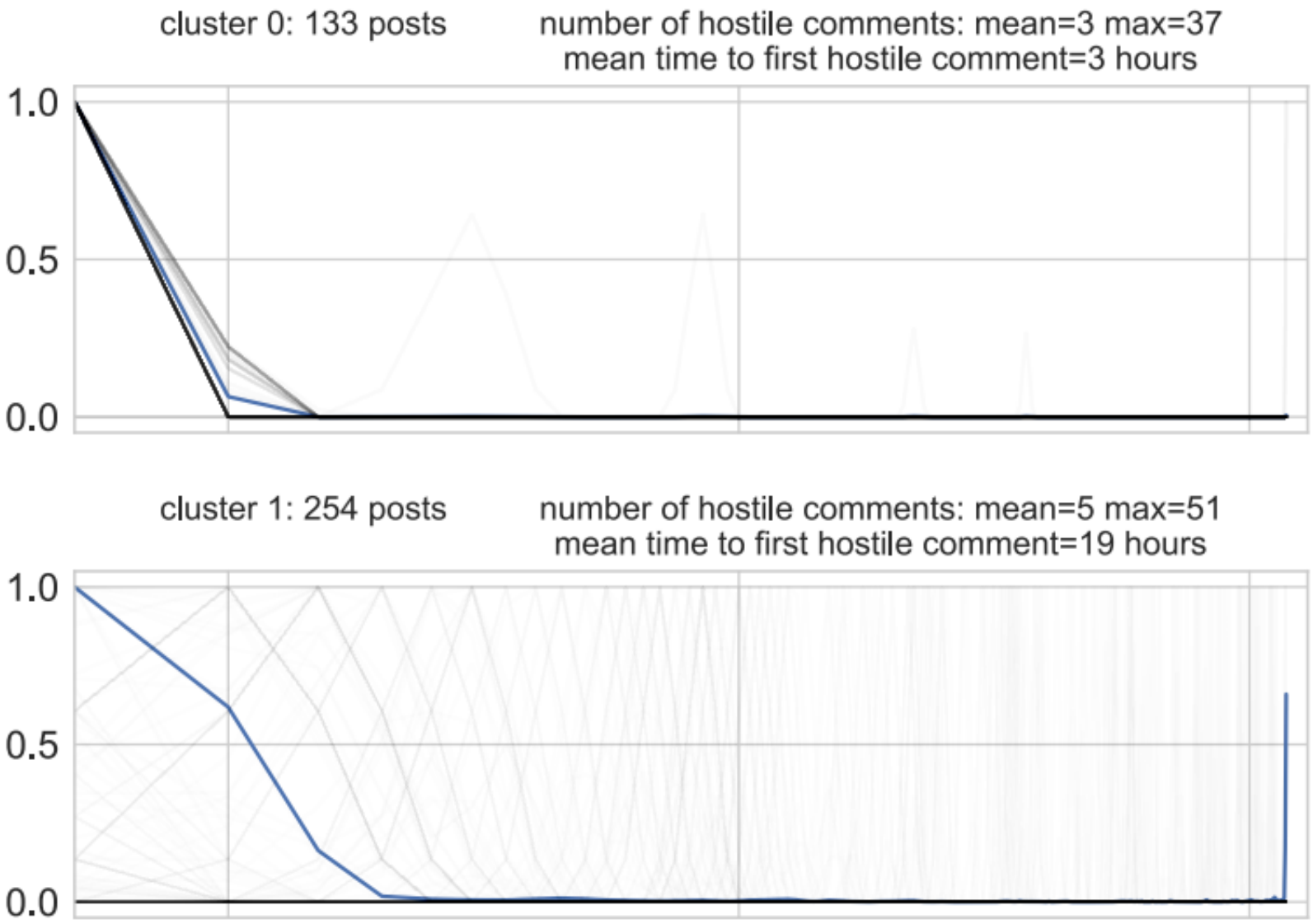
hostile

Type	Text	User	Time
Caption	Lets get it 🔫👊	User1	11:30 AM
Comment1	Bitch stfu	User2	12:44 PM
Comment2	u not my fuckin mom	User1	12:50 AM
Comment3	Let's get it then bitch	User3	12:51 AM

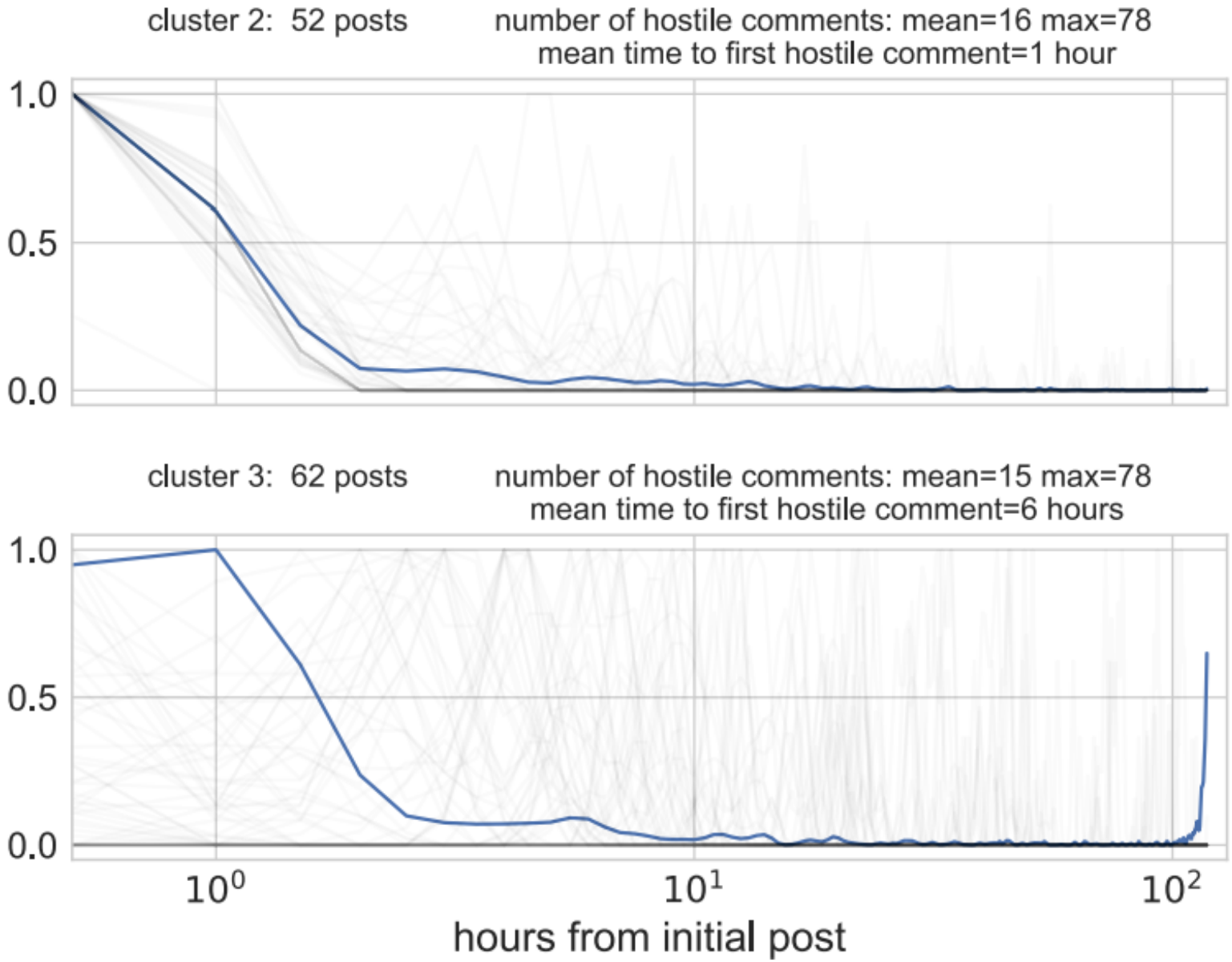
non-hostile

Type	Text	User	Time
Caption	Awww my Mommy Saturday😍	User1	8:30 AM
Comment1	Super cute picture 😍😍	User2	10:14 AM
Comment2	Nice 👍👍	User3	10:30 AM
Comment3	I love them two	User4	11:22 PM

Time series analysis
K - Spectral Centroid - Yang and Leskovec (2011)



Time series analysis
K - Spectral Centroid - Yang and Leskovec (2011)



Linguistic and Social features

- Unigram (U)
- Word2vec (w2v) - Mikolov et al. 2013
- N-gram Character word2vec (n-w2v) - Bojanowski et al. 2016
- Hatebase/ProfaneLexicon (lex)
  - www.hatebase.org
  - www.cs.cmu.edu/~biglou/resources/

Linguistic and Social features

- Final comment (final-com)
- Previous comments (prev-com) - Cheng et al. 2017
- Previous post (prev-post)
- Trend Feature (trend)
- User activity (user) - Dovidio et al. 2013

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

Unigram (U)

0	0	1	…	0	0	0

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

Word2vec (w2v)

Mikolov et al. 2013

0.24	-0.12	0.33	…	-0.78	0.40	-0.02

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

N-gram Character word2vec

Bojanowski et al. 2016

0.24	-0.12	0.33	…	-0.78	0.40	-0.02

Linguistic and Social features

Hatebase/ProfaneLexicon

www.hatebase.org

www.cs.cmu.edu/~biglou/resources/

Hatebse						ProfaneLexicon
occurrence	disability	ethnicity	gender	nationality	religion	occurrence
1	1	0	0	0	0	1

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	**your a faggot honestly**

Final comment (final-com)

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

Previous comments

Cheng et al. 2017

Their last comments from previous other posts

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

Previous post

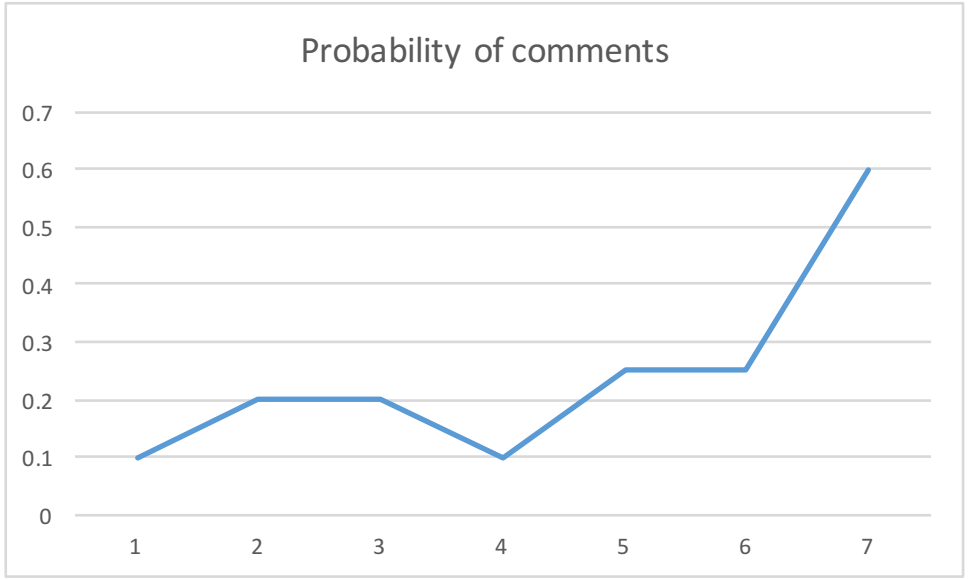
The author's the most recent post and the comments on it

Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

proba > 0.3	ratio of > 0.3	Maximum slope	max - min
1	1/7	0.35	0.5

Trend Feature



Linguistic and Social features

User	Text
author	Pain
user1	😢
user2	
author	Love you guys
user4	when white people think they black
user5	stop his friend passed away
user6	your a faggot honestly

User activity

1. Ratio of unique users in the conversation
2. Ratio of "directed" comments (those with mentions)

R unique	R mentioned
6/7	0

Our tasks

- **Hostility presence forecasting**
  - Given the initial sequence of non-hostile comments in a post, predict whether some future comment will be hostile
- **Hostility intensity forecasting**
  - Given the first hostile comment in a post, predict whether the post will receive more than N hostile comments in the future

Leading time T could vary

	Index	Text
Observed	Caption	Pain
	Comment1	😢
	Comment2	
	Comment3	Love you guys
	Comment4	when white people think they black
Leading time T	……	……
Future hostility presence	Comment7	your a faggot honestly
	Comment8	mind your own buisness kid
	Comment9	lemme block you Latino bucket
	Comment10	lol go ahead no one stoping ya

Experiments - Task 1 - Presence

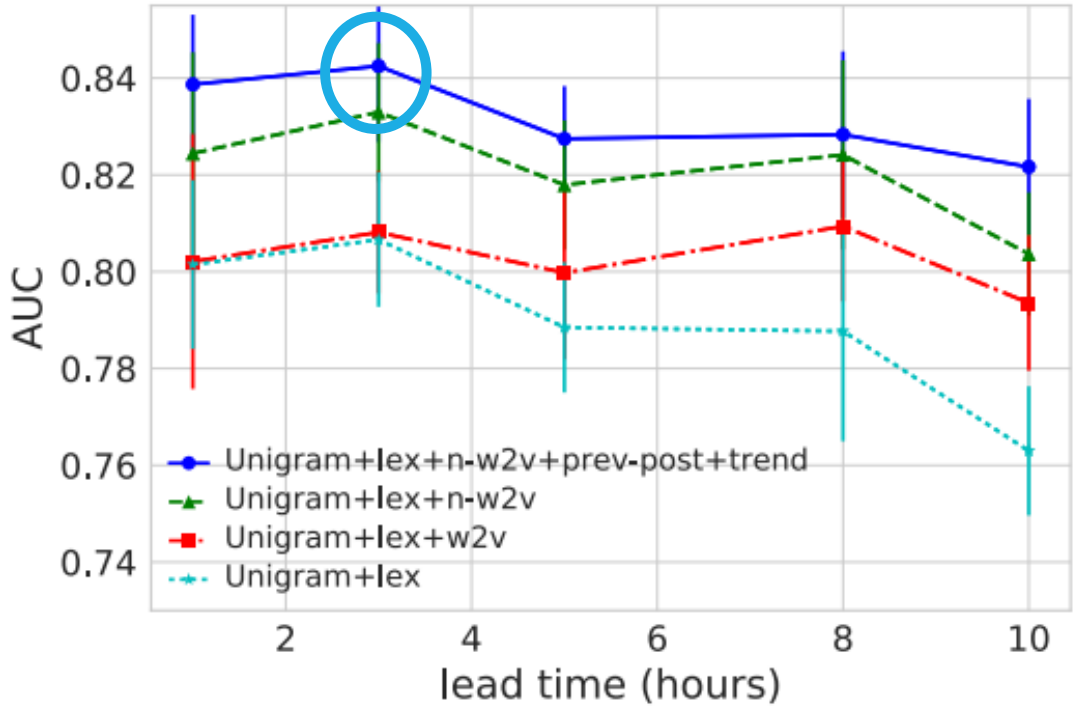


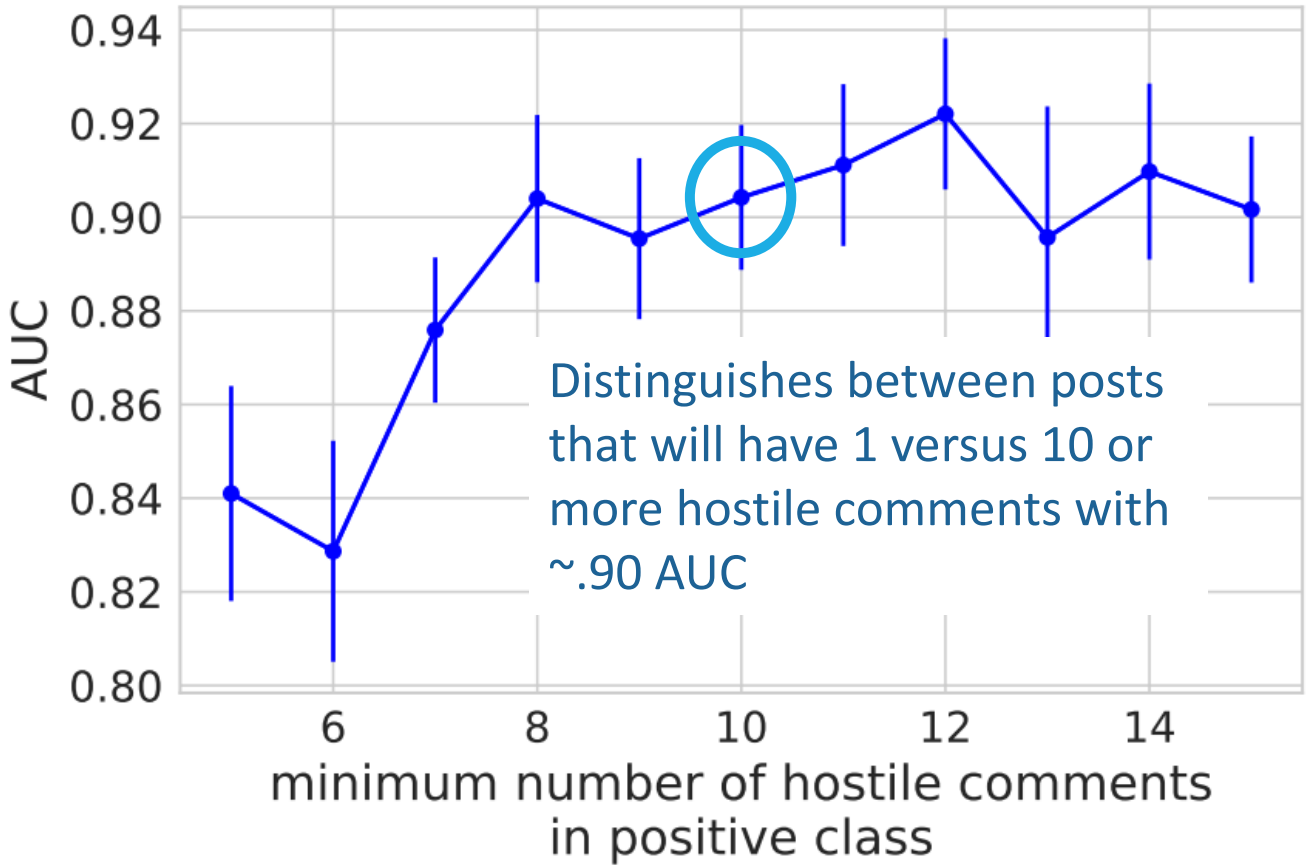
Figure 3: Hostility presence forecasting accuracy as lead time increases.

Experiments - Task 1 - Presence

Features	AUC	F1	Pre - Pos	Rec - Pos
Unigram	0.790	0.737	0.723	0.778
U + prev-com	0.707	0.672	0.660	0.693
U + final-com	0.779	0.729	0.716	0.757
U + trend	0.790	0.739	0.725	0.776
U + user	0.791	0.742	0.729	0.778
U + lex	0.792	0.750	0.732	**0.788**
U + w2v	0.794	0.725	0.715	0.749
U + n-w2v	0.810	0.736	0.725	0.746
U + prev-post	0.828	0.761	0.756	0.765
Best	**0.843**	**0.765**	**0.755**	0.778

Table 2: Forecasting accuracy of Task 1 (lead time = 3 hours).
The best combination uses all features except for w2v and prev-com.

Experiments - Task 2 - Intensity



Experiments - Task 2 - Intensity

Features	AUC	F1	Pre - Pos	Rec - Pos
Unigram	0.808	0.747	0.741	0.673
U + w2v	0.753	0.696	0.662	0.673
U + prev-com	0.786	0.701	0.694	0.605
U + user	0.817	0.761	0.752	0.695
U + n-w2v	0.821	0.775	0.781	0.711
U + trend	0.825	0.778	0.782	0.721
U + lex	0.827	0.776	0.785	0.705
U + prev-post	0.842	0.782	**0.829**	0.688
U + final-com	0.879	0.792	0.805	0.722
Best	**0.913**	**0.805**	0.785	**0.772**

Table 3: Forecasting accuracy of Task 2 (N=10).
The best feature combination is trend/user/final-com.

Experiments - Task 2 - Intensity

Features	AUC	F1	Pre - Pos	Rec - Pos
Unigram	0.808	0.747	0.741	0.673
U + w2v	0.753	0.696	0.662	0.673
U + prev-com	0.786	0.701	0.694	0.605
U + user	0.817	0.761	0.752	0.695
U + n-w2v	0.821	0.775	0.781	0.711
U + trend	0.825	0.778	0.782	0.721
U + lex	0.827	0.776	0.785	0.705
U + prev-post	0.842	0.782	**0.829**	0.688
U + final-com	0.879	0.792	0.805	0.722
Best	**0.913**	**0.805**	0.785	**0.772**

Table 3: Forecasting accuracy of Task 2 (N=10).
The best feature combination is trend/user/final-com.

Coefficient indicators

1. 

2. 

3. The abbreviation “stfu”

4. Singular second- and third-person pronouns like “you” and “she”

False negative example

Index	Text
Caption	Happy birthday mom
Comment1	Aww that's so sweettt!!!
Comment2	awh tell her I said happy birthday 💞
Comment3	Happy birthday
Comment4	🎉🎉🎉
Comment5	Happy B Day !!
……	……
Comment50	**attention whore**
Comment51	stop being a fxxking asshole
Comment52	fxxk off
Comment53	How about you stop being a racist asshole and get a life

Limitation

- **Data volume limitations**
- **Data demographic bias**

Conclusion

- Proposed models for two hostility forecasting tasks
- Discovered several predictors of future hostility
  - the post's author has received hostile comments in the past
  - the use of user-directed profanity
  - the number of distinct users participating in a conversation
  - trends in hostility thus far in the conversation
- Provided new ways to prioritize specific posts for intervention
- Code: https://github.com/tapilab/icwsm-2018-hostility