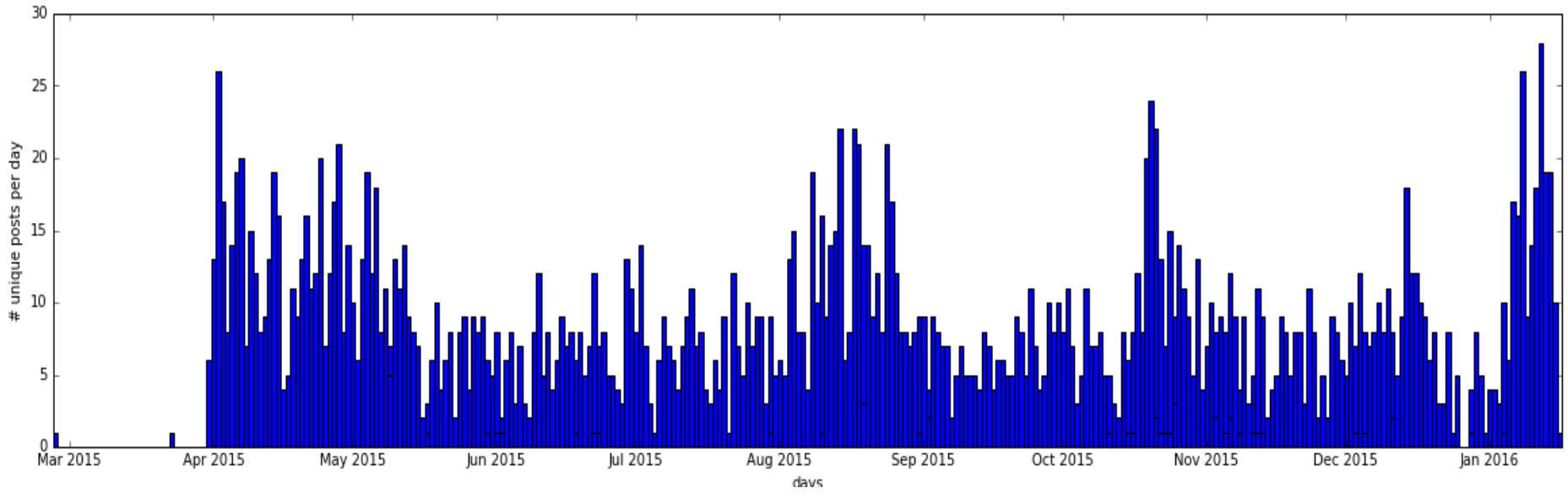
Assignment 1: Linguistic Analysis of Georgia Tech campus well-being

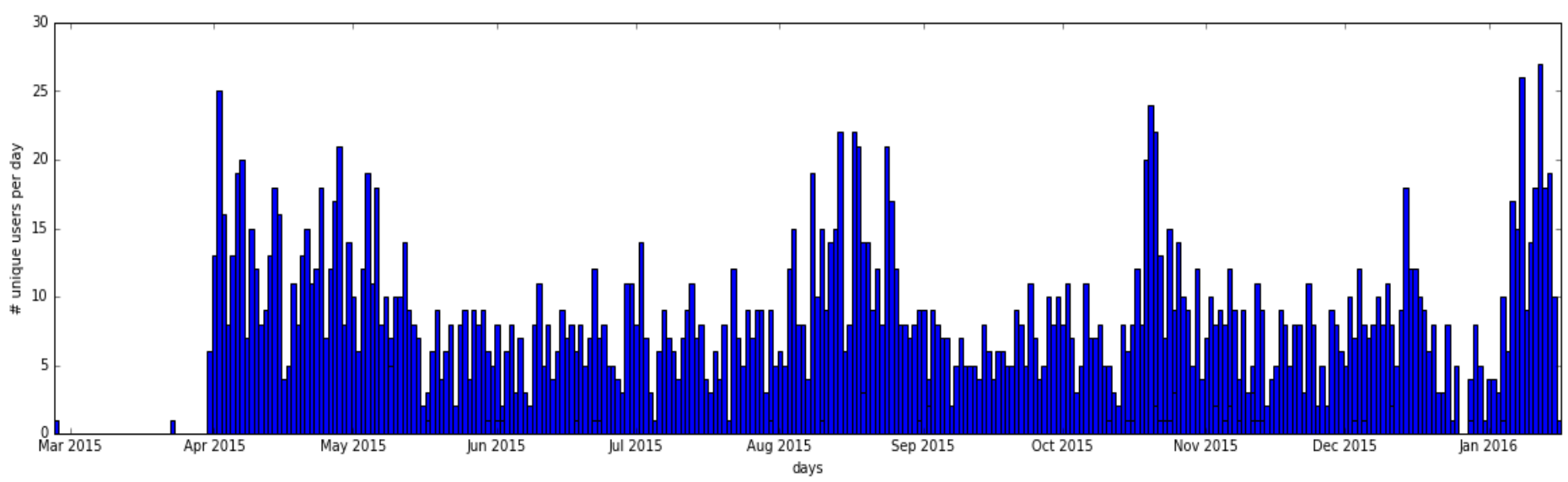
-

Sarthak Ghosh, 903048253

Descriptive Statistics

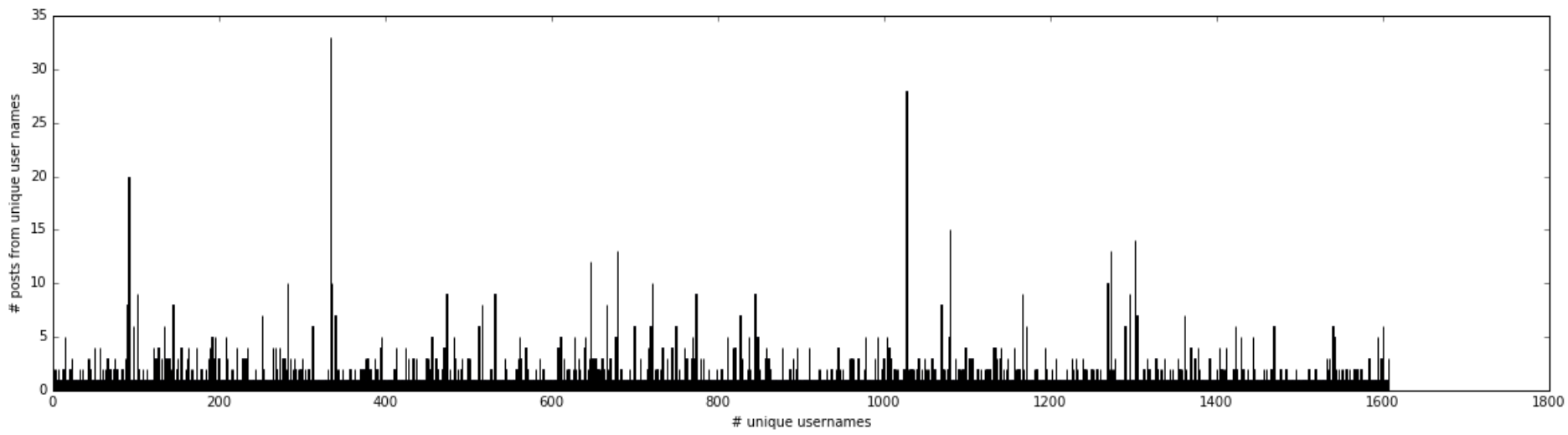
1)





Discussions

From the above given graphs we see a striking similarity between the distribution of the number of posts per day and the number of users per day. The volume of posts varies considerably depending upon which day it is. We see comparatively higher posts in April-May 2015, Aug-Sep 2015 and in Jan 2016. This pattern may be correlated to the times when semesters end/begin. Moreover there are not much differences in the number of posts per day and number of unique users per day. This hints at the fact that there are not too many overly active users. A plot of the user-names versus the number of posts they have (below) reveal that only a handful of users are extremely active on Reddit as compared to the others. There are 3 users who are outliers and about 13-15 who have moderate volumes of posts. All others are fairly average users of Reddit.



2)

Mean length of all posts: 67.621969697 words

Mean length of all titles: 6.88143939394 words

Standard deviation of length of posts: 81.0845301675

Standard deviation of length of all titles: 4.54872153743

-

3)

Bi-grams for posts below mean

bi-gram	raw freq	normalized freq
('i', 'am')	209	0.098678
('i', 'have')	197	0.0930123
('in', 'the')	186	0.0878187
('and', 'i')	169	0.0797923
('for', 'the')	156	0.0736544
('i', 'was')	143	0.0675165
('but', 'i')	137	0.0646837
('of', 'the')	125	0.0590179
('for', 'a')	121	0.0571294
('if', 'you')	118	0.0557129
('want', 'to')	114	0.0538244
('have', 'a')	114	0.0538244
('does', 'anyone')	113	0.0533522
('i', 'can')	113	0.0533522
('to', 'get')	109	0.0514636
('looking', 'for')	101	0.0476865
('is', 'there')	100	0.0472144
('on', 'the')	98	0.0462701
('wondering', 'if')	95	0.0448536
('to', 'take')	91	0.0429651
('anyone', 'know')	90	0.0424929
('so', 'i')	90	0.0424929

Bi-grams for posts above mean

bi-gram	raw freq	normalized freq
('i', 'am')	401	0.10227
('in', 'the')	349	0.0890079
('i', 'have')	338	0.0862025
('of', 'the')	308	0.0785514
('and', 'i')	251	0.0640143
('for', 'the')	215	0.054833
('if', 'you')	206	0.0525376
('want', 'to')	205	0.0522826
('but', 'i')	194	0.0494772
('on', 'the')	192	0.0489671
('to', 'be')	184	0.0469268
('for', 'a')	180	0.0459067
('have', 'a')	178	0.0453966
('i', 'was')	165	0.0420811
('to', 'get')	161	0.041061
('to', 'the')	160	0.0408059
('if', 'i')	158	0.0402958
('will', 'be')	153	0.0390207
('would', 'be')	152	0.0387656
('i', 'would')	146	0.0372354
('that', 'i')	145	0.0369804
('so', 'i')	140	0.0357052
('georgia', 'tech')	133	0.0339199
('at', 'the')	130	0.0331548
('i', "don't")	129	0.0328998

tri-grams for posts below mean

tri-gram	raw freq	normalized freq
('was', 'wondering', 'if')	66	0.0311615
('does', 'anyone', 'know')	63	0.029745
('i', 'was', 'wondering')	61	0.0288008
('i', 'want', 'to')	54	0.0254958
('i', 'have', 'a')	45	0.0212465
('i', 'need', 'to')	44	0.0207743
('is', 'there', 'a')	38	0.0179415
('and', 'i', 'was')	35	0.016525
('looking', 'for', 'a')	34	0.0160529
('does', 'anyone', 'have')	32	0.0151086
('would', 'like', 'to')	31	0.0146364
('i', 'have', 'to')	30	0.0141643
('is', 'there', 'any')	29	0.0136922
('be', 'able', 'to')	29	0.0136922
('i', 'am', 'looking')	27	0.0127479
('are', 'there', 'any')	27	0.0127479
("i'm", 'trying', 'to')	26	0.0122757
('anyone', 'have', 'any')	24	0.0113314
("i'm", 'looking', 'for')	24	0.0113314
('wondering', 'if', 'anyone')	24	0.0113314
('if', 'you', 'have')	23	0.0108593
('if', 'you', 'are')	23	0.0108593
('anyone', 'know', 'if')	22	0.0103872
('let', 'me', 'know')	22	0.0103872
('a', 'lot', 'of')	22	0.0103872

tri-grams for posts above mean

tri-gram	raw freq	normalized freq
('i', 'want', 'to')	89	0.0226983
('a', 'lot', 'of')	69	0.0175976
('i', 'have', 'a')	68	0.0173425
('i', 'am', 'a')	52	0.0132619
('looking', 'for', 'a')	51	0.0130069
('be', 'able', 'to')	49	0.0124968
('if', 'you', 'have')	45	0.0114767
('i', 'need', 'to')	43	0.0109666
('if', 'you', 'are')	41	0.0104565
('i', 'was', 'wondering')	39	0.00994644
('would', 'like', 'to')	36	0.00918133
('i', "don't", 'know')	36	0.00918133
('one', 'of', 'the')	34	0.00867126
('going', 'to', 'be')	33	0.00841622
('i', 'am', 'looking')	31	0.00790615
('i', 'feel', 'like')	31	0.00790615
('i', 'have', 'to')	31	0.00790615
('i', 'will', 'be')	31	0.00790615
('was', 'wondering', 'if')	30	0.00765111
('and', 'i', 'am')	29	0.00739607
("i'd", 'like', 'to')	28	0.00714104
('the', 'end', 'of')	28	0.00714104
('does', 'anyone', 'know')	27	0.006886
('i', 'wanted', 'to')	27	0.006886
('to', 'get', 'a')	27	0.006886

uni-grams for posts below mean

uni-gram	raw freq	normalized freq
('i',)	2118	1
('to',)	1977	0.933428
('the',)	1923	0.907932
('a',)	1502	0.70916
('and',)	1320	0.623229
('for',)	1037	0.489613
('is',)	841	0.397073
('in',)	768	0.362606
('of',)	680	0.321058
('it',)	607	0.286591
('have',)	592	0.279509
('if',)	565	0.266761
('on',)	533	0.251653
('my',)	520	0.245515
("i'm",)	495	0.233711
('or',)	448	0.21152
('this',)	442	0.208687
('but',)	437	0.206327
('that',)	430	0.203022
('anyone',)	423	0.199717
('you',)	411	0.194051
('any',)	408	0.192635
('be',)	402	0.189802
('are',)	360	0.169972
('with',)	355	0.167611

uni-grams for posts above mean

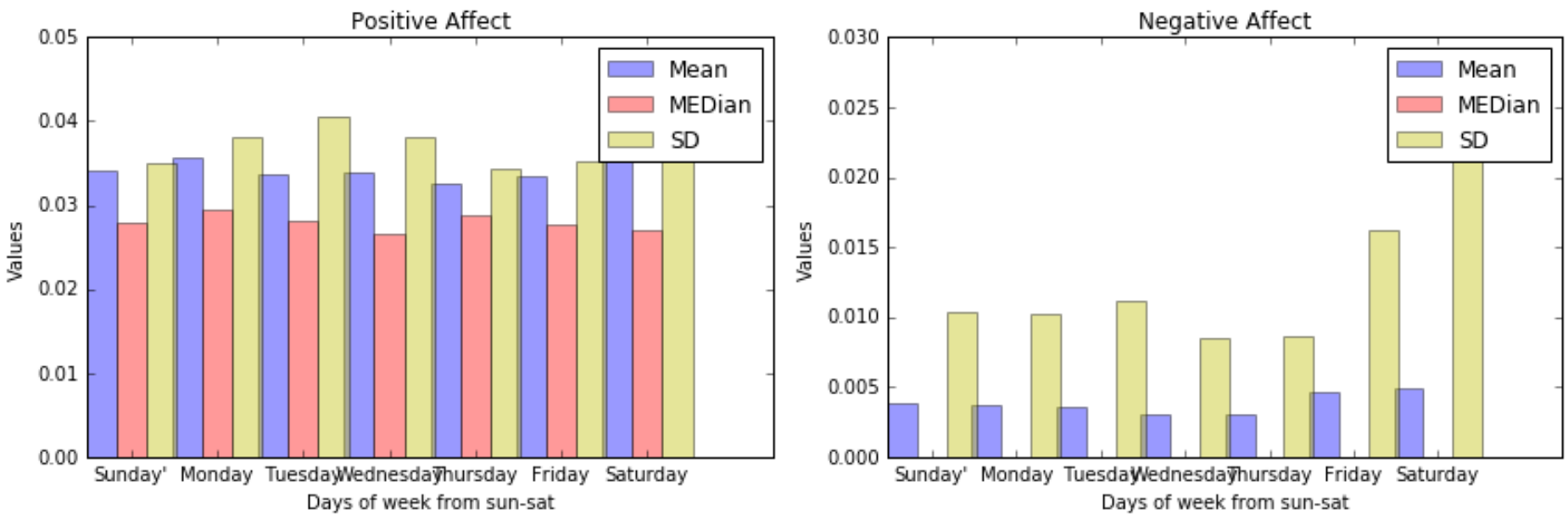
uni-gram	raw freq	normalized freq
('the',)	3921	1
('i',)	3712	0.946697
('to',)	3703	0.944402
('a',)	2962	0.75542
('and',)	2778	0.708493
('of',)	1793	0.457281
('for',)	1595	0.406784
('in',)	1587	0.404744
('is',)	1370	0.349401
('my',)	1192	0.304004
('that',)	1149	0.293037
('have',)	1002	0.255547
('on',)	930	0.237184
('be',)	912	0.232594
('it',)	892	0.227493
('you',)	872	0.222392
('this',)	865	0.220607
('if',)	803	0.204795
('but',)	800	0.20403
('with',)	734	0.187197
('at',)	723	0.184392
("i'm",)	698	0.178016
('or',)	661	0.168579
('are',)	620	0.158123
('as',)	593	0.151237

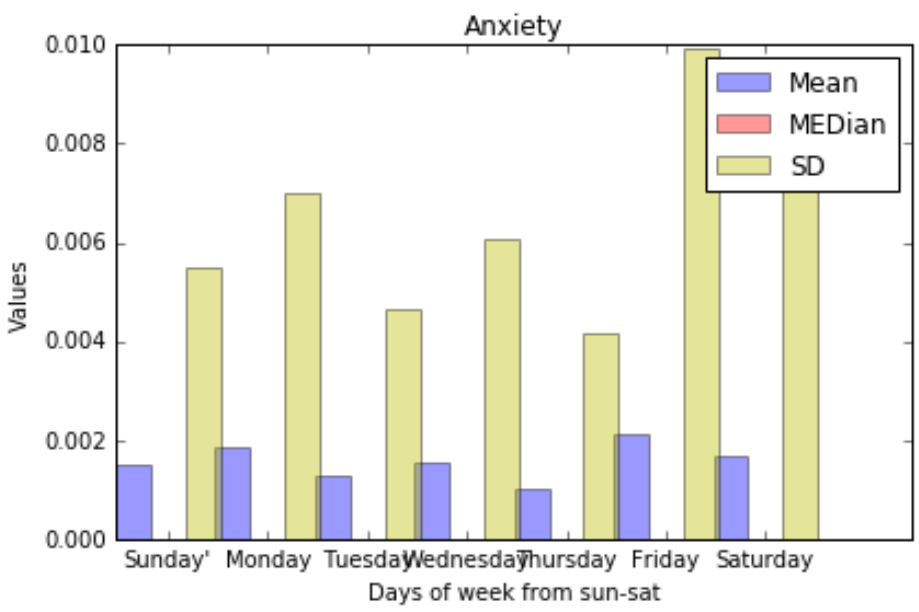
Discussions

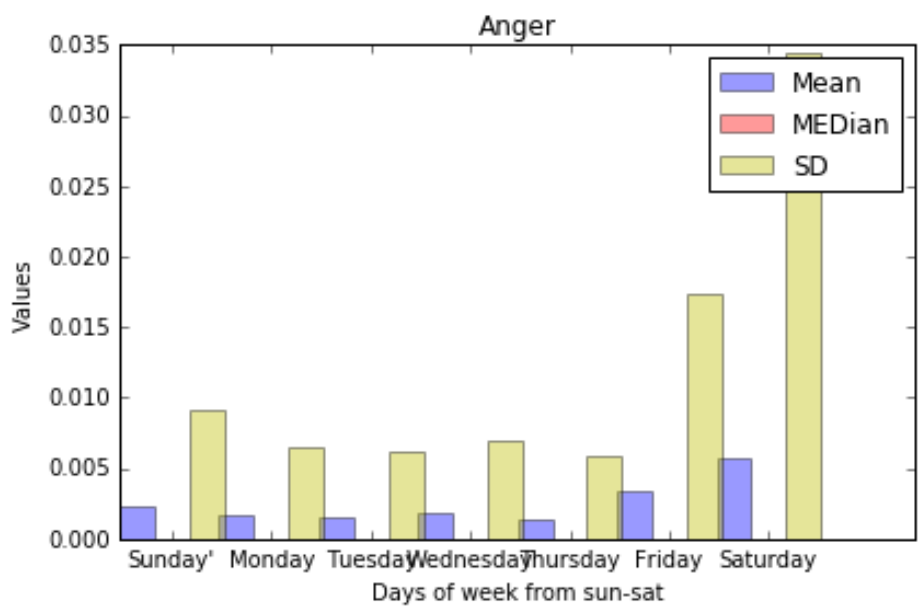
- Unigrams have highest raw frequencies in both the above mean-length posts and the below mean-length posts
- Most of the unigrams are articles like "The" "a" or words like "this", "that", "of" etc.
- In both pronoun "I" occurs high up in the unigram tables. This can hint at the fact that most posts on Reddit are ego-centric and personal.
- While several common bi-grams can be seen in the tables for posts above mean-length and posts below mean-length, posts below mean length feature bi-grams like " does anyone", "looking for", "wondering if" and "is there". On the other hand, in the posts above the mean length, we see bigrams like "I was", "I am", "will be", "would be", "Georgia tech" etc. These differences indicate that shorter posts are more of personal questions or concerns, whereas longer posts seem to be more of revealing of personal opinions and statements.
- A similar observation can be made from the tri-gram tables. Tri-grams like "was wondering if", "does anyone know" rank high up in the table for posts below mean length. This reaffirms the fact that shorter posts include more questions and personal musings. On the other hand, longer posts with trigrams like "I have a", "I feel like", "I have to", include more candid conversational statements. Longer posts may indicate more candid and honest discourse.

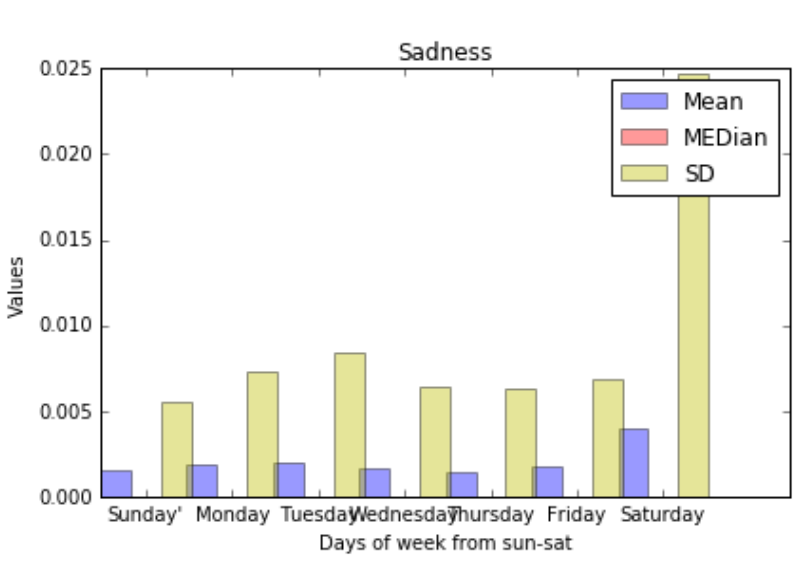
Campus Affect

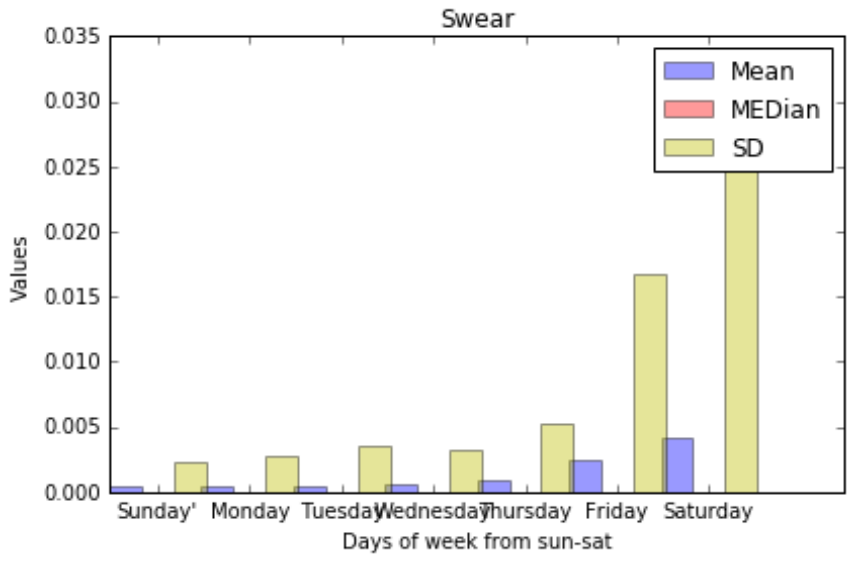
1)

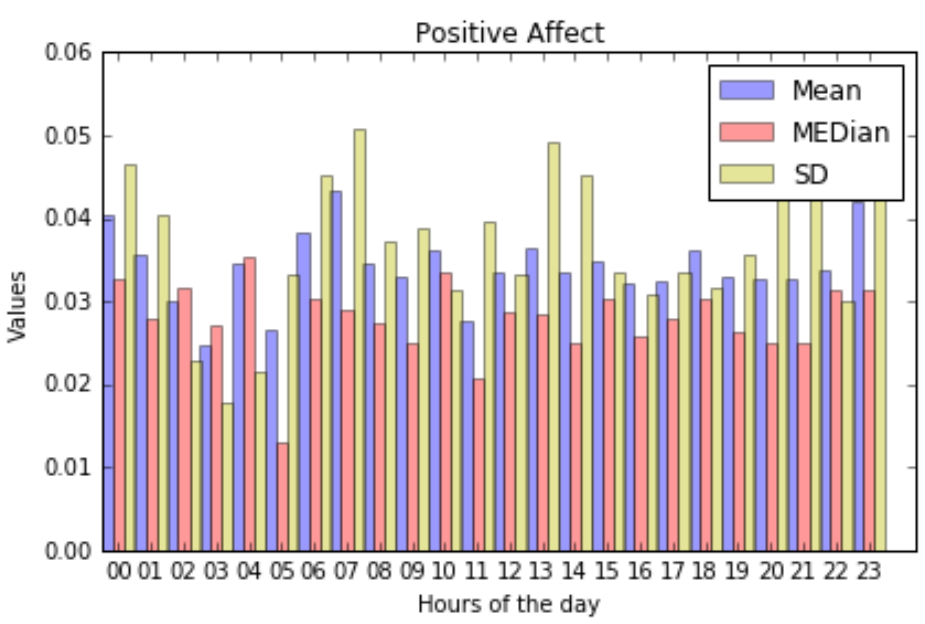


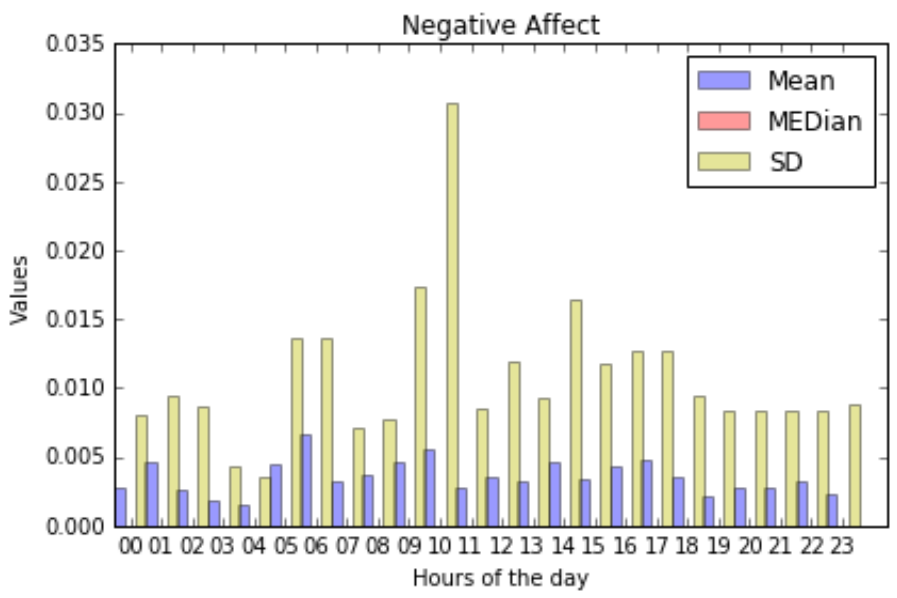


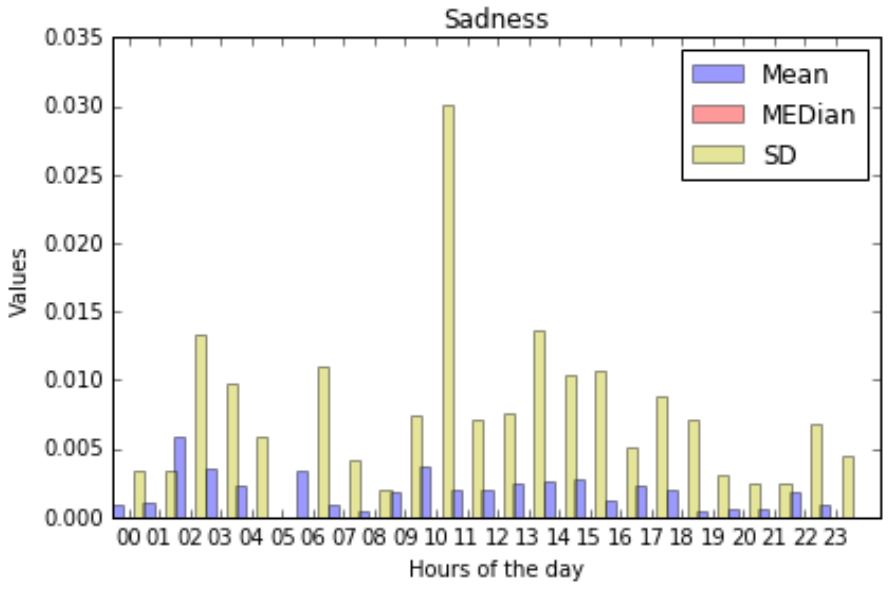


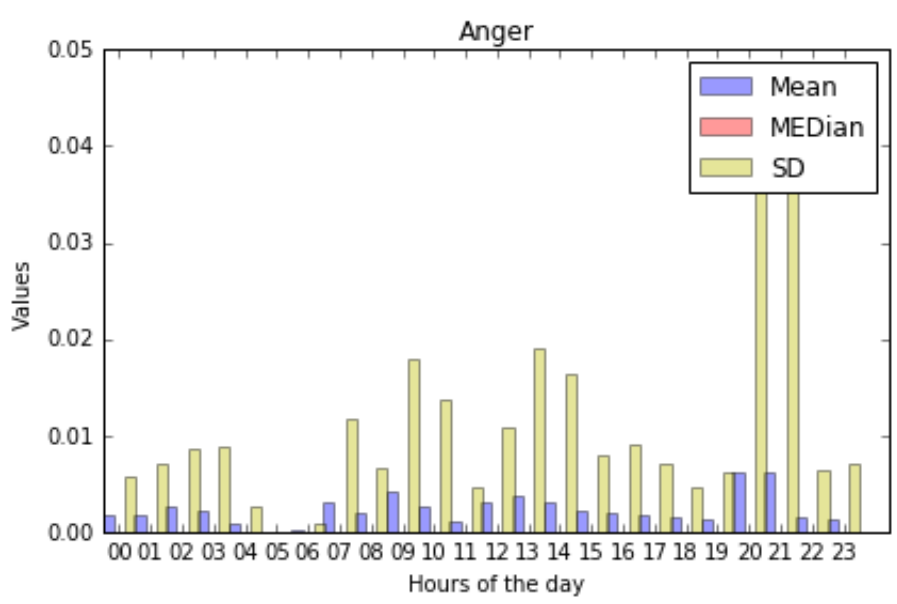


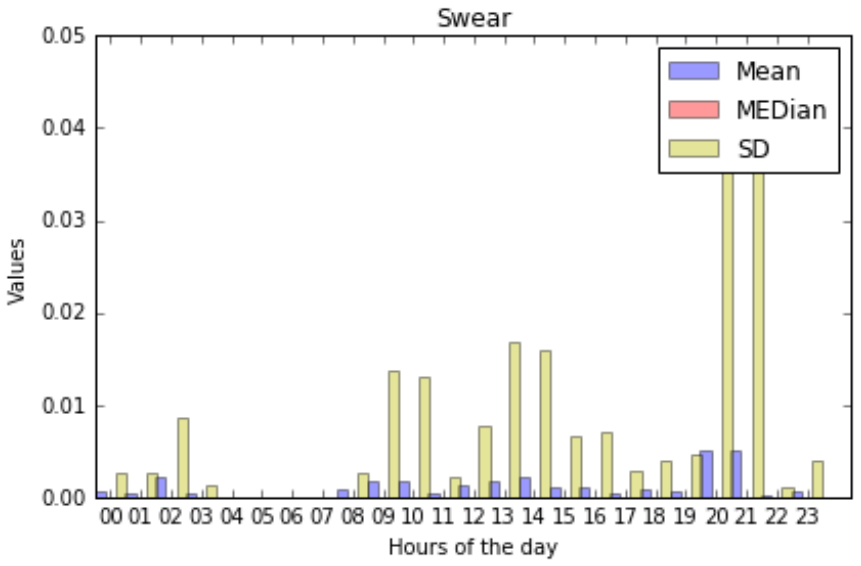


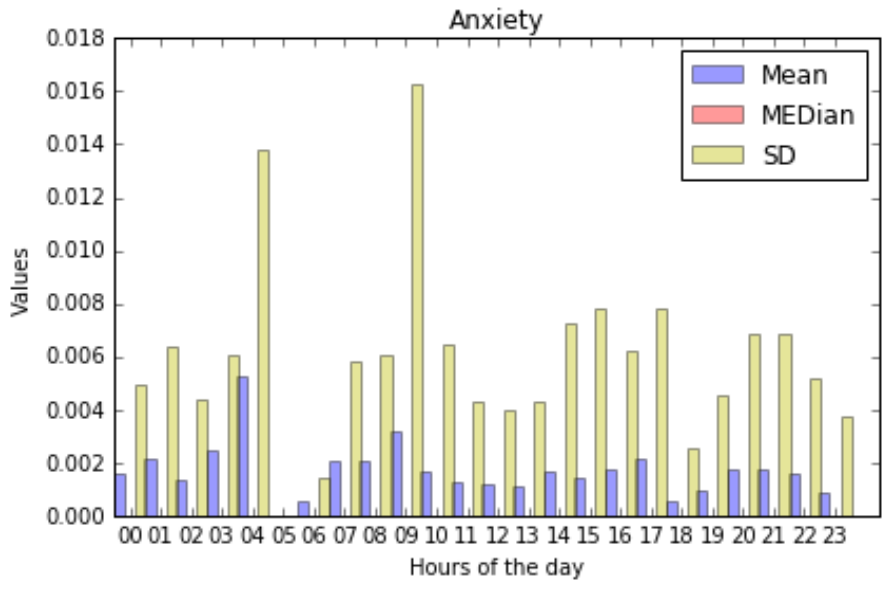












Discussions

Positive Affect: Positive affect is relatively higher on Saturdays, Sundays and Mondays. On a hourly basis, Positive affect reaches a high point in the morning (06-07 hours) and then drops during the day. It reaches local maxima points during 12-14 hours and 16-17 hours. Highest values are noticed during the 23rd and 00th hours.

Negative Affect: Negative affect shows very little variations over the week. This reveals that the posts probably contain more of campus related queries and neutral discourse, than revelations of personal frustrations or complaints. The hourly distribution of negative affect shows a little more variations. High points in Negative affect are seen at 05-06 hr, 10th hr, and 1th hour. This can be indicative of the fact that students do not like to wake up early in the morning, especially after long nights indicated by 1th hour negative posts.

Anxiety: Anxiety reaches a high on Monday, which is understandable as it is the start of another academic week. It again goes up on Fridays. This is counterintuitive as Fridays should be less anxious. However, feelings of "not having done enough during the week" may be abundant. Anxiety remains fairly low during the day and shows a high peak in the late night hours- 03-04 hrs.

Anger: Weekly and Hourly distributions of anger are fairly uniform and remain low through-out the day. Slightly greater values are seen during the 20-21 hours and on Fridays and Saturdays. It is interesting to note hoe the values decrease almost uniformly as the progresses till about 19th hour, and then there is a maximum peak.

Sadness: In the hourly trend of sadness three high points are noted at 02, 02 and 10th hours. The values reduce to a minimum during the end of a working day at 19th hour. The weekly trend of sadness, remains more or less uniform and reaches a maximum value on Saturday. A

Swear: The only notable swear values are seen during the night at 20th and 21st hours. The values remain very close to zero throughout the day. A similar trend is seen in the weekly plot. Close to zero values are seen throughout the week, with peaks on Fridays and Saturdays. This can be indicative of the fact that swearing is used less as a means of complaining or abusing, more as a means of letting go.

The higher positive affect values and lower negative affect values show that the campus discourse is generally positive

How do the findings align with your general perceptions of the campus student body?

The findings align well with my general perceptions of the student body. The overall attitude of the students are very positive. They are used to handling stress and not much negative emotions are expressed about it. However the anxiety of not doing enough is always there, considering the competitive environment of Georgia Tech. This can be resulting in the high values of anxiety, sadness and anger on Fridays and Saturday, where students might express frustrations and anger about the week that went by and the time they might have wasted (according to them). In the hourly distributions we see quite low values (in anxiety, anger, swear, sadness) during the hours of the day when school might be on. This is consistent with my perception of how busy student life is on campus. There may hardly be any time to express anxiety or anger through online posts.

Comparison with Dodds et al.

In Dodds et al. we see a very pronounced weekly cycle of average happiness. Saturday has the highest happiness value, followed by Friday and Sunday. The average happiness then declines over the week, with a low on Tuesday. There are small increases on both Wednesday and Thursday, and then a jump on Friday. Saturday has the highest average happiness.

Dodd et al find that the happiest hour of the day is 5 to 6 am, after which they notice a steep decline until midday followed by a more gradual descent to the on-average low of 10 to 11 pm. They also saw negativity decreasing well into the night

From the LIWC analysis done on the reddit posts from Georgia tech, we see slightly different trends. Positive affect is high on Saturday and Sunday. However unlike, we see high positive affect on Mondays as well, which deviates from the findings of Dodds et. al. Moreover, instead of Tuesday, we see a low point of positive affect on Thursday, which deviated from Dodds et al, but conforms with the findings of "Pulse of the nation: U.S. Mood Throughout the Day inferred from Twitter, 2011". In the hourly trends, we also notice positive affect being highest during the early morning hours of 07-08 and then going down to an afternoon low, but then jumping up again during the 12-13th hours which are typically lunch hours. We also see both negative affect and positive affect decreasing late into the night, which can be taken as a partial conformation to the findings of Dodds et al.

The difference in the findings can arise due to several reasons such as:

1) The nature of Twitter and Reddit data differ significantly from each other. Twitter is more in the moment and may contain more personal emotional revelations than what can be seen on a reddit thread discussion.

2) The data that Dodds et al. used came from a wide range of users from throughout the US. On the other hand, we are only concerned with the discussions related to our campus.
3) As we saw in the n-grams analyses, much of the discussion is based around questions and confusions regarding the system/ life at Georgia Tech and the volume is seen to go up during the ends/ beginnings of semesters. This shows that he discussion is more of an information exchange and is less likely to contain expressions that reveal any sort of personal emotions or anger, or frustration.
4) The overall positive nature of the posts give rise to high values of positive affect, however they may not be truly indicative of the average happiness of the campus, which might be a reason why the findings do not totally match with the findings of Dodds et al.

Campus Vibe

1)

moral_dimensions	mean	median	sdev
HarmVirtue	0.000760084	0	0.00584284
HarmVice	0.000319771	0	0.0039447
FairnessVirtue	0.000574536	0	0.00362975
FairnessVice	5.84777e-05	0	0.000834351
IngroupVirtue	0.00136292	0	0.00668019
IngroupVice	0.000151991	0	0.00228983
AuthorityVirtue	0.00518539	0	0.014777
AuthorityVice	6.45588e-05	0	0.00176042
PurityVirtue	0.000368885	0	0.0046454
PurityVice	4.82805e-05	0	0.000954401
MoralityGeneral	0.00354047	0	0.0128365

2)

The morality of the campus (dimension id :11) The campus shows relatively high moral values as seen from the moralityGeneral dimension value.
The tone of the campus (dimension id 01-02): High mean value of HarmVirtue and lower mean value of HarmVice suggests that the tone of the campus is generally caring and kind and more virtuous.
Do people tend to fair and impartial (dimension id 03-04)?: High mean value of FairnessVirtue compared to the very low value of FairnessVice suggests that the people tend to be fair and impartial.
Is the tone collective or authoritative (dimension id 05-08): Higher values of authoritativeVirtue show a more authoritative tone than collective.
Does the tone show degrading tendency or disgust (dimension id 09-10): With lower values of PurityVice, it can be said that fairly less amount of disgust or a degrading tendancy can be noticed in the tone of the posts.

3) Sample of posts with low morality general scores :
"i just swtiched majors at the end of the fall semester but careerbuzz still shows my old major, when does it change? i tried to change it manually but it looks like it updates from the schools system. thanks for your help :) 0.0"

"just imagine if all georgia tech students are certified to use hand guns. this will forever alter the expected narrative of our once sad and depressing clery alerts. it would also make clery alerts few and far between." 0.0

"i was super interested in studying abroad at hkust in a couple years, as they offer a ton of my required classes, and i have just always dreamed of visiting hong kong. i was wondering if anyone here has studied at hkust, and if so, if you could share any info on the difficulty of classes / overall experience there. would you recommend it to a cs major? thanks in advance!" 0.0

“due to some recent troubles i am no longer able to pay the full amount of tuition for next semester. i am supposed to recieve ~3750 in federal subsidized and unsubsidized loans for this next semester. if i decide to drop down to 6 hours, will that have any effect on the dispersal of the loan money?” 0.0
“hey guys, i have an opportunity for younger djs to play out on thursday night at a place on edgewood. this is an event that allows you to showcase your music and talent as a dj. it's not a huge party, it's more of an intimate setting with a crowd typically from 10-40 people. really depending on the night, thursday's are pretty volatile. we've been doing this event for 2 years now and want to start bringing in younger talent. however i am having difficulties finding the talent since i'm a bit older and my friend group just doesn't have the young connects anymore! i'm looking for more funk, house, techno, tech house, style djs. not looking for banging edm, hip hop, or indie music djs. even if your style may not fit in the above but you're interested reach out to me anyways with your sc or mixcloud and i'll take a listen. i won't give out all the details about the event here but if you reach out to me i will provide more info. thanks guys! “0.0

high morality general scores :

“exactly how bad is the internet at gtl?” 0.125
“gatech: good at mars, intermediate at football.” 0.142857142857
“just changed my flair. feels good, man” 0.142857142857
“what's the good word” 0.25
“was it worth it?” 0.25

Examine on how your qualitative examination of this sample aligns with/deviates from what the dimensional value actually indicates; thereby reflecting on the utility and limitations of dictionary based approaches of detecting a psychological attribute in a community/population.

An examination of 5 of the lowest moralityGeneral posts reveal a few posts that talk about guns and dropping school. However most of them are fairly neutral in tone and do not express low morality in anyway.
An examination of 5 of the topmost moralityGeneral posts show that posts can be just small questions which can have a negative annotation as well- for example “ Was it worth it?”. Because of the word “worth” the analysis will generate a good moralityGeneral score for it, whereas in reality it can be treated as a fairly neutral post in terms of morality or positivity as well. Similarly we also see the failure of the system to understand a sarcastic comment : “ Gatech: good at Mars, intermediate in football” .Thus a dictionary based approach can have its own limitations and pluspoints depending on the type of data available. I feel it can be more useful where the posts are more personal and less informative.