Language Dynamics in Social Media

Animesh Mukherjee

Department of Computer Science and Engg., IIT Kharagpur

Suman

Binny

Abhik

Jasabanto

Pawan

Monojit

Chris

Location of IITs. 18 functioning IITs (green). 5 planned IITs (blue)

CNERG

CNeRG
Complex Networks Research Group

CONTACT
+91-3222-282255

LOCATION
CSE, IIT Kharagpur

Faculty@CNeRG

Pawan Goyal

NLP, OSN

Bivas Mitra

N/W Sc.,
Transport N/W

Niloy Ganguly

N/W Sc., OSN
Transport N/W

Sourangshu Bhattacharya

ML

Sandip Chakraborty

IoT, Mobile N/W
Transport N/W

Saptarshi Ghosh

OSN, Comp.
Journalism

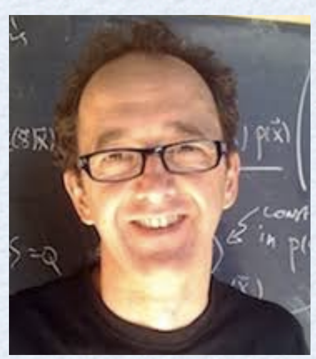
Collaborators@CNeRG

Chris Biemann

UHH

Monojit Choudhury

MSR, India



Matteo Marsili

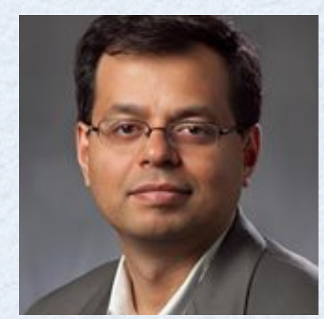
ICTP

Krishna Gummadi

MPI-SWS

Soumen Chakrabarti

IIT Bombay



Romit Roy Choudhury

UIUC

Prasenjit Mitra

PennState

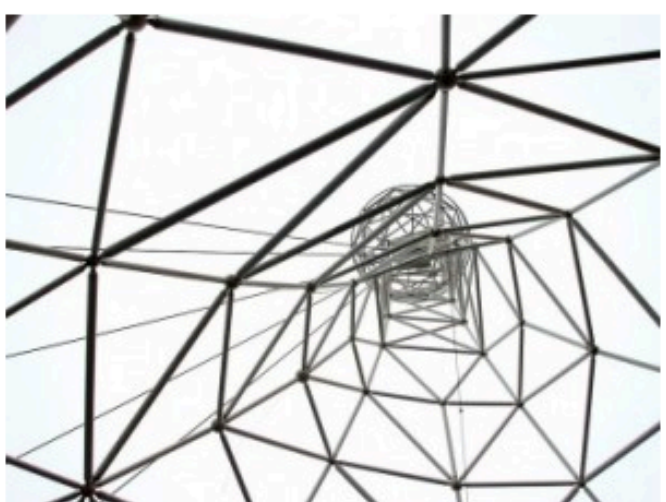
Research@CNeRG


Internet of Things


Systems


Application domains


Web and Social


Network Science & ML

Natural Language Processing

Research Contributions

Today's talk

- Language Dynamics **(1 PhD completed, 2 PhDs pursuing)**

ACL' 14* ICWSM '15 (1+1)* CSCW '16* ICWSM '17* EMNLP '17* COLING '18* CSCW '18*

PNAS, USA '12 PRE '12, '13 Euro. Ph. Lett. '13 JNLE '15 IEEE T CSS '18**

- Citation Analysis **(1 PhD completed, 2 PhDs pursing)**

JCDL' 14 ASONAM '14 JCDL '15 CIKM '15* KDD 2015* PAKDD '16 CIKM '16* KDD 17*

J. of Informetrics '15 Communications of the ACM '15* J. of Informetrics '16 IJDL '18

- Networks Science **(1 PhD complete, 1 PhD pursuing)**

ASONAM' 13 KDD '14* ASONAM '16 CIKM '16* ASONAM '18 ECML/PKDD '18* CIKM '18*

IEEE JSAC '13* Nature Sci. Rep. '13* IEEE TKDE '16* ACM TKDD '16* ACM Comp. Surv. 17*

* Most prestigious / top-tier publications


LinkedIn
LIVEJOURNAL
foursquare
vimeo
YouTube
facebook
Blogger
my
twitter
flickr
tumblr.
Like
yelp

Why care about language?

Business Review

The Reason Twitter's Losing Active Users

by Umair Haque

FEBRUARY 12, 2016

I think the answer's hidden in plain sight. In case you don't use Twitter, the unfortunate fact is that today, it's less like a town square and something more like a mosh pit. Not just freewheeling, irreverent, and rambunctious, but plagued by harassment, abuse, bullying, intimidation, threats — a ceaseless flickering hum of low-level emotional violence.

Opinion conflicts: An effective route to detect incivility in Twitter

TechXplore   Topics   Week's top   Latest news

Home / Internet

OCTOBER 3, 2018 FEATURE

65

40

Share

Email

A new convolutional neural network model to detect abuse and incivility on Twitter

by Ingrid Fadelli , Tech Xplore

CSCW 2018

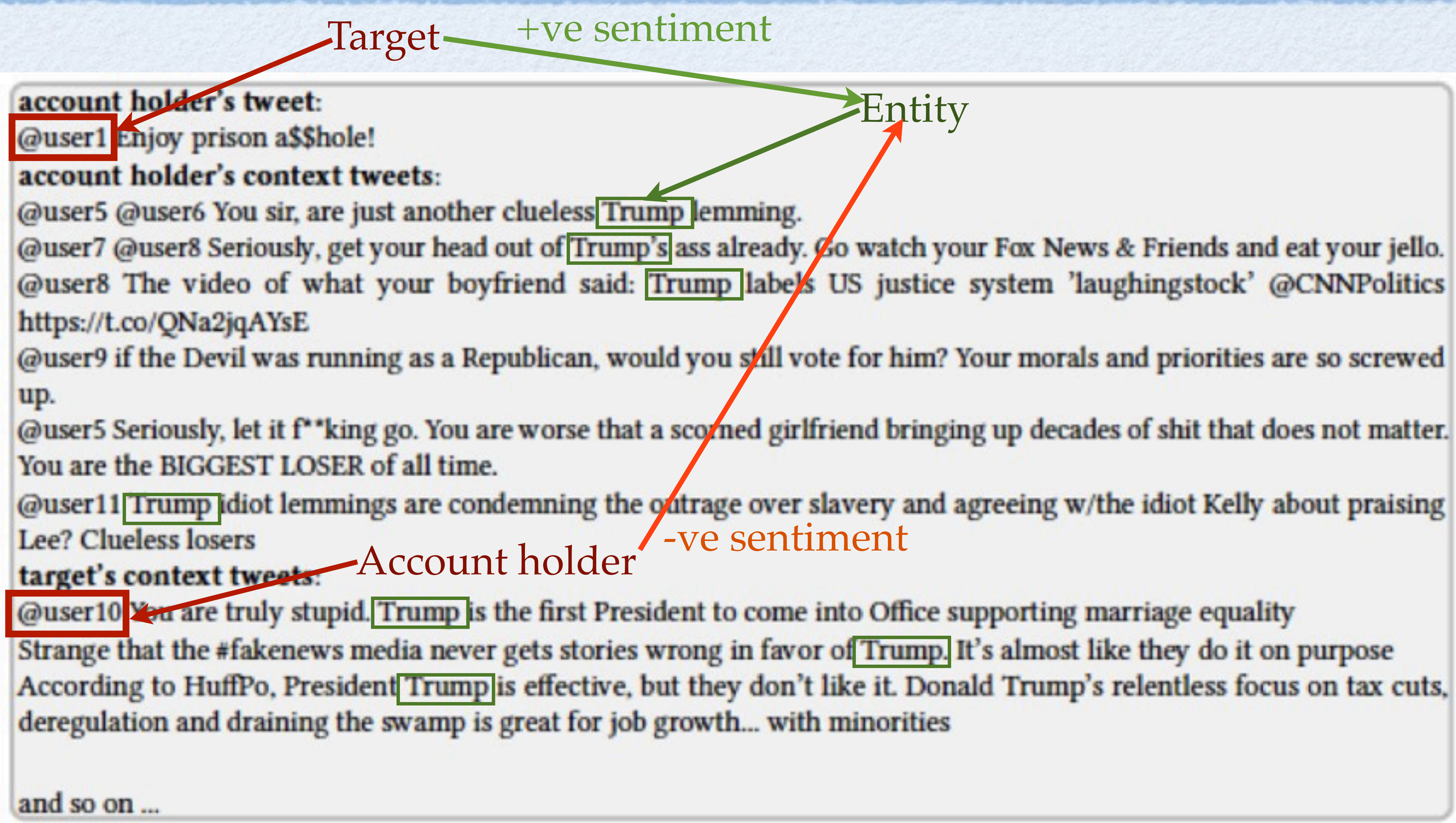
INCIVILITY

- Incivility involves sending harassing or threatening messages (via text message or e-mail), posting derogatory comments about someone on a website or a social networking site (such as Facebook, Twitter etc.), or physically threatening or intimidating someone in a variety of online settings

Our Hypothesis

- If the victim/target has expressed strong sentiments toward an entity mention (i.e., fanship or rivalry) then this might result in the offender/account holder to indulge into incivil/abusive behavior -- we call this ***opinion conflicts***
- This could be observed from the tweets posted by the target and the account holder immediately before/after the abuse happened -- the incivility context
- Can this help in detecting incivility?

An Example -
the Incivility Context

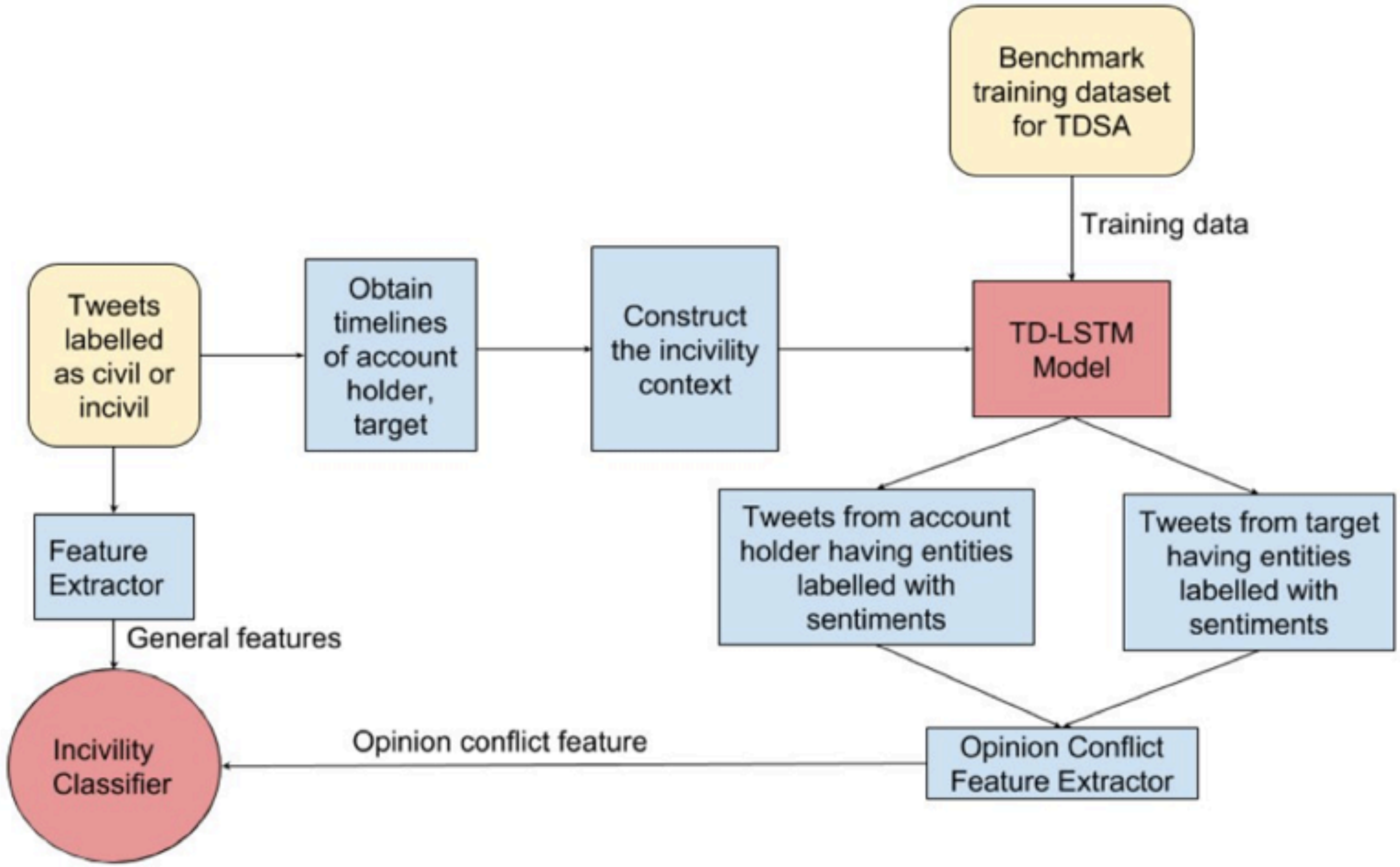


account holder's tweet:
@user1 Enjoy prison a$$hole!
account holder's context tweets:
@user5 @user6 You sir, are just another clueless Trump lemming.
@user7 @user8 Seriously, get your head out of Trump's ass already. Go watch your Fox News & Friends and eat your jello.
@user8 The video of what your boyfriend said: Trump labels US justice system 'laughingstock' @CNNPolitics https://t.co/QNa2jqAYsE
@user9 if the Devil was running as a Republican, would you still vote for him? Your morals and priorities are so screwed up.
@user5 Seriously, let it f**king go. You are worse that a scorned girlfriend bringing up decades of shit that does not matter. You are the BIGGEST LOSER of all time.
@user11 Trump idiot lemmings are condemning the outrage over slavery and agreeing w/the idiot Kelly about praising Lee? Clueless losers
target's context tweets:
@user10 You are truly stupid. Trump is the first President to come into Office supporting marriage equality
Strange that the #fakenews media never gets stories wrong in favor of Trump. It's almost like they do it on purpose
According to HuffPo, President Trump is effective, but they don't like it. Donald Trump's relentless focus on tax cuts, deregulation and draining the swamp is great for job growth... with minorities

and so on ...

Target: Pro Trump, praises US economy
Account holder: Anti-Trump, pro Obama
Opinion conflict leads to incivil behavior

OUR FRAMEWORK FOR INCIVILITY DETECTION



Operationalizing the defn.

- Obtain a list of offensive words (https://www.cs.cmu.edu/~biglou/resources/bad-words.txt)
    - `Swear/profane words: f**k, a**le, b**ch etc.`
    - `Negative words: die, hell, death etc.`
    - `Others: enemy, drug etc.`
- Use the above list to filter the Twitter data stream from August to December 2017 (~2M)
- Mention based filtering -- since we are interested in conversations (~300K)
- Randomly sample ~25K tweets, manually label, perfect agreement between two annotators -- 24271 tweets (incivil: 8800, civil: 15441)

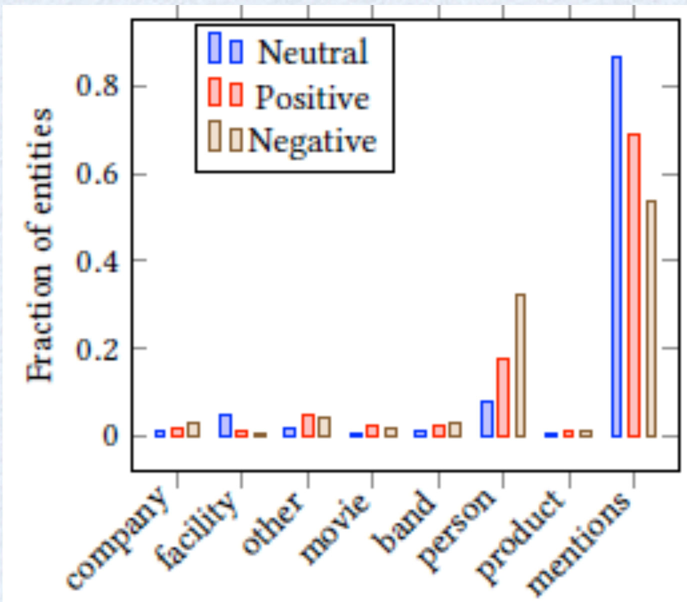
Criteria for labeling

- Blackmails or threats: physical or psychological threats to the victims. Eg., I'll smash you in the face when I see you.
- Insult: insults that are abusive for the victim. Eg., You are such an a**le.
- Cursing: wishing that some grave misfortune befalls the victim. Eg., You'll die and burn in hell.
- Sexual harassment: unwanted sexual talk which might be derogatory. Eg., Post a naked pic u s**t!
- Hey b**hes, feel like seeing a movie tonight? ~~Incivil~~

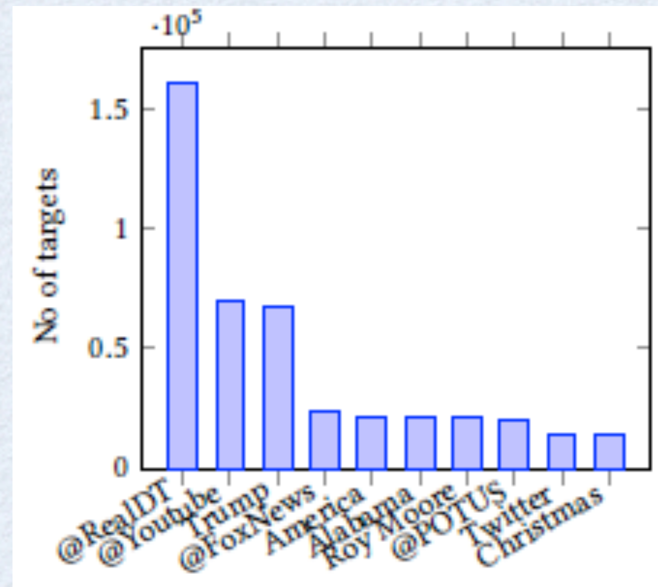
Analyzing Aggression Context

- NER: Twitter NLP tool (https://github.com/aritter/twitter_nlp)
- We find targeted sentiments of named entities using TD-LSTM model.
  - I bought a new mobile phone.The display is amazing but battery life is poor.

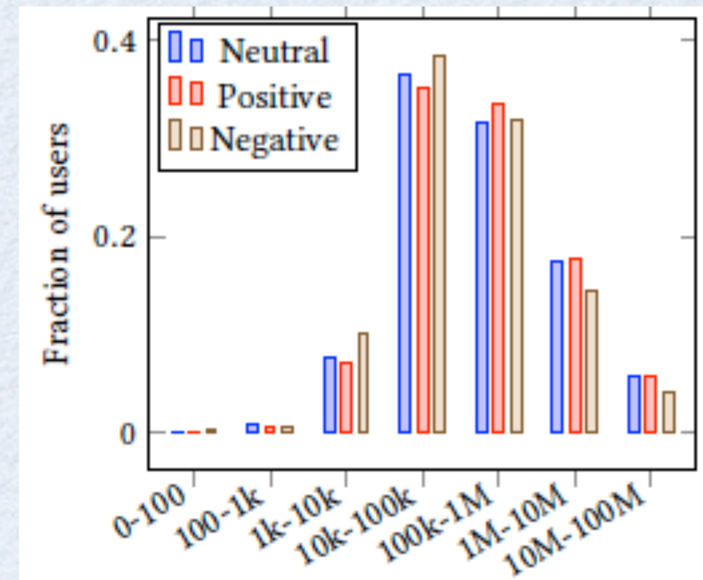
Key observations



Sentiments toward named entity classes



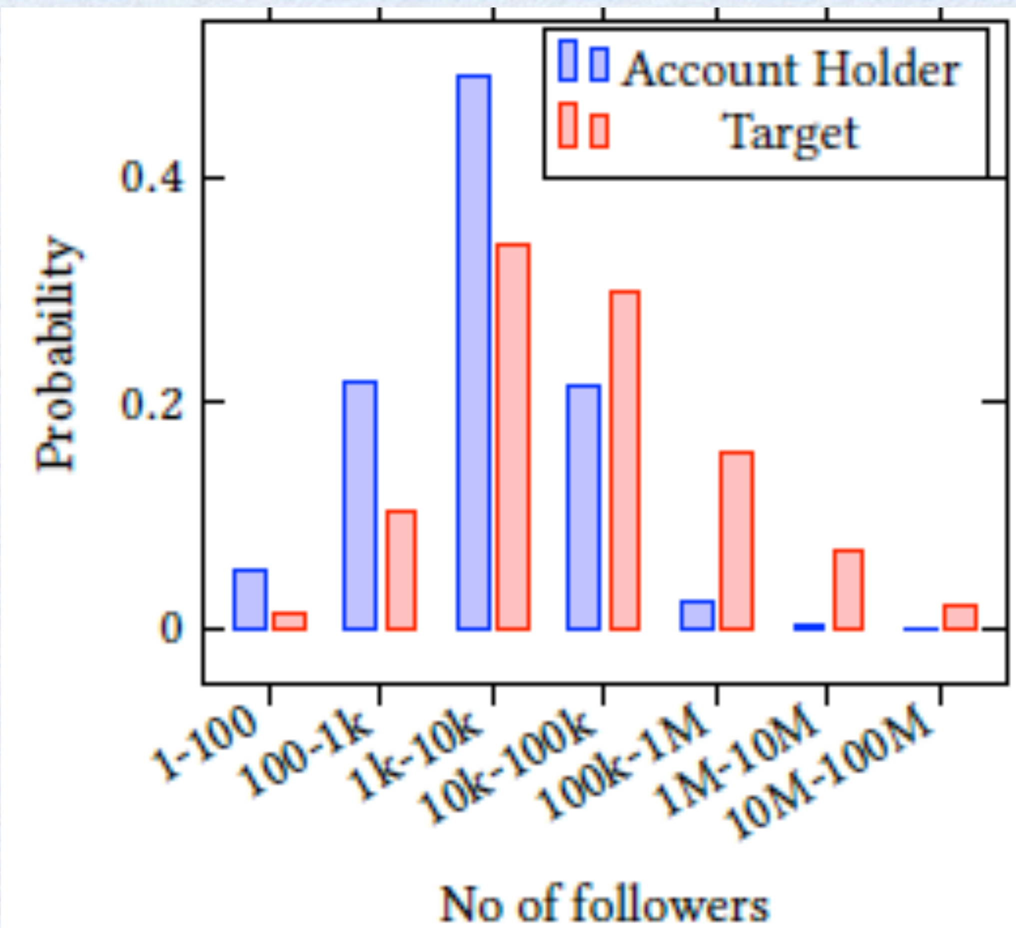
Popular entities



Sentiment vs followership

- Many users express negative sentiments toward person and mention classes
- Users express negative sentiments more often than positive sentiments
- Many targets express opinion about Trump, Youtube and Fox News
- Users with moderately low followership express more negative sentiments toward entities.
- Users with high followership express more positive sentiments toward entities.

Social Prestige

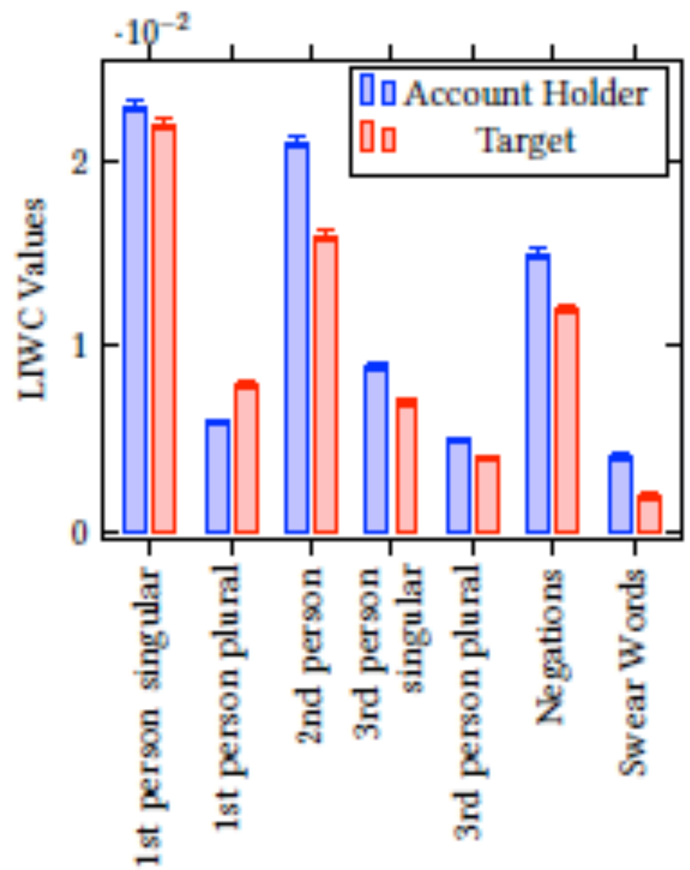


Target accounts are much more popular than the offender accounts (p-value $< 10^{-5}$)

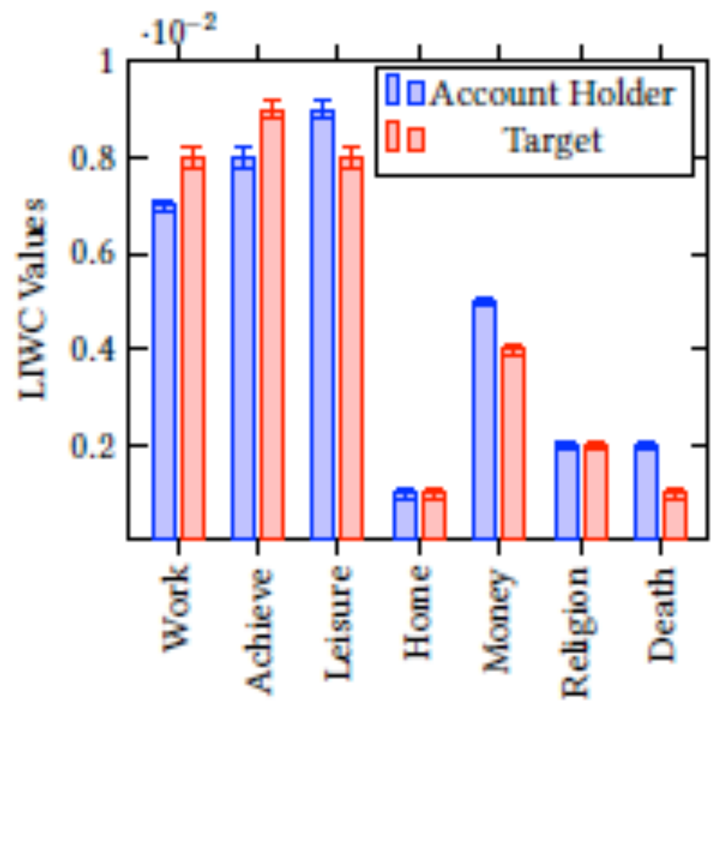
Typically news media anchors accounts like Fox News, Sky Sports, Telegraph, CBS Sweden, CNN, News 18 etc. get victimized

@user1 @newsmedia1 stupid English bitch asking ïs there people on that plane?... No you thick cunt it's like google cars....
Aggression context:
A Dubai firefighter has died of injuries sustained putting out fire after plane crash landing - Emirates chairman https://t.co/i25sjAKfOC

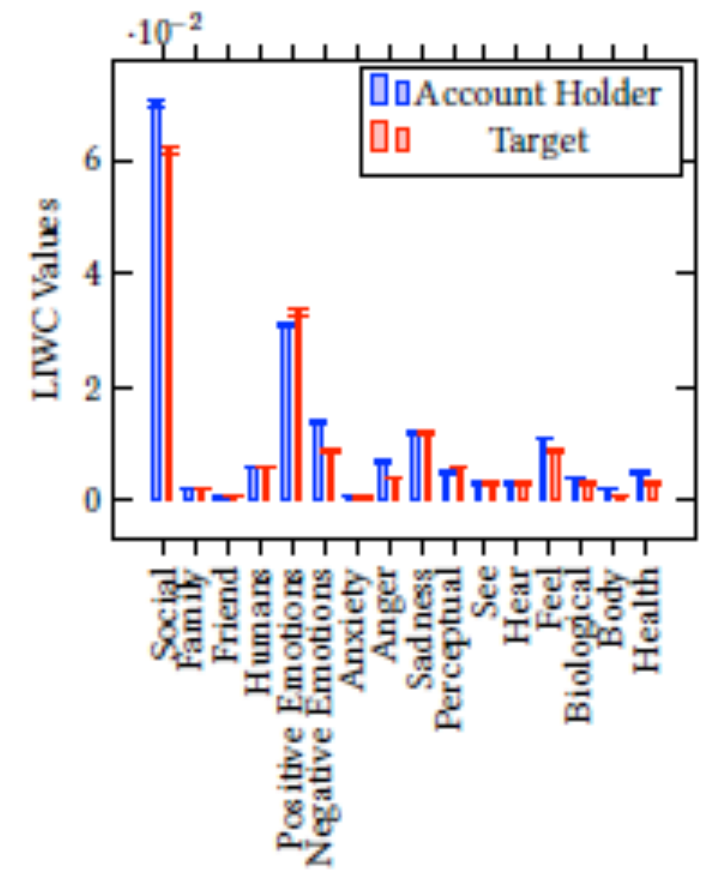
LIWC ANALYSIS



Linguistic categories



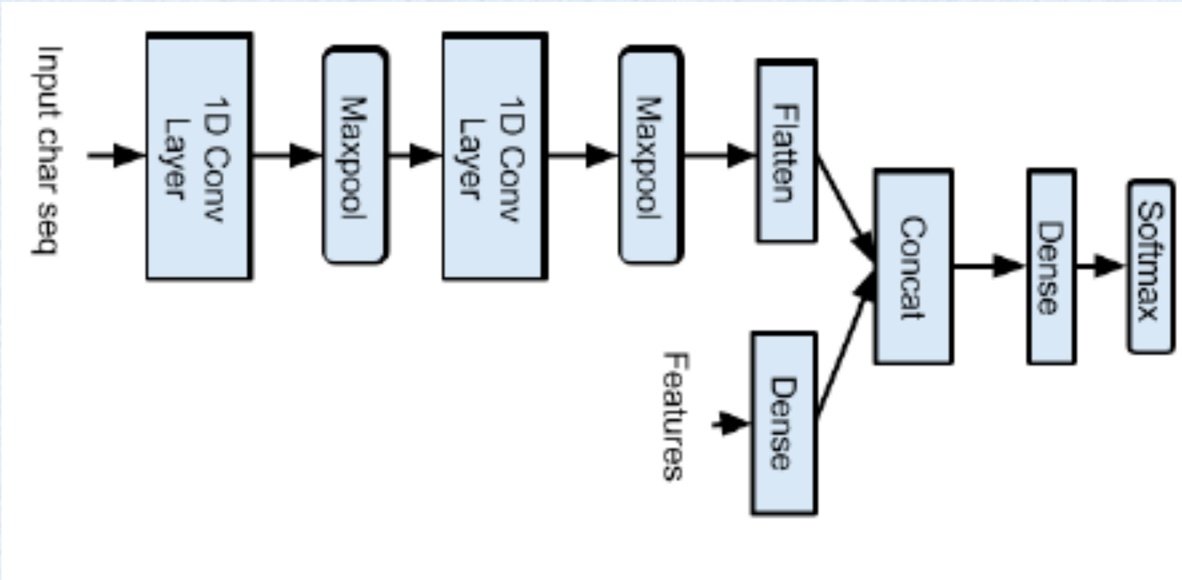
Personal concerns categories



Cognitive categories

- Offenders use first person more frequently than targets
- Offenders use more negation and swear words
- Targets tweet more about work and achievements; offenders speak more about monetary aspects.
- Offenders use “religion” and “death” related words more often.
- Offenders are more expressive about their opinions; tweet more in the categories “body” and “sexual” and express negative emotions.
- Targets express positive emotions more.

OUR PRESCRIPTION

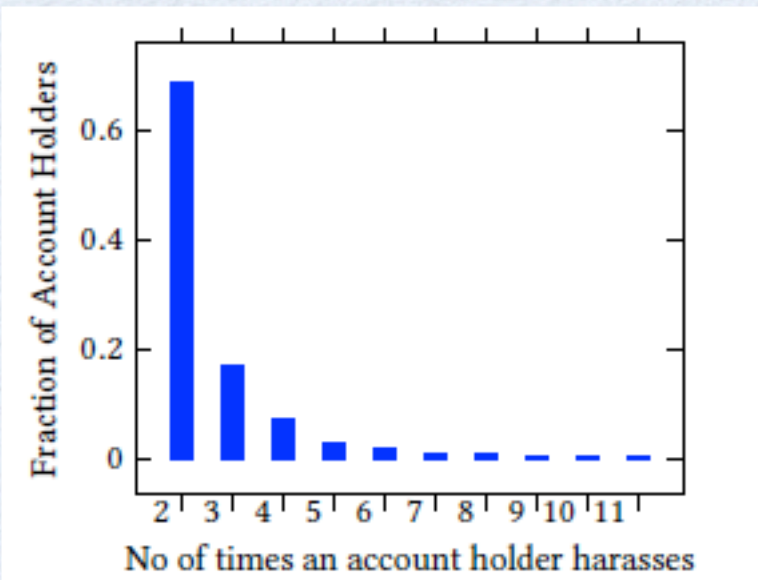


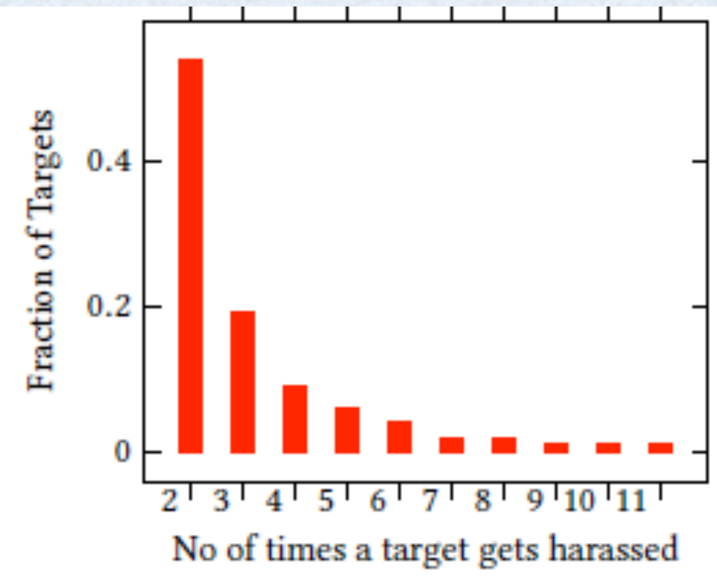
char-CNN + opinion conflict feature based model for incivility detection.

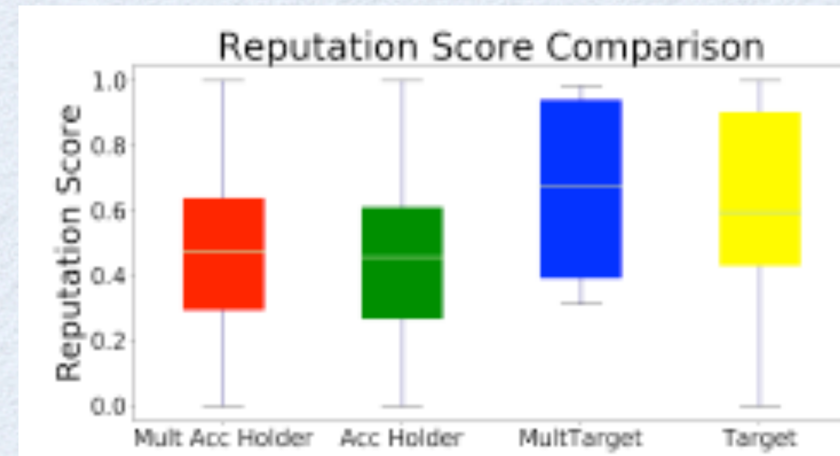
F-score improvement is statistically significant (p-value < 0.05)

Method	Accuracy	F1-Score	ROC-AUC
Content based features			
Bommerson et al. [10])	71.1%	0.27	0.61
n-gram based features			
unigrams	86.4%	0.73	0.91
unigrams (automatic feature selection)	85.6%	0.72	0.91
bigrams	86.7%	0.66	0.79
bigrams (automatic feature selection)	88.4%	0.69	0.81
trigrams	82.3%	0.18	0.54
trigrams (automatic feature selection)	83.6%	0.29	0.59
unigrams + bigrams	86.7%	0.74	0.91
unigrams + bigrams (automatic feature selection) (Xu et al. [104])	87.5%	0.67	0.79
unigrams + trigrams	86.7%	0.63	0.62
unigrams + trigrams (automatic feature selection) (Hosseinmardi et al. [57])	86.7%	0.64	0.62
bigrams + trigrams	86.8%	0.66	0.79
bigrams + trigrams (automatic feature selection)	88.8%	0.76	0.93
unigrams + bigrams + trigrams	86.8%	0.73	0.91
unigrams + bigrams + trigrams (automatic feature selection)	88.9%	0.77	0.92
baseline 1 + opinion conflict	76.1%	0.37	0.61
unigrams + bigrams + trigrams + opinion conflict	80.9%	0.77	0.90
unigrams + bigrams + trigrams + textual features (Chen et al. [21])	86.7%	0.77	0.92
char-LSTM and char-CNN Models [1]			
char-LSTM	88.9%	0.80	0.84
char-LSTM+attention (Pavlopoulos et al. [81])	78.5%	0.53	0.74
char-LSTM+attention (Pavlopoulos et al. [81])+opinion conflict	78.6%	0.54	0.75
char-CNN	93.0%	0.81	0.88
char-CNN + opinion conflict	93.3%	0.82	0.89

Post-hoc Analysis







$$\frac{\#followers}{\#followers+\#friends}$$

@user1 Hillary was a skankb*tch,and they make a cream for your butt-hurt condition (account holder: user3)
@user1 Get ur a$$ off here, u stupid b*tch (account holder: user4)

@user5 you are the biggest f**king piece of scum there is. Karma is a b*tch (account holder: user6)
@user5 you are the biggest scumbag (account holder: user6)

All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media.

EMNLP 17, IKDD 17 DATA CHALLENGE

Lexical Borrowing

- When a word or a phrase from a foreign language is used as a part of the native vocabulary. E.g. This is totally *karma*, she told herself angrily. (from English to Hindi): *botal* from the bottle, *kaptaan* from the captain, *afsar* from officer

- Can be seen in conversations of *monolingual* people

- Borrowed words may attain *native language accent*.

- They finally become part of the *native language vocabulary*.

K. Bali, J. Sharma, M. Choudhury, and Y. Vyas. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In First workshop on Computational approaches to code-switching, *EMNLP*, page 116, 2014.

“Socially” Quantify Borrowing

- Each word is language-tagged. The different tag a word can have is L1 (native), L2 (foreign), NE and Other.
- Tweet level tag based on word level tag
  - L1: Almost every word (> 90%) in the tweet is tagged as L1.
  - L2: Almost every word (> 90%) in the tweet is tagged as L2.
  - CML1: Code-mixed tweet but majority (i.e.,> 50%) of the words are tagged as L1.
  - CML2: Code-mixed tweet but majority (i.e., > 50%) of the words are tagged as L2.
  - CMEQ: Code-mixed tweet having very similar number of words tagged as L1 and L2 respectively.
  - Code Switched: There is a trail of L1 words followed by a trail of L2 words or vice versa.

"Socially" Quantify Borrowing

- The *Unique User Ratio* (*UUR*) for word usage across languages is defined as follows:
  UUR*(*w*) = (*U_L1* + *U_CML1*) / *U_L2
  Where *U_L1* (*U_L2* , *U_CML1*) is the number of unique users who have used the word *w* in a *L1* (*L2* ,*CML1*) tweet at least once.

- The *Unique Tweet Ratio* (*UTR*) for word usage across languages is defined as follows:
  UTR*(*w*) = (*T_L1* + *T_CML1*) / *T_L2
  Where *T_L1* (*T_L2* , *T_CML1*) is the total number of *L1* (*L2* , *CML1*) tweets which contain the word *w*.

Datasets and Pre-processing

- We consider native language *L1* as *Hindi* and foreign language *L2* as *English*
- Considered language tagged data presented by Rudra et al., 2016
- Crawled data over 28 hashtags (Nov 2015 to Jan 2016) spanning over subjects sports, religion, movies and politics.
- This leads to a sum of 811981 tweets.
- After language tagging we got 3982 users who at least code mixed once in their tweets
- We crawled their timeline (Feb 2016 to March 2016)
- Using this two step process we collected a total of 1550714 distinct tweets.
- Filtered tweets having only URLs, written in non-romanized scripts and tweets having empty content. Finally 787606 tweets.

Candidate/Target Words

- Baseline-biased words – ’thing’, ’way’, ’woman’, ’press’, ’wrong’, ’well’, ’matter’, ’reason’, ’question’, ’guy’, ’moment’, ’week’, ’luck’, ’president’, ’body’, ’job’, ’car’, ’god’, ’gift’, ’status’, ’university’, ’lyrics’, ’road’, ’politics’, ’parliament’, ’review’, ’scene’, ’seat’, ’film’, ’degree’

- Randomly selected words – ’people’, ’play’, ’house’, ’service’, ’rest’, ’boy’, ’month’, ’money’, ’cool’, ’development’, ’group’, ’friend’, ’day’, ’performance’, ’school’, ’blue’, ’room’, ’interview’, ’share’, ’request’, ’traffic’, ’college’, ’star’, ’class’, ’superstar’, ’petrol’,’uncle’

Bali et al. 2014

Relied on native language newspaper based signals

Baseline Metric: ***log(F_L2 / F_L1)***

F_L2: Frequency of *L1* transliterated form of a word *w* in the standard *L1* newspaper corpus. (*w* is a *L2* word)

F_L1: Frequency of the *L1* translation of the word *w* in the same newspaper corpus

Baseline rank list: is created by ranking *L2* candidate words in non increasing order of baseline metric.

Example:

L1: Hindi *L2*: English

Candidate word: *film* (*chalachitra* in Hindi)

L1 news paper corpus: Hindi Jagaran

Metric: log(frequency(film)/ frequency(chalachitra))

Ground-Truth Preparation

- Online survey done over 58 participants
- 3 choices were given to participants for each of 57 target word
- *Language preference factor* (*LPF*): (Count_En – Count_Hi) where Count_Hi (Count_En) refers to the number of survey participants who preferred the sentence containing the Hindi translation of the target word (target word itself)
- More positive values of LPF denotes higher usage of target word as compared to its Hindi translation and therefore higher likeliness of the word being borrowed.

CORRELATION/RE-ANNOTATION

Rank-List$_1$	Rank-List$_2$	$\rho - hlws$	$\rho - mws$	$\rho - full$
UUR	Ground truth	**0.67**	**0.64**	**0.62**
UTR	Ground truth	**0.66**	**0.63**	**0.63**
UPR	Ground truth	**0.66**	**0.64**	**0.62**
Baseline	Ground truth	0.49	0.14	0.26

Word-rank	Context	$\mu_{E \to H}$	$\sigma_{E \to H}$
TOP	H_{all}	**0.91**	0.15
TOP	H_{most}	**0.85**	0.23
MID	H_{all}	0.58	0.28
MID	H_{most}	0.61	0.34
BOT	H_{all}	0.13	0.18
BOT	H_{most}	0.16	0.21

On Avg. 88% annotators changed the language tag for Top bucket across Hall and Hmost

#Bieber + #Blast = #BieberBlast: Early Prediction of Popular Hashtag Compounds

CSCW ’16 (Best Paper Honorable Mention Award)

Lexical compounding

Prevalent all through over the history of evolution of any language

wheel + chair = wheelchair
white + wash = whitewash
in + so + far = insofar

In some cases, meanings of the compound also get altered:

book + worm = bookworm
light + house = lighthouse

HASHTAG COMPOUNDING

Justin Bieber Agent @JBieber_Agent · 6h
Justin Bieber's Stepmom-To-Be Is Only 7 Years Older Than Him #**bieber**
dragplus.com/post/id/337114...
2 2

Naeha Sareen @NaehaSareen · 10h
Yassss! Check out my friend Dan! His #**bieber** cover is 🙌🏽🙌🏽😍😍

> **Dan Talevski** @dantalevski
> Check out my new @justinbieber #LoveYourself cover!
> youtube.com/watch?v=ezkn6_...

2 10

Matt Yoakum @Matt4Music · 22h
Shooting a music video today with @RebeccaZamolo and @filmmanian
#**bieber** #purpose
14 133

Janet ❤️ @Bieber_rose · Feb 28
Justin is the cutest Boy☺️ #justin #**bieber** #smile

#bieber

Jay arc Mwesigwa @Jayarc13 · Feb 19
Happy **Birthday** @mkavine. Forever my boo! #**Blast**
1 1

iRajeshChauhan™ @irajesh81 · Feb 16
This #types of dish go to another world #**birthday** #party #**blast** #pavbhaji ♡.♡
#antiquity 🍺 pic.twitter.com/hfw1ob1OgR
2
View photo

Bill Janisse @BillJanisse · Feb 16
Having a #**blast** at #mumm in #Napa drinking #bubbles and enjoying the #view
, best #**birthday**... instagram.com/p/BB0yF01neAJ/
1 3

Nichole Long @nikkifumasta · Feb 14
#Happy #**Birthday** Audrina! We went to #BuildABear where the kids had a
#**blast**!! @ Build-A-Bear... instagram.com/p/BBv7TIUj00Q/

Chukkuvellam @chukkutweets · Feb 10
vichuu_ : #gang #**birthday** #gangsters #**blast** #freak #cake #cutting @
Thangasseri-Kollam Kerala-India. instagram.com/p/BBnLDYXF43H/ ...

#blast

Edelyn Rafael @QueenOfJBiebs · 13 Nov 2015
My neighbor's will be blessed because i will be playing @justinbieber's
#Purpose on repeat. #**BieberBlast**

HE FINALLY FOLLOWED @mylovelyjdb · 13 Nov 2015
CANNOT WAIT TO GO HOME AND BLAST PURPOSE OMG #**BIEBERBLAST**
#PURPOSE #ITSFINALLYHERE I can't breath
1

briana/PINNED TWEET! @SLICKSPRAYBERRY · 13 Nov 2015
My 10 year old BIEBER Fever infested self is back
#**bieberblast** #purpose #BieberIsBack
1

Leon Purvis @LeonPurvis · 13 Nov 2015
#**BieberBlast**! #Purpose #NewAlbum instagram.com/p/-Az4dNgjBP/
1 1

Marciee @Jxstinxluke · 12 Nov 2015
GUYS REMEMBER WHEN WE USE TO #**Bieberblast** new albums!!! ;)
@justinbieber

#bieberblast

Coalition of Concepts

#Wikipedia + #Blackout = #WikipediaBlackout

general concepts

specific concept

#CSCW + #2016 = #CSCW2016

Use of hashtag compounds

- **Marketing** - #AmazonPrimeDay, #USDEUR
- **Expressing communicative intent** – affective expression, political persuasion, humor etc. E.g., #PresidentTrump, #BlackLivesMatter
- **Spontaneous pressure** - #TheBestFeelingInARelationship, #YouKnowItsRealWhen, #RelationshipTips

INFLUENCED VS SPONTANEOUS

Spontaneous

Influenced

Hashtag compounds	no.of mentions per tweet	no. of retweets per tweet	no. of collocations per tweet	no. of mentions in first 50 tweets
#TheBestFeelingInARelationship	0.074	0.370	0.148	2
#10WorstFeelings	0.041	0.434	0.097	0
#YouKnowItsRealWhen	0.071	0.372	0.208	3
#RelationshipTips	0.023	0.498	0.175	1
#CMTAwards	0.446	0.225	0.401	12
#JessicaForTheWin	0.324	0.294	0.853	11
#SmartGalaxyS3	0.495	0.430	0.183	16
#BringBackToonami	0.565	0.214	0.159	20

- Spontaneous compounds have lesser no. of mentions per tweets and lesser no. of collocations with other hashtags
- Influenced hashtag compounds spread via multiple mentions in early stage of propagation unlike the spontaneous ones.

Adoption of hashtag compounds

- **Popular:** frequency of occurrences of #AB is higher than both #A and #B
- **Unpopular:** frequency of occurrences of #AB is lower than both #A and #B

Frequency after 10 months from compounding

Popular	Formation
#HighSchoolMemories (21700)	#HighSchool (395) + #Memories (4178)
#OperationLegalizeWeed (3978)	#Operation (18) + #LegalizeWeed (12)
#WikipediaBlackout (2638)	#Wikipedia (202) + #Blackout (524)
#CNNDebate (2615)	#CNN (1637) + #Debate (125)
#GoldenGlobes (2581)	#Golden (125) + #Globes (61)
Unpopular	**Formation**
#LoveOomf (1)	#Love (14525) + #Oomf (142299)
#YOLOForJesus (1)	#YOLO (47056) + #ForJesus (4)
#HateCanada (3)	#Hate (1622) + #Canada (2399)
#SweetBabyJesusThatsGood (1)	#SweetBabyJesus (45) + #ThatsGood (27)
#LiquidationMonday (3)	#Liquidation (51) + #Monday (965)

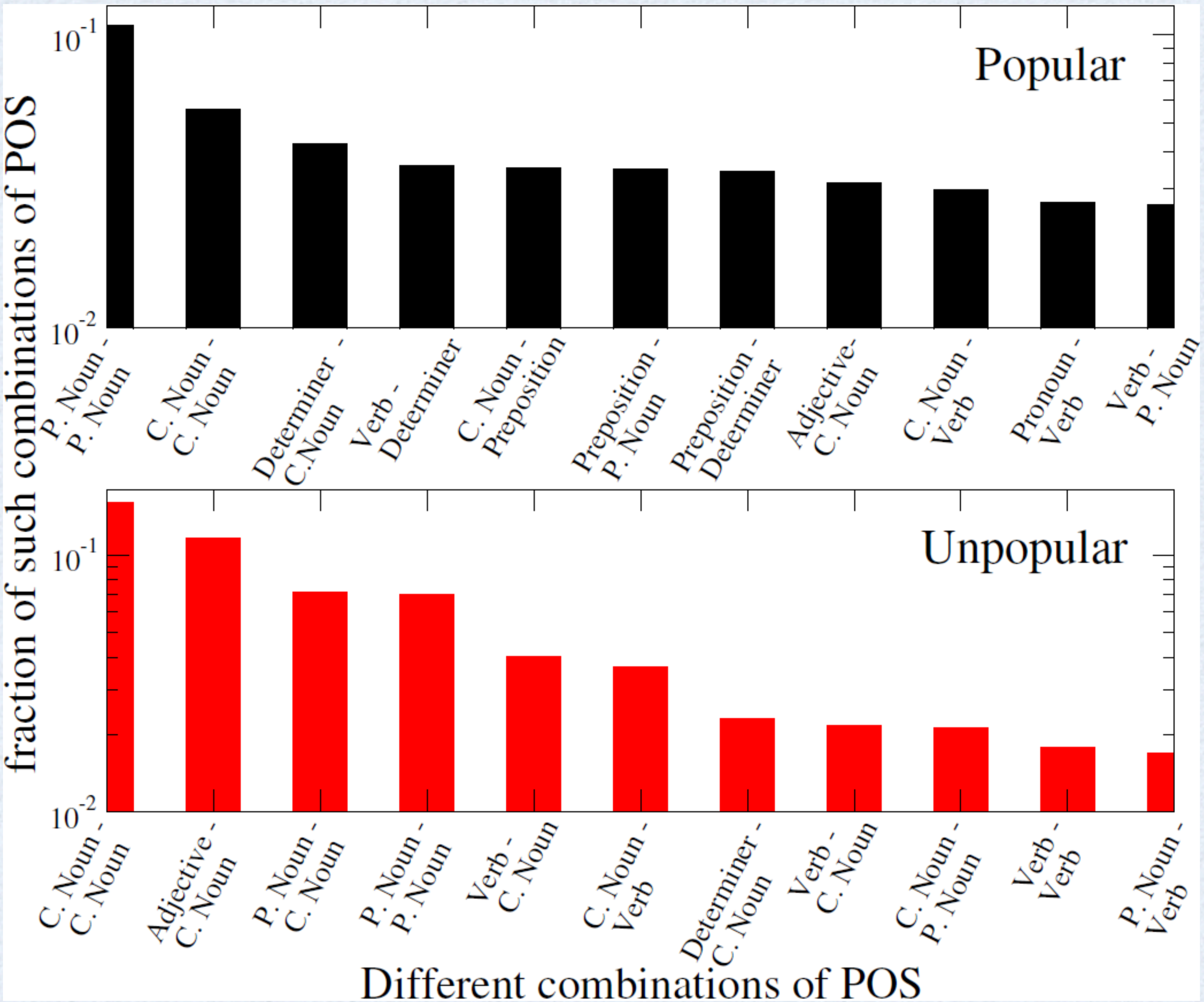
Research Questions

- Can we systematically obtain early signals from the data that differentiates the popular from the unpopular compounds?
- What features makes a hashtag compound "popular"?
- Can we predict with high accuracy whether a compound will eventually become popular early in time, i.e., even before the compounding have taken place?

Linguistic Analysis

- Which parts-of-speech (POS) combinations are more frequent?
- Which named entity (NE) combinations are more frequent?
- What kind of words combine more? (dictionary words/ new words like celebrity names or chattish words)

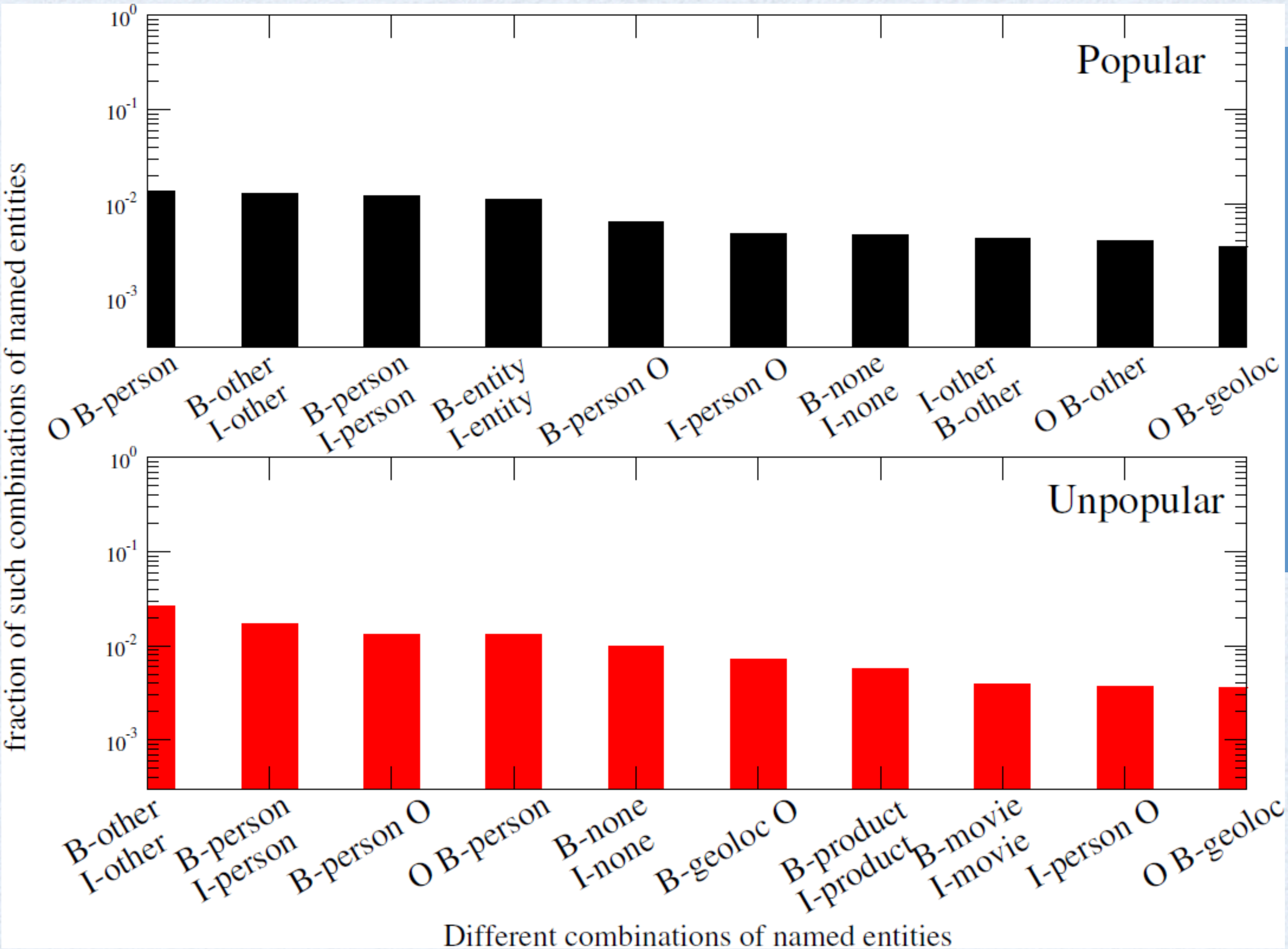
Part-of-speech combination



In natural language compounding, **noun-noun, verb-noun, noun-verb, verb-verb** are some common forms of compounding

For popular compounds : proper noun-proper noun, common noun-common noun, determiner-common noun and verb-determiner; **For unpopular compounds:** common noun-common noun, Adjective-common noun, determiner-common noun, proper noun-proper noun etc

Named-entity combination



- **The most prevalent NE :**
  person-person
  product-product
  movie-movie
- **But fractions differ** between popular and unpopular ones.

Vocabulary effects

Popular		Unpopular	
Combinations	%	Combinations	%
INV-INV	43.9	INV-INV	66.9
OOV-OOV	20.7	OOV-INV	14.0
INV-OOV	19.8	INV-OOV	13.6
OOV-OOV	15.6	OOV-OOV	5.5

Marked distinction

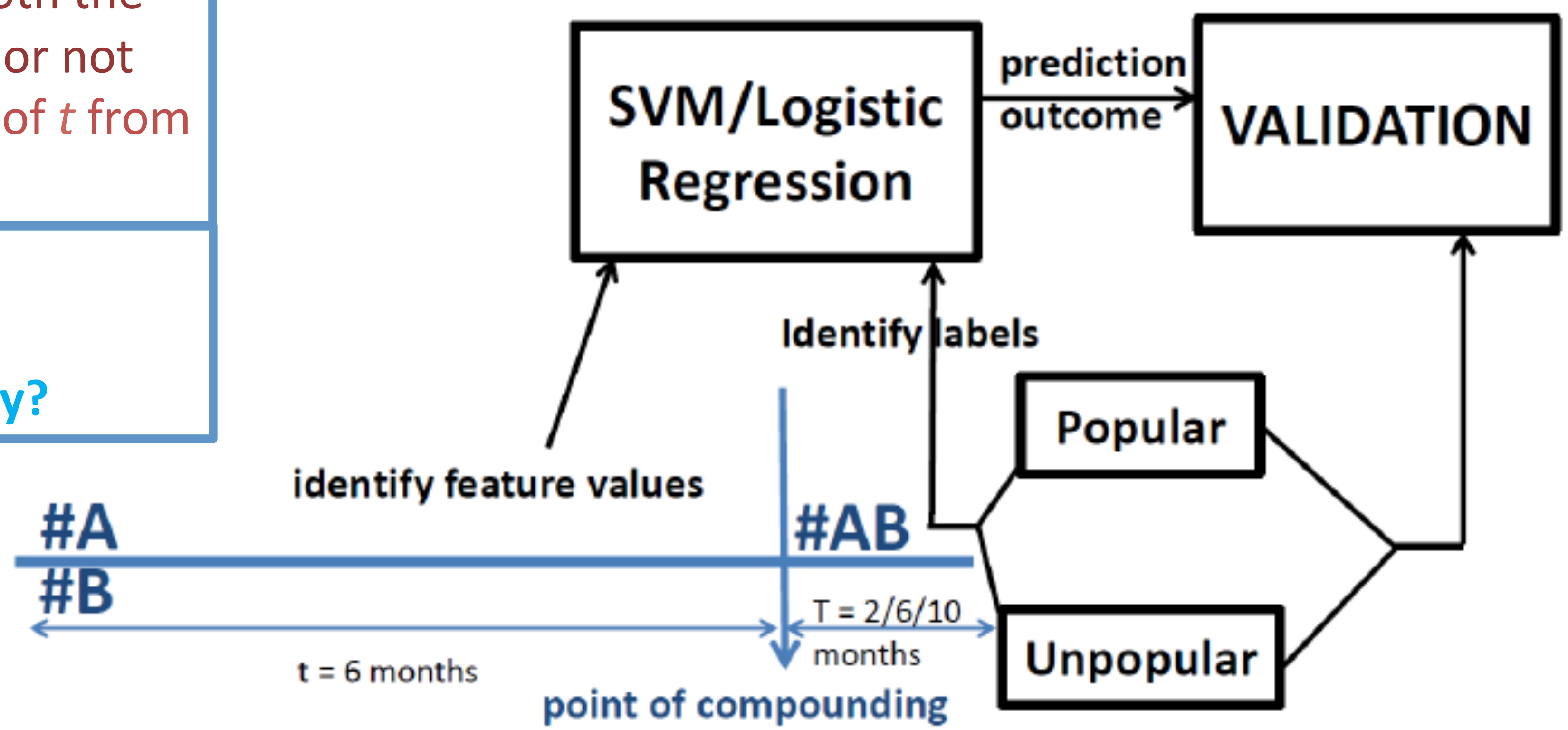
-- rank order in which the combinations are used

-- individual % of use

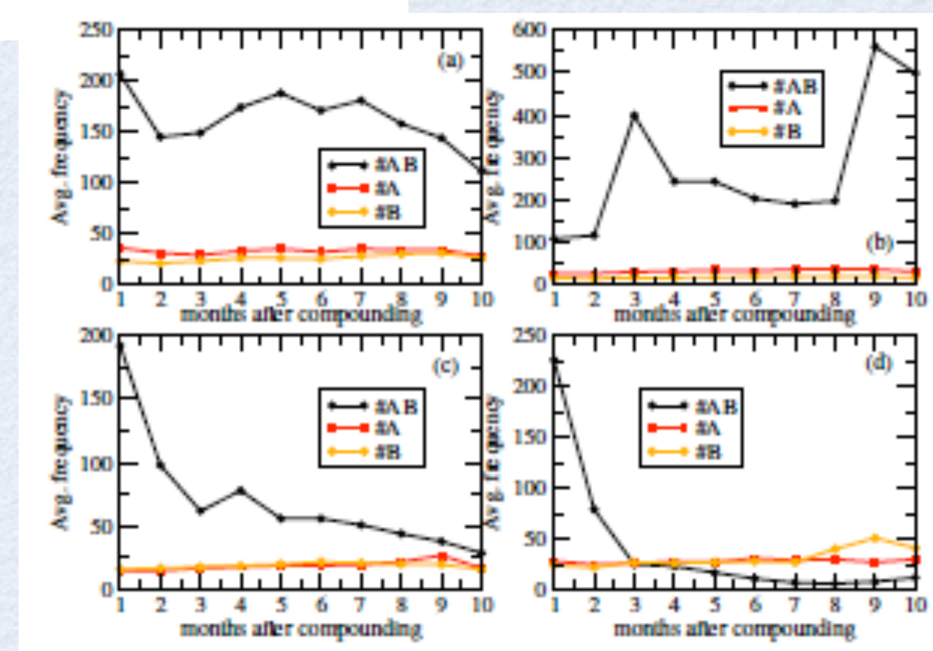
PREDICTION FRAMEWORK

Predict
-- whether a particular hashtag compound will be more popular than both the constituent hashtags or not
-- after a time period of *t* from the point of merging

#AmazonPrimeDay
or
#Amazon #PrimeDay?



Observation period



Prediction Features

Hashtag Content features

- Character length of the compound hashtag
- Presence of n-grams in English texts
- Part-of-speech combination
- Named entity combination
- OOV/INV combination

Tweet Content features

- Word overlap
- n-gram overlap
- Collocation frequency of the compounding pair
- Word diversity of the compounding hashtags
- Average topic overlap among the compounding hashtags

User features

- no. of unique users tweeting the individual hashtags
- no. of unique users being mentioned in tweets containing the individual hashtags
- no. of common users mentioned in same tweets containing both #A and #B
- no. of retweets using #A / #B

Performance Evaluation

Time period	Classifier	K	Accuracy	Precision	Recall	F-Score	ROC Area
$T = 2$ months	SVM(10-fold cross validation)	10	76.18	0.762	0.762	0.762	0.762
		20	76.42	0.764	0.764	0.764	0.764
		30	**77.07**	**0.771**	**0.771**	**0.771**	**0.771**
		40	76.37	0.764	0.764	0.764	0.764
		50	76.72	0.767	0.767	0.767	0.767
	Logistic Regression(10-fold cross validation)	10	76.13	0.761	0.761	0.761	0.836
		20	76.43	0.764	0.764	0.764	0.839
		30	**76.48**	**0.765**	**0.765**	**0.765**	**0.841**
		40	76.27	0.763	0.763	0.763	0.838
		50	76.42	0.764	0.764	0.764	0.837
	SVM(seperate train and test set)	30	77.7	0.777	0.77	0.772	0.771
	Logistic Regression(seperate train and test set)	30	77.5	0.781	0.775	0.776	0.834

We achieve 77.07% accuracy for predicting after 2 months. For long term predictions, we achieve 77.5% and 79.13% for 6 and 10 months respectively

Importance of Feature Groups

Feature model	Accuracy
All	**77.07%**
tweet content + user	75.9%
tweet content + hashtag content	75.12%
hashtag content + user	72.4%
tweet content	74.1%
user	68.18 %
hashtag content	65.04%

Tweet content features comes first followed by the users features

Correspondence Analysis

✦ 600 randomly selected hashtag compounds from 2000 compounds used for classification

✦72 participants; 25 questions; each question is answered by 3 participants

Avg. frequency

	Popular			Unpopular		
	#AB	#A	#B	#AB	#A	#B
Correctly judged by both *HE* and *AF*	1324.25	130.63	117.62	2.4	1369.34	1297.47
Wrongly judged by both *HE* and *AF*	1610.33	433.25	136.58	5.77	180.7	250.23
Correctly judged by only *HE*	1644.4	460.24	332.5	1.48	576.05	949.33
Correctly judged by only *AF*	259.3	46.3	73.3	12.24	1130.91	849.34

- human evaluators can correctly label those cases where the hashtag compound have the highest frequency for the popular class and lowest for the unpopular class
- the automatic prediction framework can identify the popular hashtag compounds whose frequency values are not very different from the constituent hashtags

WE ALSO PARTY!

Visit us at: http://www.cnergres.iitkgp.ac.in/