QUESTIONS ABOUT QUESTIONS: AN EMPIRICAL ANALYSIS OF INFORMATION NEEDS ON TWITTER

ABSTRACT

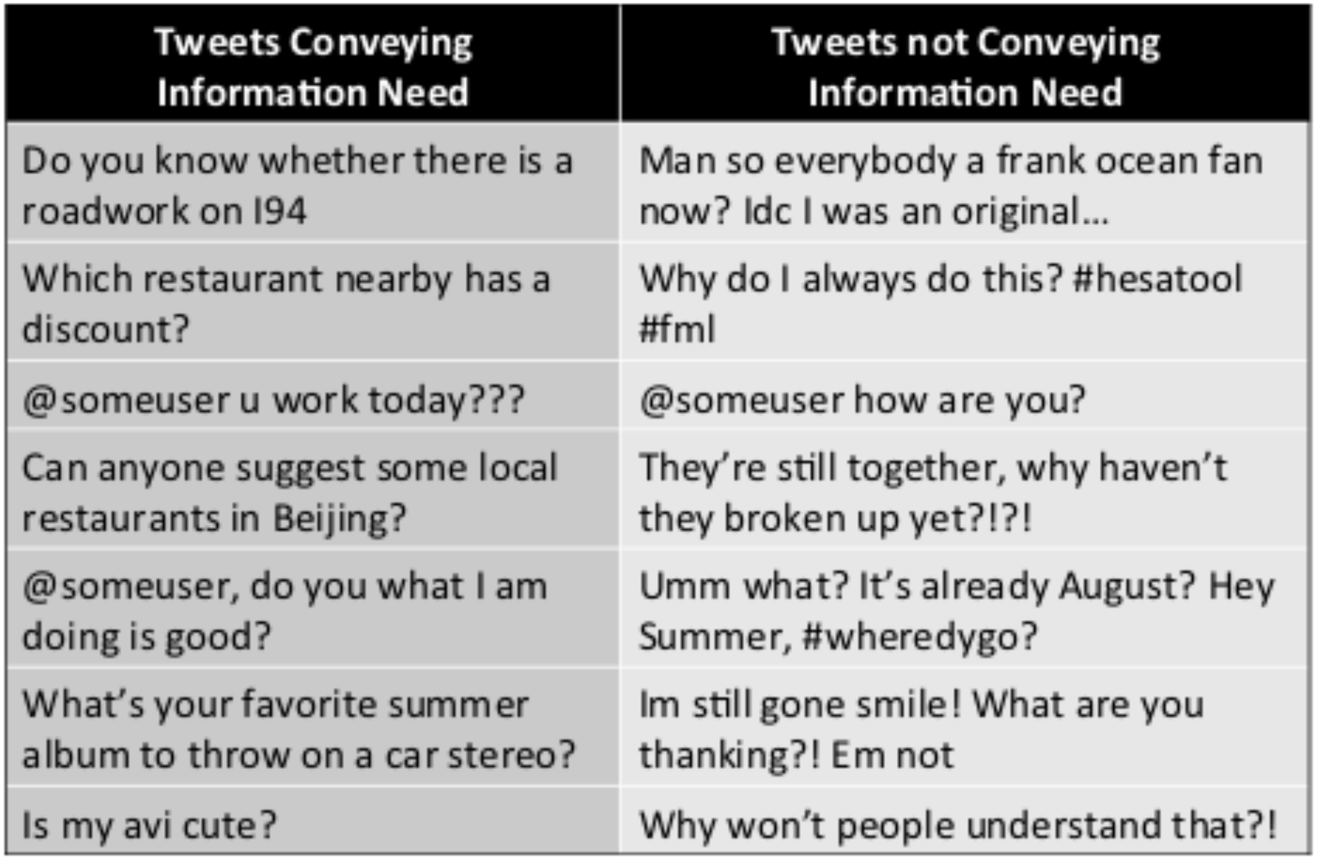
- We take the initiative to **extract and analyze information needs** from billions of online conversations collected from Twitter.
- We can accurately **detect real questions** in tweets
- We then present **a comprehensive analysis** of the large-scale collection of information needs we extracted.

INTRODUCTION

- 13% of a random sample of tweets were questions.

1. Broadcasting
2. Targeting the question to particular friends.

- Information needs through social platforms present a higher coverage of topics related to human interest, entertainment, and technology, compared to search engine queries.



Tweets Conveying Information Need	Tweets not Conveying Information Need
Do you know whether there is a roadwork on I94	Man so everybody a frank ocean fan now? Idc I was an original…
Which restaurant nearby has a discount?	Why do I always do this? #hesatool #fml
@someuser u work today???	@someuser how are you?
Can anyone suggest some local restaurants in Beijing?	They're still together, why haven't they broken up yet?!?!
@someuser, do you what I am doing is good?	Umm what? It's already August? Hey Summer, #wheredygo?
What's your favorite summer album to throw on a car stereo?	Im still gone smile! What are you thanking?! Em not
Is my avi cute?	Why won't people understand that?!

Figure 1: Instances of tweets conveying an information need, and those which don't.

CONTRIBUTION

- We present the first very large scale and longitudinal study of information needs in Twitter.
- We prove that information needs detected on Twitter have a considerable power of predicting the trends of search engine queries.
- Through the in-depth analysis of various types of time series, we find many interesting patterns related to the entropy of language and bursts of information needs.

EXPERIMENT SETUP

- The collection covers a period of 358 days, from July 10th, 2011 to June 31st 2012.
- 4,580,153,001 tweets

- We focus on tweets that contain at least one question mark.
- 81.5% of information needs asked through social platforms were explicitly phrased as questions and included a question mark.
- In our collection of tweets, 10.45% of tweets contain explicit appearance of question mark(s).

DETECTING INFORMATION NEEDS

- A text classification problem.

1. Give a formal definition of this problem and generate a set of labeled tweets as training/testing examples.
2. Introduce a classifier trained with these examples, using the state-of-the-art machine learning algorithms and a comprehensive collection of features.

DEFINITION AND RUBRICS

- “real questions” :
- A tweet conveys an information need, or is a real question, if it expects an informational answer from either the general audience or particular recipients.

1. it requests for a piece of factual knowledge, or a confirmation of a piece of factual knowledge
2. it requests for an opinion, idea, preference, recommendation, or personal plan of the recipient(s), as well as a confirmation of such information.

HUMAN ANNOTATION

- two human annotators
- sampled 5,000 tweets
- 3,119 tweets are labeled as real tweets
- 1,595 are labeled as conveying an information need and 1,524 are labeled not conveying an information need

- The inter-rater reliability measured by Cohen’s kappa score is 0.8350

TEXT CLASSIFICATION

- **Feature Extraction**
- Lexical features / the semantic knowledge base WordNet / syntactical features

- four different types of feature from each tweet:
  - lexical ngrams, synonyms and hypernyms of words(obtained from the WordNet), ngrams of the part-of-speech (POS) tags, and light metadata and statistical features such as the length of the tweet and coverage of vocabulary,etc..

TEXT CLASSIFICATION

- **Lexical Features**
- We included unigrams, bigrams, as well as trigrams.

- For example tweets beginning with the 5Ws(who, when, what, where, and why) are more likely to be real questions.

- 44,121lexicalfeatures.

TEXT CLASSIFICATION

- **WordNet Features**

- synonyms
- hypernyms

- By doing this, our algorithm can also handle words that haven’t been seen in the training data.

- 23,277 WordNet features are extracted

TEXT CLASSIFICATION

- **Part-of-Speech Features**
- Capture light syntactic information.

1. given a tweet with n words, w 1 ;w 2 ; … ;w n , we extract grams from the part-of-speech sequence of the tweet, is t 1 ;t 2 ;…;t n ,
2. Extract unigrams, bigrams and trigrams from this part-of-speech sequence as additional features of the tweet.

- 3,902 POS features are extracted in total

TEXT CLASSIFICATION

- **Meta Features**
- 6 meta data features and simple statistical features of the tweet
- such as the length of the tweets, the number of words, the coverage of vocabulary, the number of capitalized words, whether or not the tweet contains a URL, and whether or not it mentions other users.

TEXT CLASSIFICATION

Feature Selection

- Reduce the dimensionality of the data

- Bi-Normal Separation

- tpr = tp= (tp + fn)
- fpr = fp= (fp + tn)

$$\|F^{-1}(tpr) - F^{-1}(fpr)\|,$$

- F is the Normal cumulative distribution function

TRAINING CLASSIFIER

1. train four independent classifiers using the Support Vector Machine
2. combine the four classifiers that represent four types of features into one stronger classier using boosting, **Adaptive Boosting**

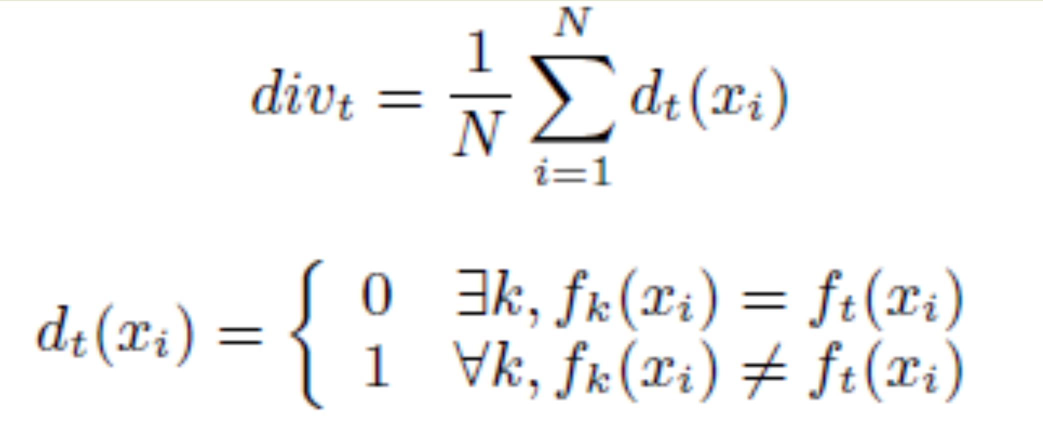
- Adaboost is an effective algorithm that trains a strong classifier based on several groups of weak classifiers.
- AdaBoost方法是一种迭代算法，在每一轮中加入一个新的弱分类器，直到达到某个预定的足够小的错误率。每一个训练样本都被赋予一个权重，表明它被某个分类器选入训练集的概率。如果某个样本点已经被准确地分类，那么在构造下一个训练集中，它被选中的概率就被降低；相反，如果某个样本点没有被准确地分类，那么它的权重就得到提高。

TRAINING CLASSIFIER

- After several iterations, when the combination of weak classifiers starts to achieve a higher performance, the diversity inside the combination is getting lower.

- add a parameter to control for the diversity of the weak learners in each iteration.

- The diversity that a new classifier could add in iteration t is defined as follows:

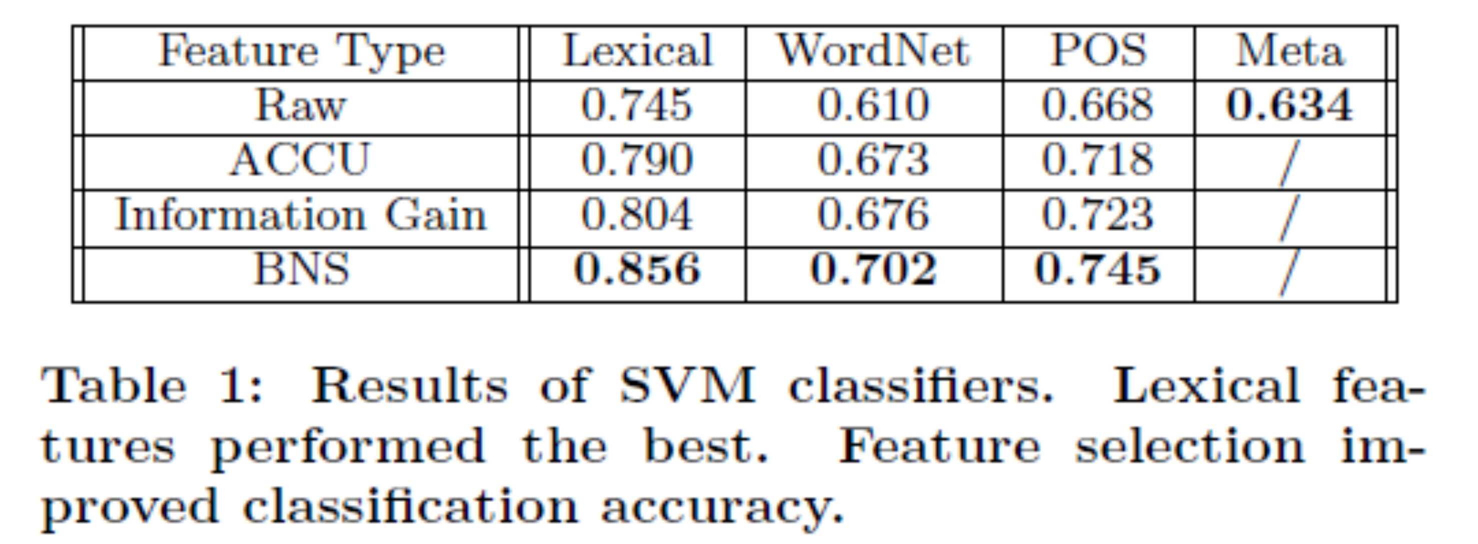


$$div_t = \frac{1}{N}\sum_{i=1}^{N} d_t(x_i)$$

$$d_t(x_i) = \begin{cases} 0 & \exists k, f_k(x_i) = f_t(x_i) \\ 1 & \forall k, f_k(x_i) \neq f_t(x_i) \end{cases}$$

- The diversity of a classier represents how much new information it could provide to a group of classifiers that have already been trained in Adaboost.

EVALUATION OF THE CLASSIFIER



Feature Type	Lexical	WordNet	POS	Meta
Raw	0.745	0.610	0.668	**0.634**
ACCU	0.790	0.673	0.718	/
Information Gain	0.804	0.676	0.723	/
BNS	**0.856**	**0.702**	**0.745**	/

Table 1: Results of SVM classifiers. Lexical features performed the best. Feature selection improved classification accuracy.

EVALUATION OF THE CLASSIFIER

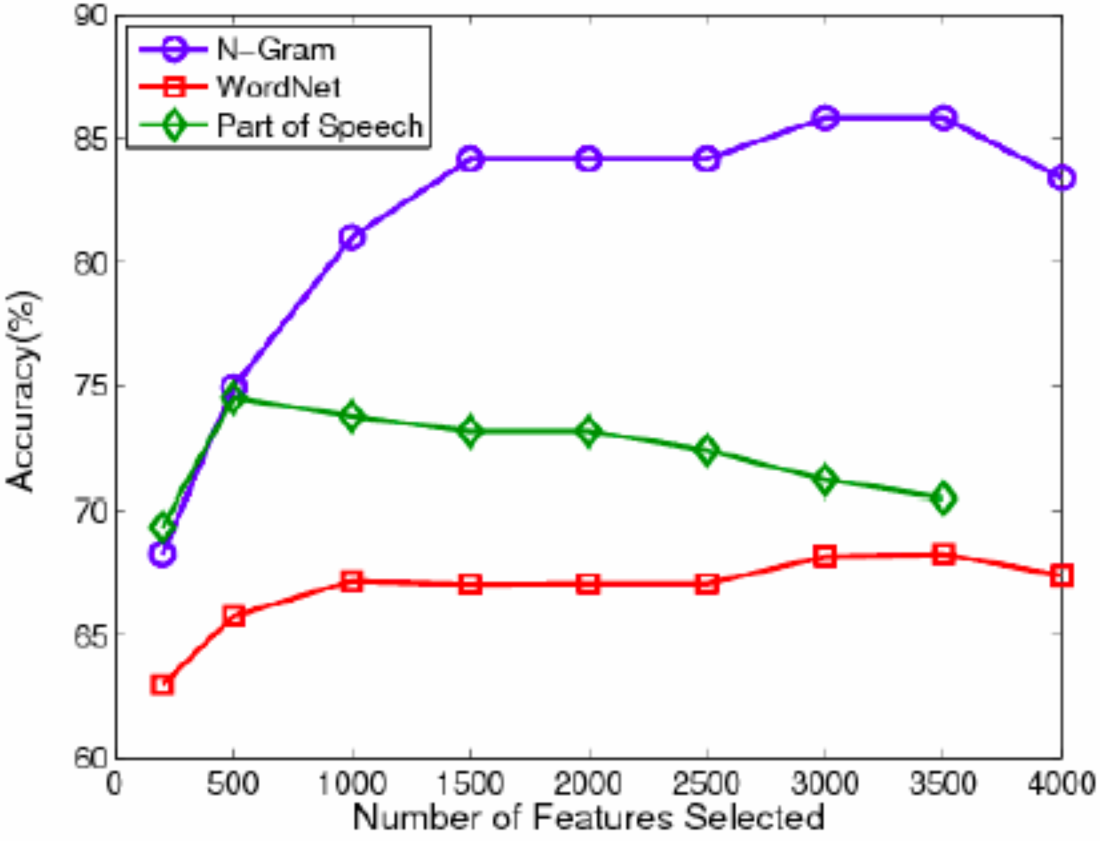


Figure 2: Feature selection using BNS

EVALUATION OF THE CLASSIFIER

- The accuracy of the classifier improved from 85.6% to 86.6%
- The small margin suggests that the lexical features are strong enough in detecting information needs, while other types of features add little to the success.

ANALYZING INFORMATION NEEDS

- 136,841,672 tweets conveying information need between July 10th 2011 to June 31st 2012.
- This is roughly a proportion of 3% of all tweets, and 28.6% of tweets with question marks.

GENERAL TREND

- first 5 months
- we normalize the time series so that the two curves are easier to be aligned on the plot

$$x_i' = \frac{x_i - \mu}{\sigma}$$

GENERAL TREND

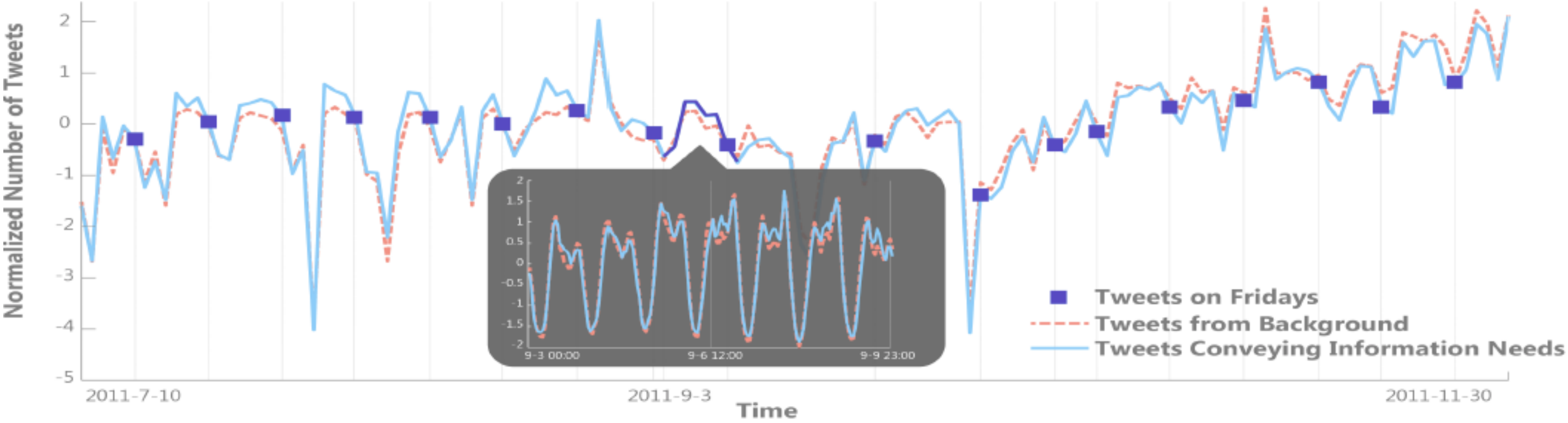
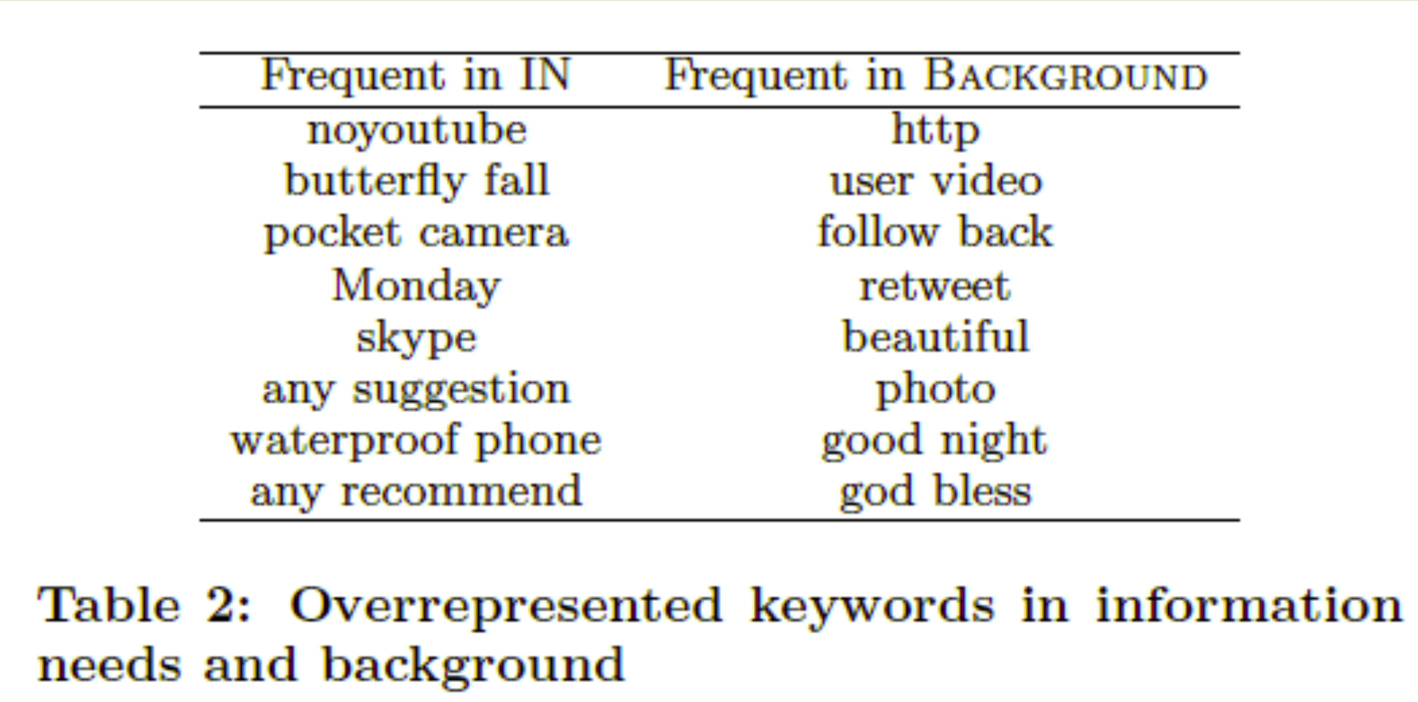


Figure 3: Questions and background tweets over time.

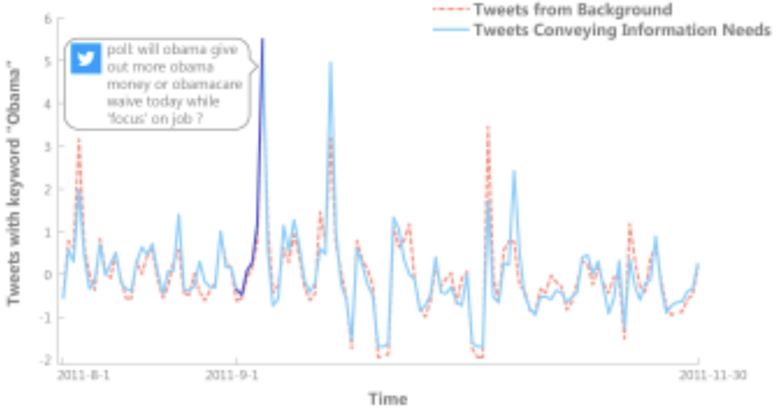
KEYWORDS

- what people ask.

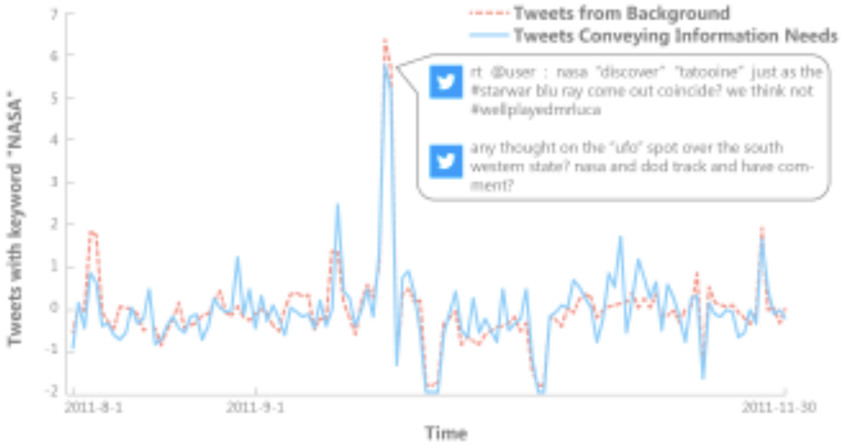


Frequent in IN	Frequent in Background
noyoutube	http
butterfly fall	user video
pocket camera	follow back
Monday	retweet
skype	beautiful
any suggestion	photo
waterproof phone	good night
any recommend	god bless

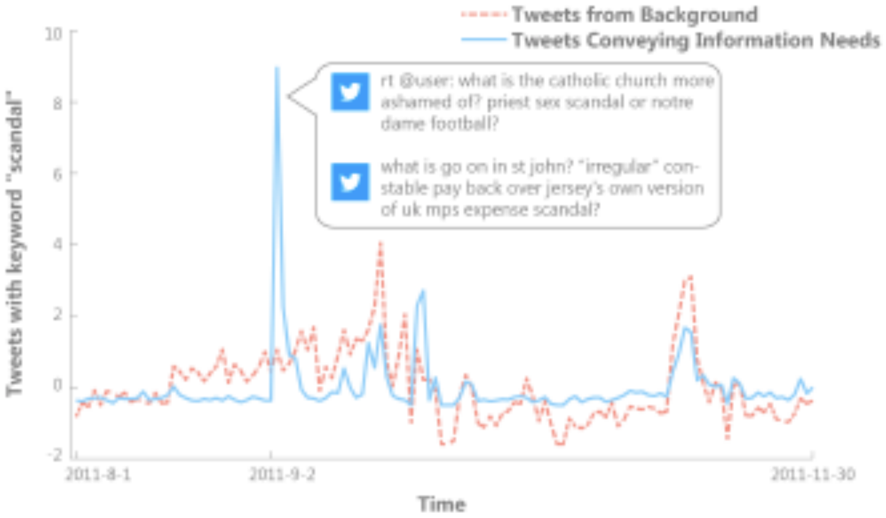
Table 2: Overrepresented keywords in information needs and background



(a) Trend of tweets conveying information need with keyword "obama"



(b) Trend of tweets conveying information need with keyword "nasa"



(c) Trend of tweets conveying information need with keyword "scandal"

BURSTINESS

- we adopt a straightforward solution to detect similar burst events in the time series of information needs and the background.

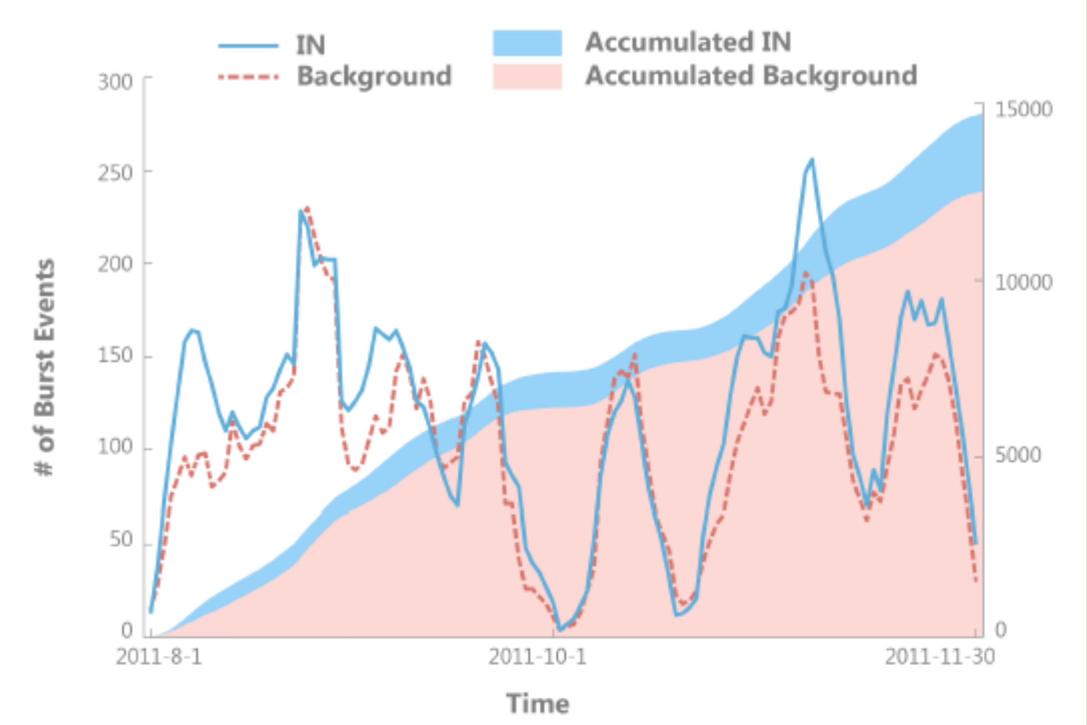
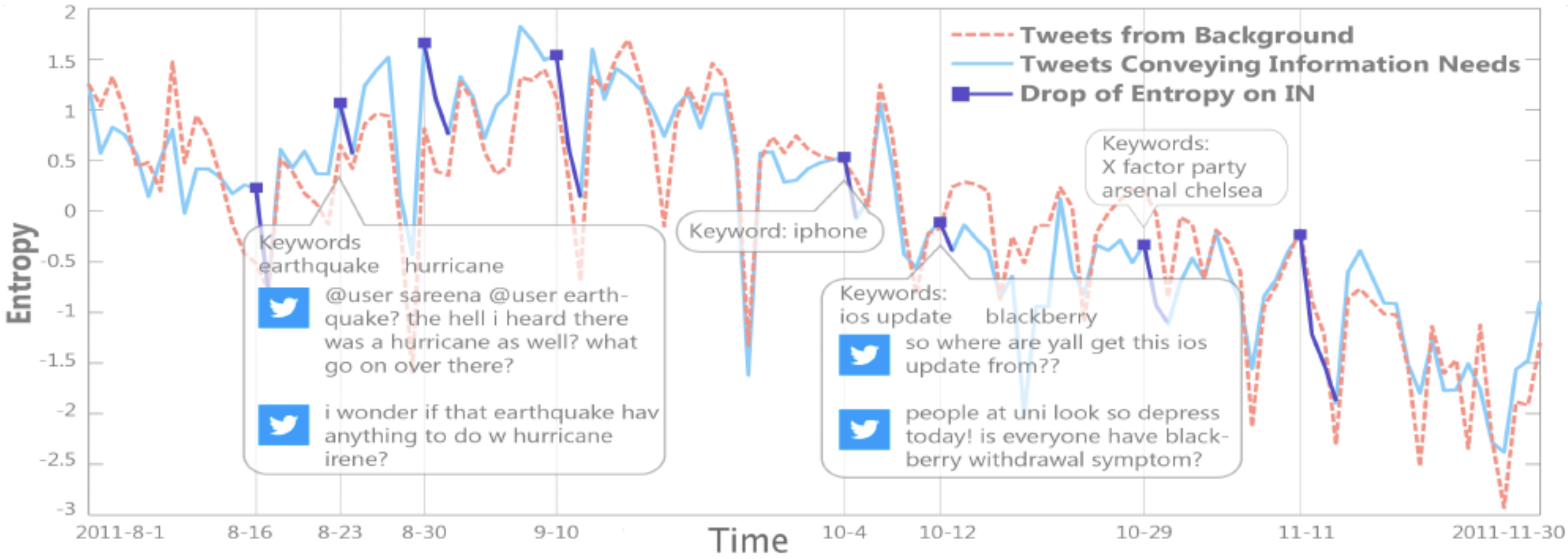


Figure 5: Bursts detected from IN and background

ENTROPY



@user sareena @user earthquake? the hell i heard there was a hurricane as well? what go on over there?

i wonder if that earthquake hav anything to do w hurricane irene?

so where are yall get this ios update from??

people at uni look so depress today! is everyone have blackberry withdrawal symptom?

Figure 6: Entropy of word distributions in questions and background

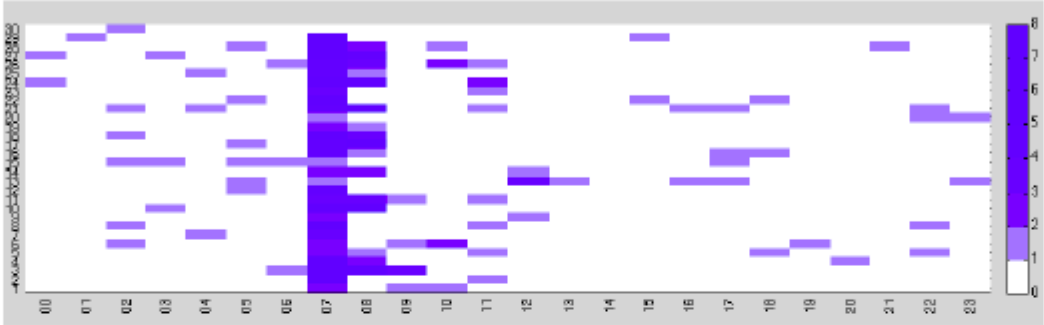


Figure 7: Questions of a user of low entropy.

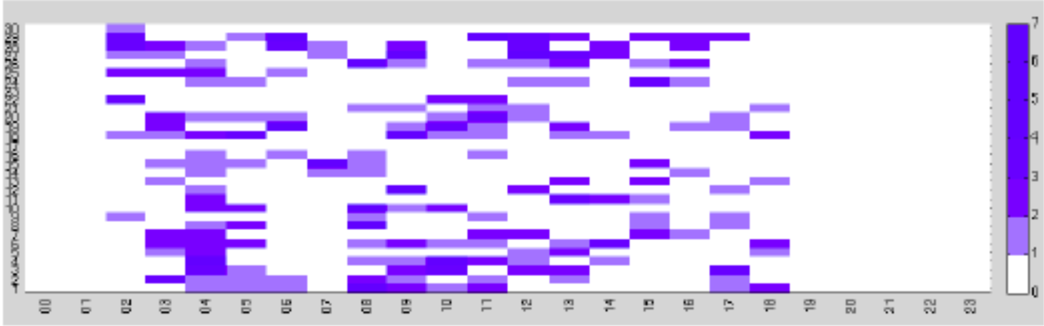
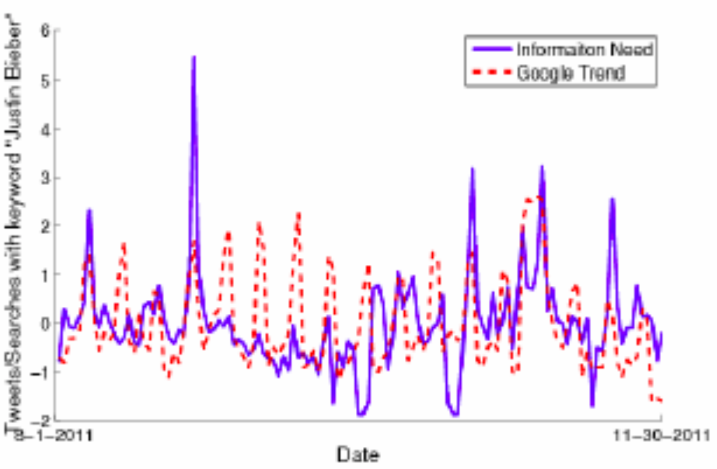
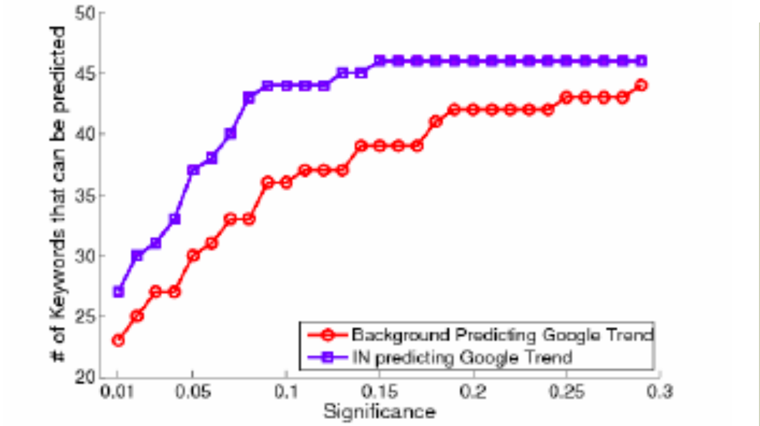


Figure 8: Questions from a user of high entropy.

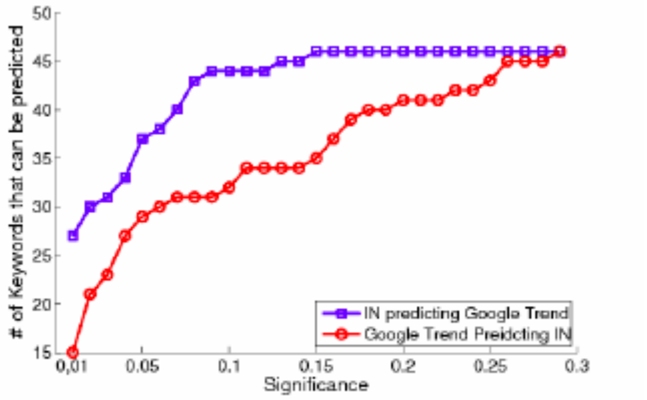
PREDICTIVE POWER



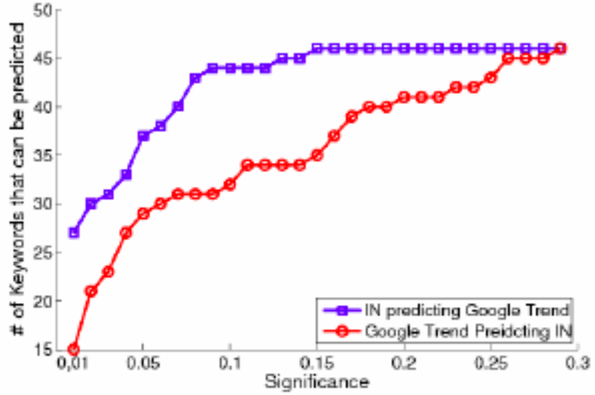
(a) Keyword: Justin Bieber



(b) Background v.s. information needs in predicting Google trends. The higher the better



(c) Information need v.s. Google trends in predicting each other. The higher the better



(c) Information need v.s. Google trends in predicting each other. The higher the better

CONCLUSION

- we present the first large-scale analysis of information needs, or questions, in Twitter.
- We proposed an automatic classification algorithm that distinguishes real questions from tweets with question marks

- We then present a comprehensive analysis of the large-scale collection of information needs we extracted.

ON PARTICIPATIONIN GROUP CHATS ON TWITTER

ABSTRACT

- To predict whether a user that attended her first session in a particular Twitter chat group will return to the group, we build 5F Model that captures five different factors: individual initiative, group characteristics, perceived receptivity, linguistic affinity, and geographical proximity .

- The research question we investigate is what factors ensure continued individual participation in a Twitter chat.

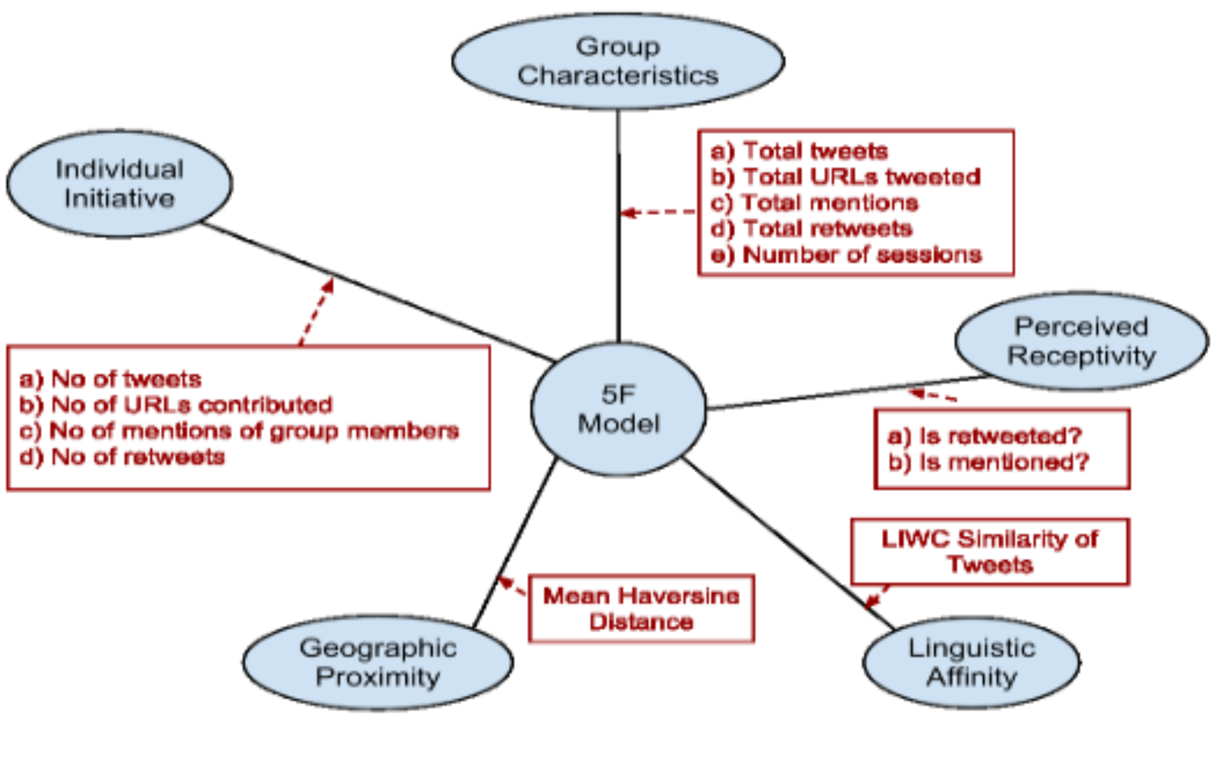


Figure 1: Overview of the *5F Model*

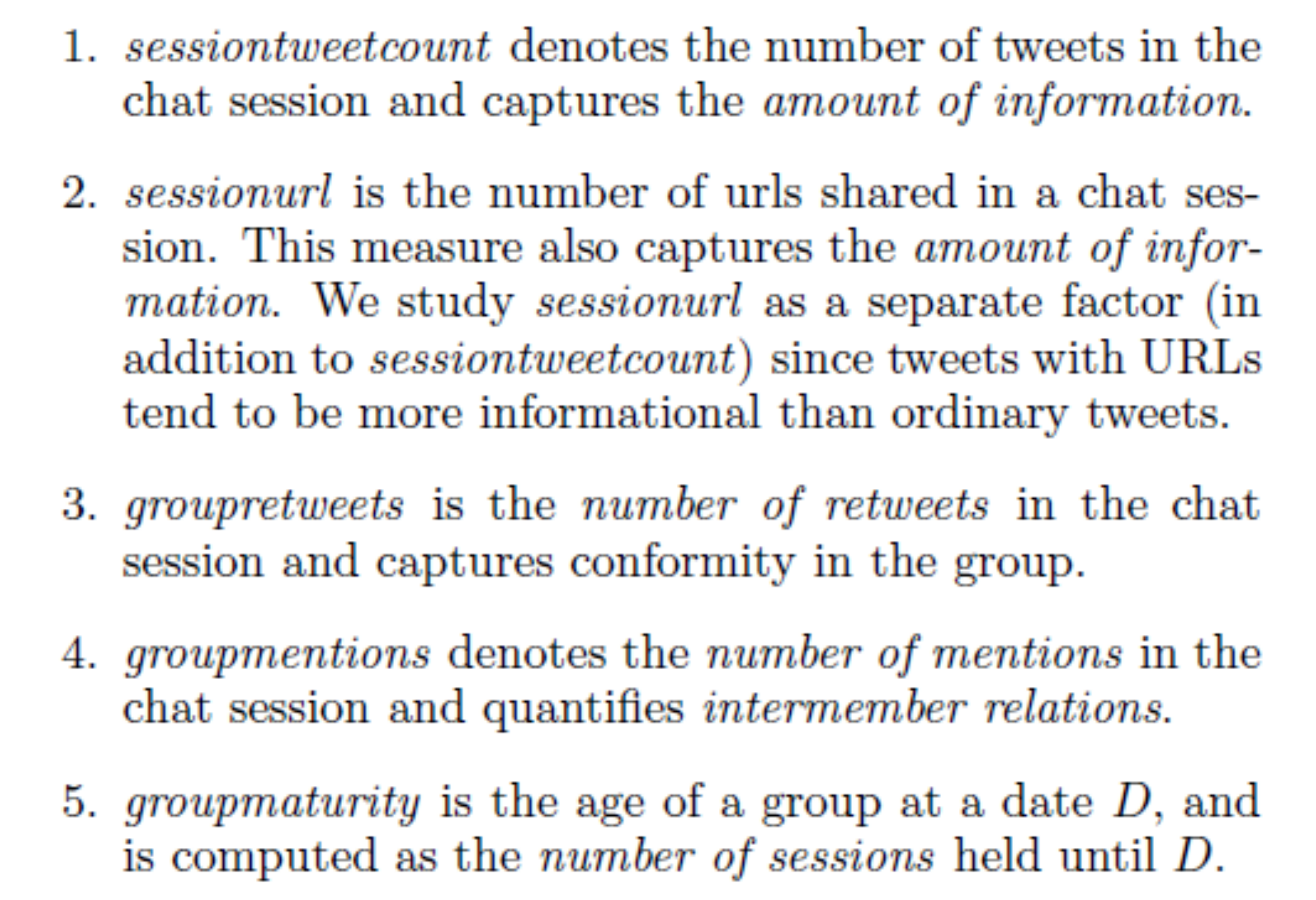
5F MODEL

- Individual Initiative

1. *usertweetcount* denotes the number of tweets the user contributes to the session.
2. *userurl* denotes the number of urls the user contributes to the chat session.
3. *usermentions* is the total number of times the user mentions another (by using @).
4. *userretweets* is the number of retweets by the newcomer user and captures the amount of information she found to be worth sharing with her followers.

5F MODEL

- Group Characteristics



1. *sessiontweetcount* denotes the number of tweets in the chat session and captures the *amount of information.*

2. *sessionurl* is the number of urls shared in a chat session. This measure also captures the *amount of information.* We study *sessionurl* as a separate factor (in addition to *sessiontweetcount*) since tweets with URLs tend to be more informational than ordinary tweets.

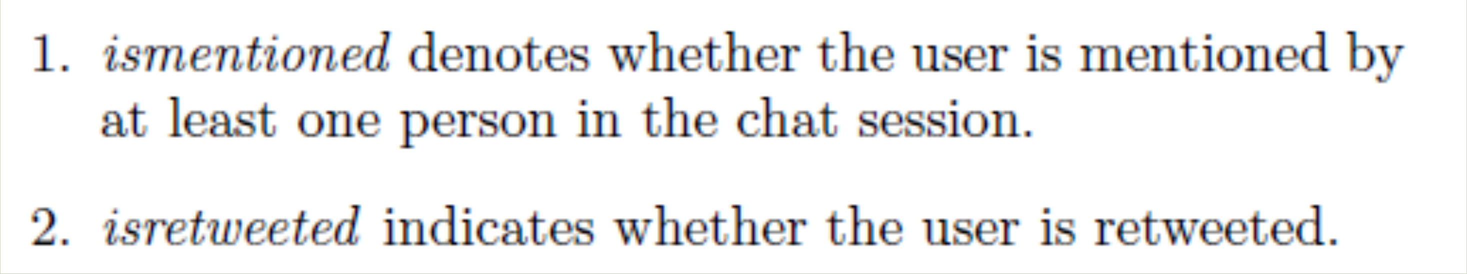
3. *groupretweets* is the *number of retweets* in the chat session and captures conformity in the group.

4. *groupmentions* denotes the *number of mentions* in the chat session and quantifies *intermember relations.*

5. *groupmaturity* is the age of a group at a date D, and is computed as the *number of sessions* held until D.

5F MODEL

- Perceived Receptivity



1. *ismentioned* denotes whether the user is mentioned by at least one person in the chat session.
2. *isretweeted* indicates whether the user is retweeted.

5F MODEL

- Linguistic Affinity

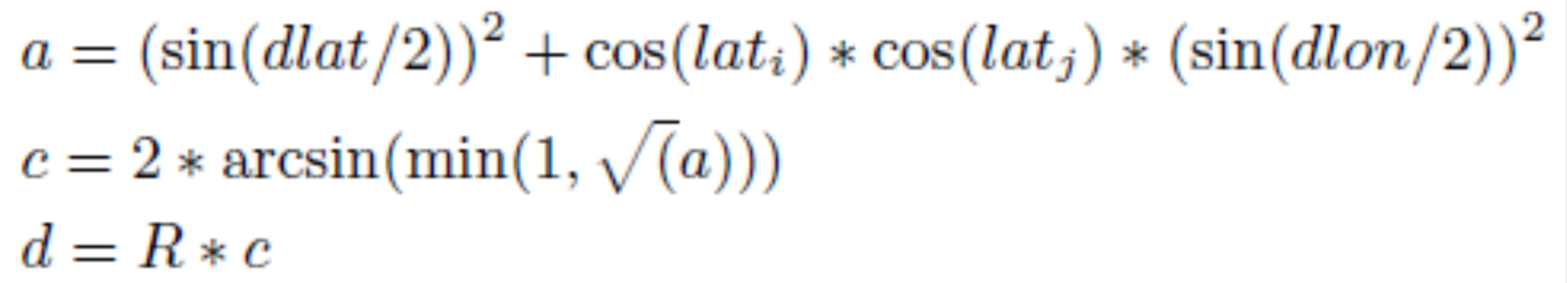
- We make use of *Linguistic Inquiry and Word Count (LIWC)* to compare linguistic markers between a user and a group.

- *LIWC* is a text analysis software that calculates the degree to which people use different categories of words across a wide array of texts

- We consider the set of tweets a user ui shares in her first session as a text document and compute the value of each linguistic marker to obtain her *LIWC-vector* for that particular session.

GEOGRAPHIC PROXIMITY

- the distance d (in meters) between two users ui and uj



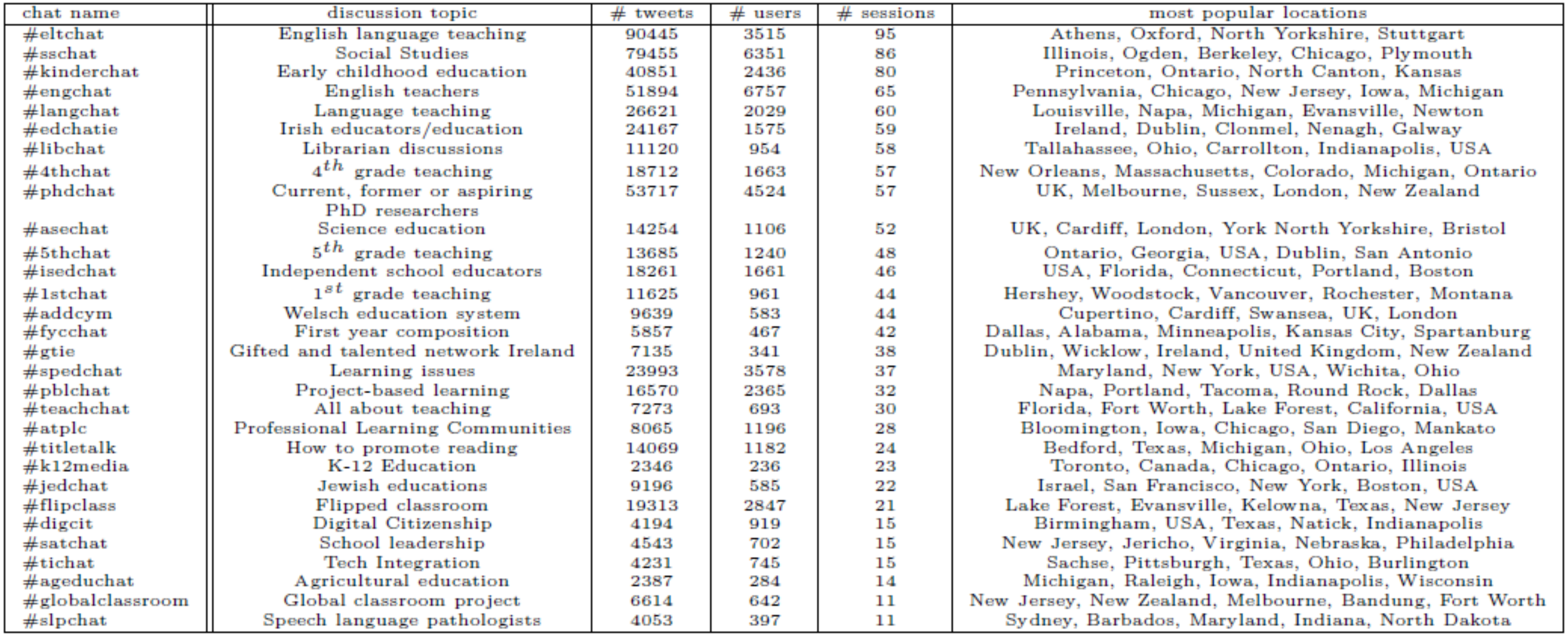
$$a = (\sin(dlat/2))^2 + \cos(lat_i) * \cos(lat_j) * (\sin(dlon/2))^2$$

$$c = 2 * \arcsin(\min(1, \sqrt{}(a)))$$

$$d = R * c$$

DATA SET

- **Group Chats Studied**
- June 2010-July 2012.
- By identifying the hours of high activity, we capture the sessions for each chat

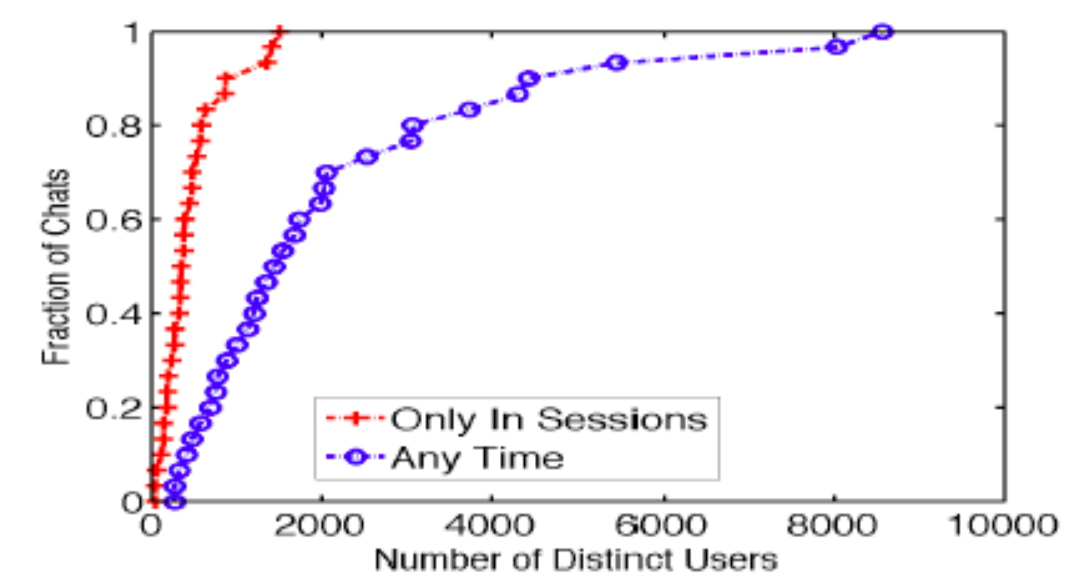


chat name	discussion topic	# tweets	# users	# sessions	most popular locations
#eltchat	English language teaching	90445	3515	95	Athens, Oxford, North Yorkshire, Stuttgart
#sschat	Social Studies	79455	6351	86	Illinois, Ogden, Berkeley, Chicago, Plymouth
#kinderchat	Early childhood education	40851	2436	80	Princeton, Ontario, North Canton, Kansas
#engchat	English teachers	51894	6757	65	Pennsylvania, Chicago, New Jersey, Iowa, Michigan
#langchat	Language teaching	26621	2029	60	Louisville, Napa, Michigan, Evansville, Newton
#edchatie	Irish educators/education	24167	1575	59	Ireland, Dublin, Clonmel, Nenagh, Galway
#libchat	Librarian discussions	11120	954	58	Tallahassee, Ohio, Carrollton, Indianapolis, USA
#4thchat	4^{th} grade teaching	18712	1663	57	New Orleans, Massachusetts, Colorado, Michigan, Ontario
#phdchat	Current, former or aspiring PhD researchers	53717	4524	57	UK, Melbourne, Sussex, London, New Zealand
#asechat	Science education	14254	1106	52	UK, Cardiff, London, York North Yorkshire, Bristol
#5thchat	5^{th} grade teaching	13685	1240	48	Ontario, Georgia, USA, Dublin, San Antonio
#isedchat	Independent school educators	18261	1661	46	USA, Florida, Connecticut, Portland, Boston
#1stchat	1^{st} grade teaching	11625	961	44	Hershey, Woodstock, Vancouver, Rochester, Montana
#addcym	Welsch education system	9639	583	44	Cupertino, Cardiff, Swansea, UK, London
#fycchat	First year composition	5857	467	42	Dallas, Alabama, Minneapolis, Kansas City, Spartanburg
#gtie	Gifted and talented network Ireland	7135	341	38	Dublin, Wicklow, Ireland, United Kingdom, New Zealand
#spedchat	Learning issues	23993	3578	37	Maryland, New York, USA, Wichita, Ohio
#pblchat	Project-based learning	16570	2365	32	Napa, Portland, Tacoma, Round Rock, Dallas
#teachchat	All about teaching	7273	693	30	Florida, Fort Worth, Lake Forest, California, USA
#atplc	Professional Learning Communities	8065	1196	28	Bloomington, Iowa, Chicago, San Diego, Mankato
#titletalk	How to promote reading	14069	1182	24	Bedford, Texas, Michigan, Ohio, Los Angeles
#k12media	K-12 Education	2346	236	23	Toronto, Canada, Chicago, Ontario, Illinois
#jedchat	Jewish educations	9196	585	22	Israel, San Francisco, New York, Boston, USA
#flipclass	Flipped classroom	19313	2847	21	Lake Forest, Evansville, Kelowna, Texas, New Jersey
#digcit	Digital Citizenship	4194	919	15	Birmingham, USA, Texas, Natick, Indianapolis
#satchat	School leadership	4543	702	15	New Jersey, Jericho, Virginia, Nebraska, Philadelphia
#tichat	Tech Integration	4231	745	15	Sachse, Pittsburgh, Texas, Ohio, Burlington
#ageduchat	Agricultural education	2387	284	14	Michigan, Raleigh, Iowa, Indianapolis, Wisconsin
#globalclassroom	Global classroom project	6614	642	11	New Jersey, New Zealand, Melbourne, Bandung, Fort Worth
#slpchat	Speech language pathologists	4053	397	11	Sydney, Barbados, Maryland, Indiana, North Dakota

Table 1: Education Chats Studied

SALIENT STATISTICS

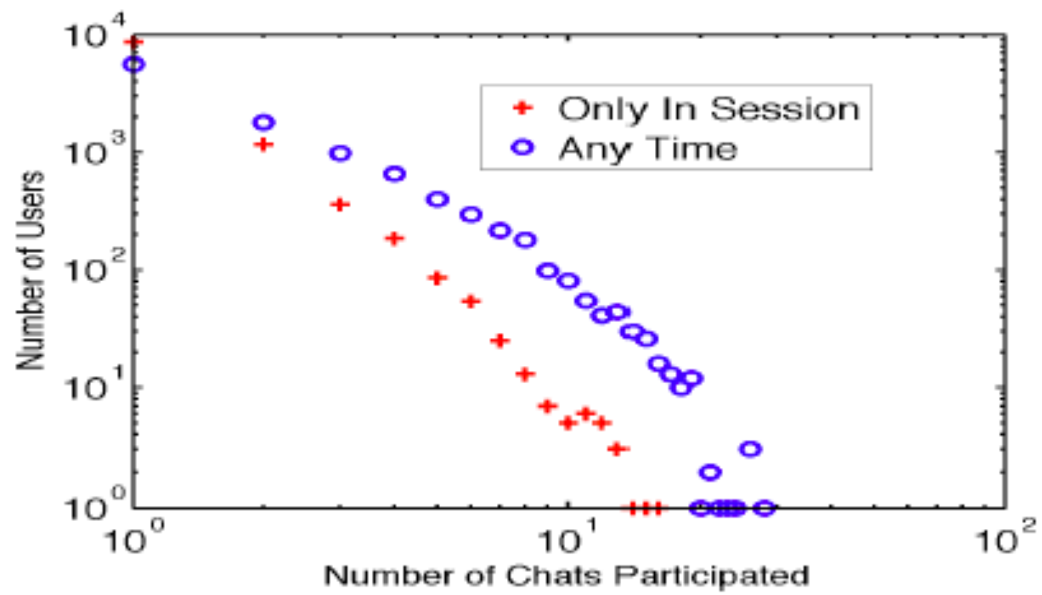
- Distribution of the number of users in and outside chat sessions:



(a) CDF of Number of users in per chat

SALIENT STATISTICS

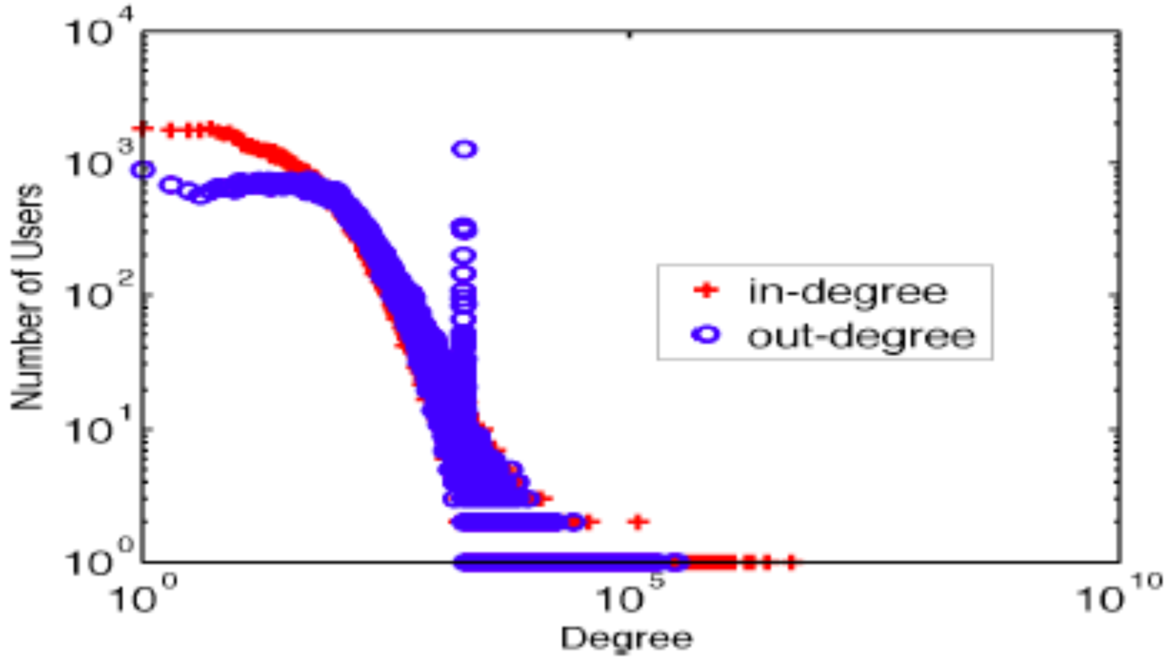
- Distribution of the number of distinct chats users
- participate in:



(b) Log-log scale plot of the number of chats per user

SALIENT STATISTICS

- Degree distribution of education chat users:



(c) Log-log scale plot of degree distribution

SALIENT STATISTICS

- Geographical distribution of education chat users

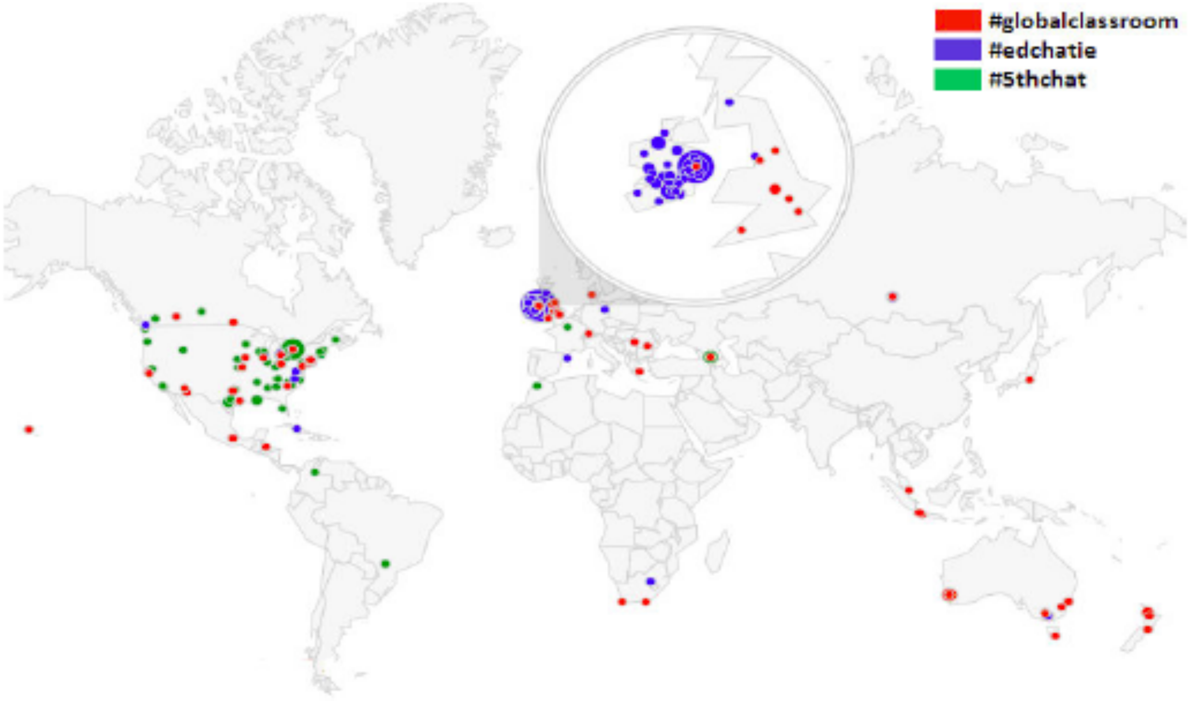
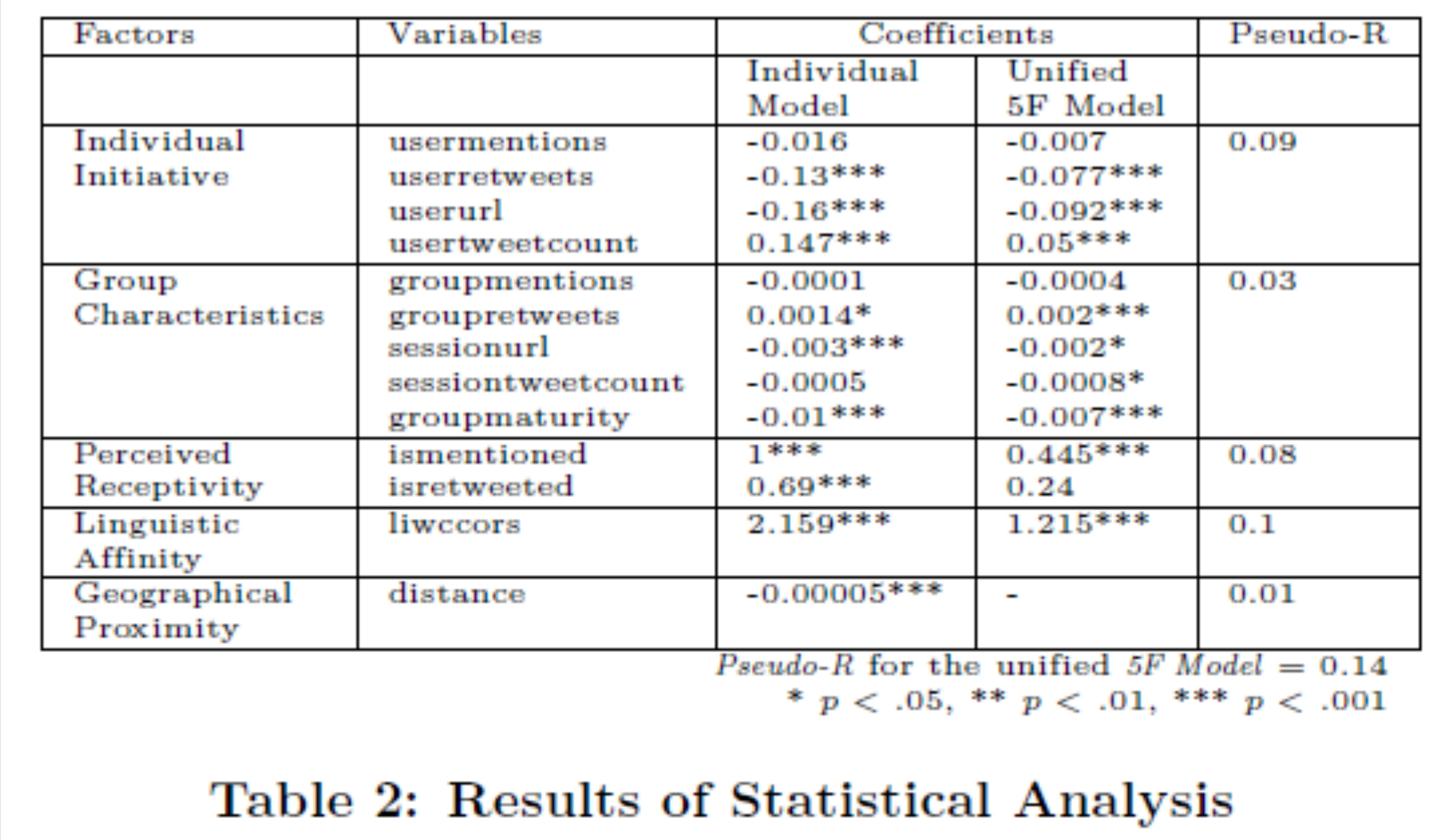


Figure 3: Geographical Distribution of Three Chats

STATISTICAL ANALYSIS



Factors	Variables	Coefficients		Pseudo-R
		Individual Model	Unified 5F Model	
Individual Initiative	usermentions	-0.016	-0.007	0.09
	userretweets	-0.13***	-0.077***	
	userurl	-0.16***	-0.092***	
	usertweetcount	0.147***	0.05***	
Group Characteristics	groupmentions	-0.0001	-0.0004	0.03
	groupretweets	0.0014*	0.002***	
	sessionurl	-0.003***	-0.002*	
	sessiontweetcount	-0.0005	-0.0008*	
	groupmaturity	-0.01***	-0.007***	
Perceived Receptivity	ismentioned	1***	0.445***	0.08
	isretweeted	0.69***	0.24	
Linguistic Affinity	liwccors	2.159***	1.215***	0.1
Geographical Proximity	distance	-0.00005***	-	0.01

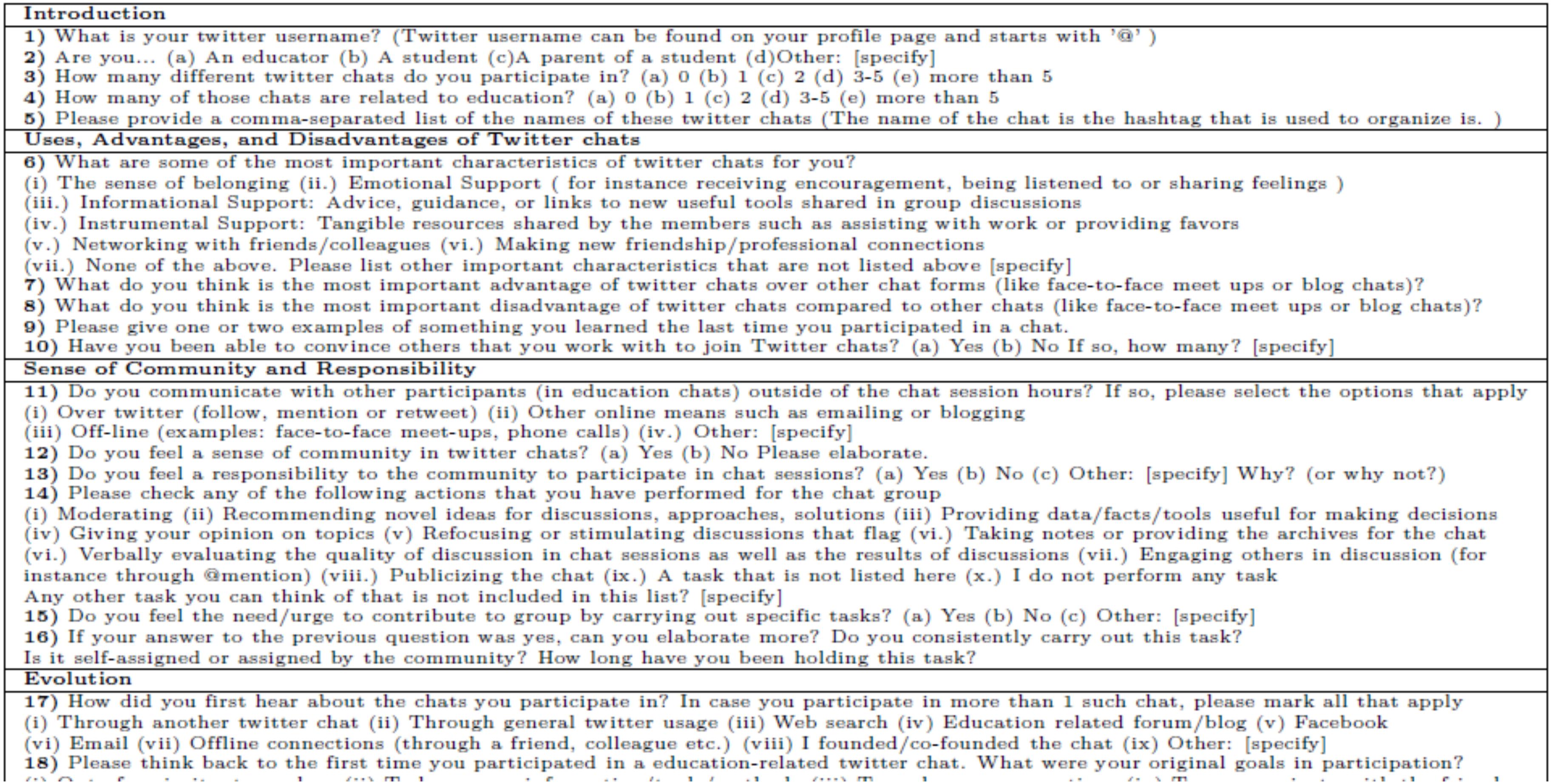
Pseudo-R for the unified *5F Model* = 0.14

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2: Results of Statistical Analysis

USER SURVEY

- an online survey of 26 questions



Introduction

1) What is your twitter username? (Twitter username can be found on your profile page and starts with '@')
2) Are you... (a) An educator (b) A student (c)A parent of a student (d)Other: [specify]
3) How many different twitter chats do you participate in? (a) 0 (b) 1 (c) 2 (d) 3-5 (e) more than 5
4) How many of those chats are related to education? (a) 0 (b) 1 (c) 2 (d) 3-5 (e) more than 5
5) Please provide a comma-separated list of the names of these twitter chats (The name of the chat is the hashtag that is used to organize is.)

Uses, Advantages, and Disadvantages of Twitter chats

6) What are some of the most important characteristics of twitter chats for you?
(i) The sense of belonging (ii.) Emotional Support (for instance receiving encouragement, being listened to or sharing feelings)
(iii.) Informational Support: Advice, guidance, or links to new useful tools shared in group discussions
(iv.) Instrumental Support: Tangible resources shared by the members such as assisting with work or providing favors
(v.) Networking with friends/colleagues (vi.) Making new friendship/professional connections
(vii.) None of the above. Please list other important characteristics that are not listed above [specify]
7) What do you think is the most important advantage of twitter chats over other chat forms (like face-to-face meet ups or blog chats)?
8) What do you think is the most important disadvantage of twitter chats compared to other chats (like face-to-face meet ups or blog chats)?
9) Please give one or two examples of something you learned the last time you participated in a chat.
10) Have you been able to convince others that you work with to join Twitter chats? (a) Yes (b) No If so, how many? [specify]

Sense of Community and Responsibility

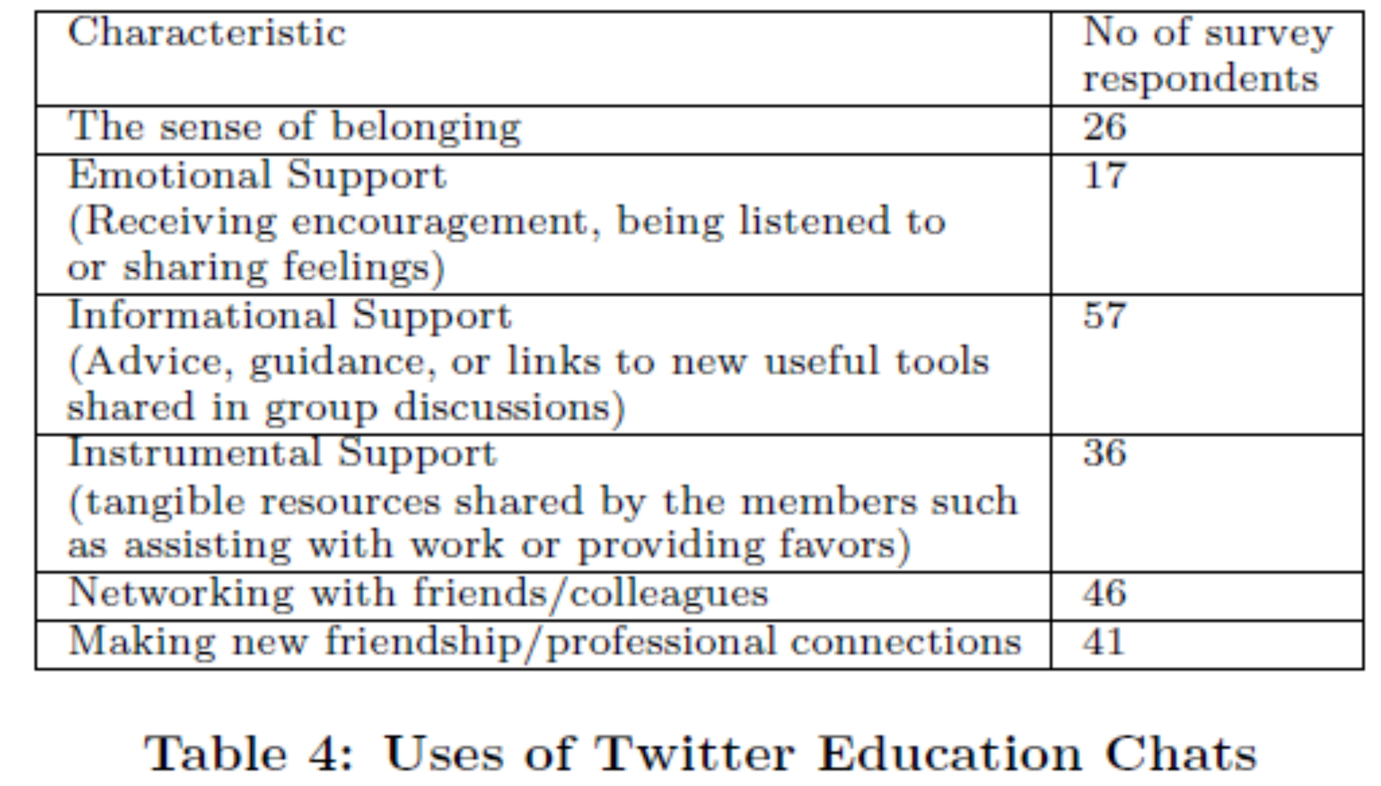
11) Do you communicate with other participants (in education chats) outside of the chat session hours? If so, please select the options that apply
(i) Over twitter (follow, mention or retweet) (ii) Other online means such as emailing or blogging
(iii) Off-line (examples: face-to-face meet-ups, phone calls) (iv.) Other: [specify]
12) Do you feel a sense of community in twitter chats? (a) Yes (b) No Please elaborate.
13) Do you feel a responsibility to the community to participate in chat sessions? (a) Yes (b) No (c) Other: [specify] Why? (or why not?)
14) Please check any of the following actions that you have performed for the chat group
(i) Moderating (ii) Recommending novel ideas for discussions, approaches, solutions (iii) Providing data/facts/tools useful for making decisions
(iv) Giving your opinion on topics (v) Refocusing or stimulating discussions that flag (vi.) Taking notes or providing the archives for the chat
(vi.) Verbally evaluating the quality of discussion in chat sessions as well as the results of discussions (vii.) Engaging others in discussion (for instance through @mention) (viii.) Publicizing the chat (ix.) A task that is not listed here (x.) I do not perform any task
Any other task you can think of that is not included in this list? [specify]
15) Do you feel the need/urge to contribute to group by carrying out specific tasks? (a) Yes (b) No (c) Other: [specify]
16) If your answer to the previous question was yes, can you elaborate more? Do you consistently carry out this task?
Is it self-assigned or assigned by the community? How long have you been holding this task?

Evolution

17) How did you first hear about the chats you participate in? In case you participate in more than 1 such chat, please mark all that apply
(i) Through another twitter chat (ii) Through general twitter usage (iii) Web search (iv) Education related forum/blog (v) Facebook
(vi) Email (vii) Offline connections (through a friend, colleague etc.) (viii) I founded/co-founded the chat (ix) Other: [specify]
18) Please think back to the first time you participated in a education-related twitter chat. What were your original goals in participation?

USER SURVEY

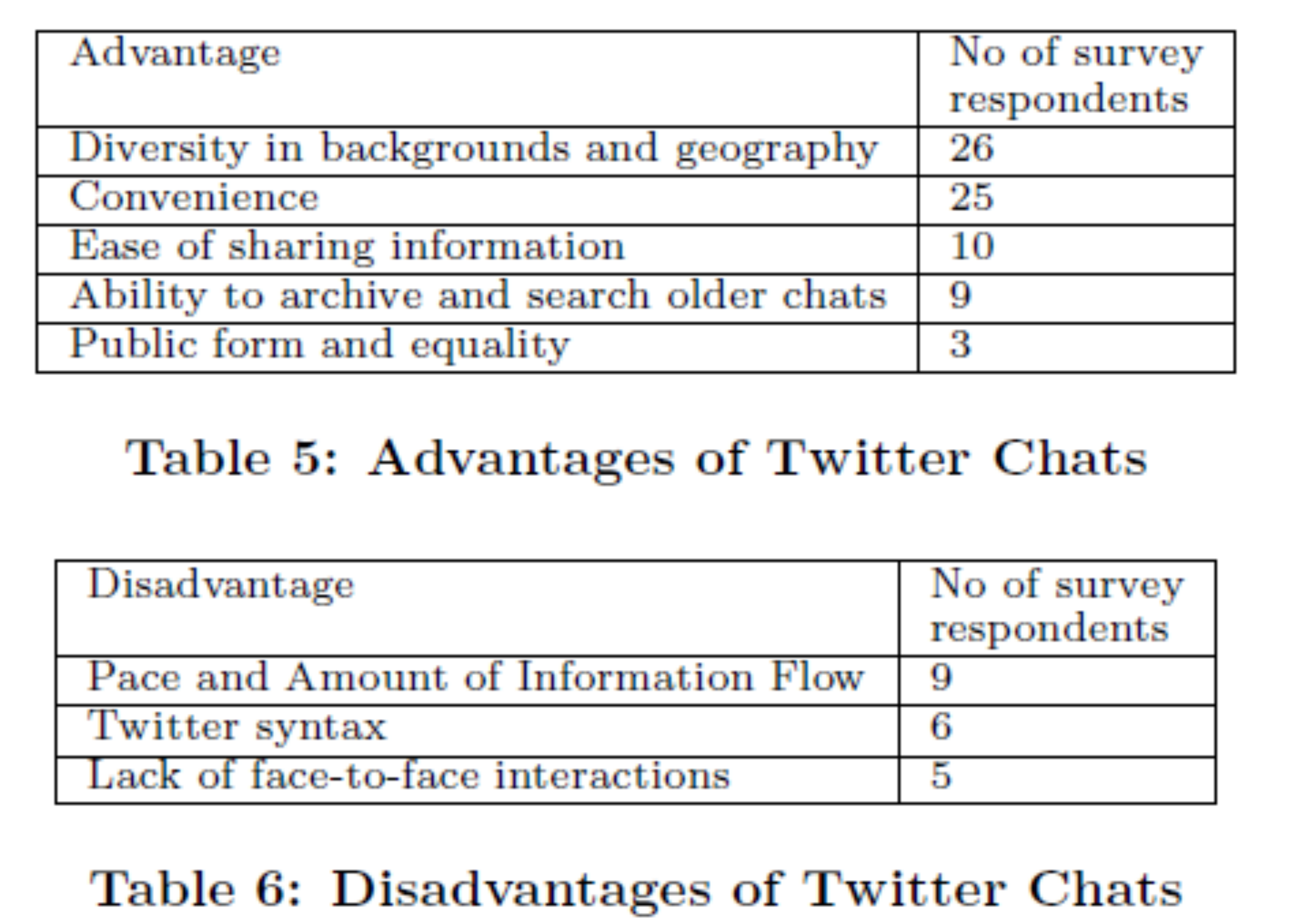
- Usage, Advantages and Disadvantages



Characteristic	No of survey respondents
The sense of belonging	26
Emotional Support (Receiving encouragement, being listened to or sharing feelings)	17
Informational Support (Advice, guidance, or links to new useful tools shared in group discussions)	57
Instrumental Support (tangible resources shared by the members such as assisting with work or providing favors)	36
Networking with friends/colleagues	46
Making new friendship/professional connections	41

Table 4: Uses of Twitter Education Chats

USER SURVEY



Advantage	No of survey respondents
Diversity in backgrounds and geography	26
Convenience	25
Ease of sharing information	10
Ability to archive and search older chats	9
Public form and equality	3

Table 5: Advantages of Twitter Chats

Disadvantage	No of survey respondents
Pace and Amount of Information Flow	9
Twitter syntax	6
Lack of face-to-face interactions	5

Table 6: Disadvantages of Twitter Chats

CONCLUSIONS

- We developed *5F Model* that predicts whether a person attending her first chat session in a particular Twitter chat group will return to the group.
- We performed statistical data analysis for thirty educational Twitter chats involving 71411 users and 730944 tweets over a period of two years.
- We also complemented the results of statistical analysis with a survey study.