A Project Report

On

Spam Detection in Social Networks

Pragathi Reddy Mallu -109748757, Vivek Nuthalapati-109747457,
Vasudevan Nagendra-108482623
Department of computer science,
Stony Brook University,

Under the guidance

Of

Professor: **Dr. Jie Gao**
Course: Social Networks
Date: 05/19/2015

ABSTRACT

The project work is titled as “Spam detection in social Networks”. The goal of this project is to identify the spam profiles in a user social network-Facebook”. This project can be considered as a tool that detects the spam profiles that are linked to a user account, when respective ego network data is given to it. Spam in social networks refers to any unwanted messages send by either legitimate or illegitimate users. The project uses naïve Bayesian algorithm to identify the spam content in the message. In order to have a clear differentiation between spam accounts and legitimate accounts, we focused on analysis of key features in Facebook such as no of friends, no of posts, ratio of friend requests send/received etc. and classified users into various clusters such as spammers who are less active, spammers who are more active, legitimate user, more active legitimate user. Data collection for an ego network is done by doing a survey on Facebook usage. In addition to this, we created spam profiles in order to know to what extent Facebook is able to detect spam profiles and these spam profiles were also processed through the tool.

TABLE OF CONTENTS

1. INTRODUCTION:

The project "Spam Detection in Social Networks" detects whether a user profile in social Network-Facebook is spam profile or not. An ego network with all the required information when given as input, tool classifies all the user accounts into the respective clusters and identify all the spam accounts that are linked with that ego network. In order to do this, all user accounts of an ego network with no of friends, no of friend requests send/received, etc. is given to the tool.

1.1 SOCIAL NETWORK-FACEBOOK:

Social Network: A social networking site is a platform to build social networks or social relations among people who share interests, activities, backgrounds or real-life connections. A social network site consists of a representation of each user (often a profile), his or her social links, and a variety of additional services. Social network sites are web-based services that allow individuals to create a public profile, to create a list of users with whom to share connections, and view and cross the connections within the system. Most social network sites are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Social network sites are varied and they incorporate new information and communication tools such as mobile connectivity, photo/video/sharing and blogging. Online community services are sometimes considered a social network service, though in a broader sense, social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, pictures, posts, activities, events, interests with people in their network.

Facebook: Facebook is a social networking site which allows user to do all the above mentioned activities. In addition to these Facebook provides others options such as privacy settings where user can enforce visibility restrictions on his posts, tags etc. There are multiple options such as public to friends, public to anyone in settings.

2. SPAM IN SOCIAL NETWORKS:

Spam is a well-known and well-studied concept in the case of messages. Spam in the case of email meant as irrelevant messages from an unintended user of no value to the legitimate user. In the case of social networks, what is spam? In the case of social networks such as Facebook, twitter or, the spam is any sort of message or post from either an unintended user or from some group or any entity part of this social media, who is actually not intended to send such messages. These messages carries no value or relevance to the actual user, while the spammer benefit financially while spreading the message. The main reason for spamming is always not required to be monetary benefits, but include be done for the purpose of fun or for some other benefits.

3. FEATURES USED BY CLASSIFIER:

Classifier analyses an ego network using the three below mentioned characteristics of a social network.

1. Network

2. Activity on Facebook

3. Type of messages

3.1 NETWORK:

Network of the user is a major categorization factor. It defines the quantity and quality of the user's network. Network of a user is determined by taking two factors into account.

No of friends: No of friends indicate how large a user network is. It is a well weighted factor to determine the network size. If the no of friends is greater than the average no of friends of that ego network, it says that user has good network, if the no of friends of a user is very less compared to the average no of friends, then user is considered to have a very poor network.

Ratio of no of friend requests received /no of friend requests send: If this ratio is very less, it signifies that user is more active in sending friend requests. Spammers usually send more no of friend requests than the number of friend requests received. As spam profile doesn't have real identity, the probability of receiving friend requests is also less. A legitimate user profile has this ratio approximately equal to 1.

3.2 ACTIVITY ON FACEBOOK:

A user is said as an active user based on his activities on Facebook. Examples for activities are posting on his/her own wall, tagging himself or his friends in a picture, sending messages etc.

No of Posts: If a user posts more than the average number of posts in his ego networks, then user is said as active user. This factor doesn't contribute much to identification of a spam account, but helps in determining the active participation of user in Facebook.

No of Tags: A user posting a video/photo and tagging his friends or himself indicates the identity of a user, whereas spammer always try to hide their identity. In addition to this, a spammer has less number of tags compared to the average number of tags in that ego network. So tags play a key role in determining the active participation of a user and differentiating between spammer account and legitimate account.

Ratio of no of messages received/send: If this ratio is less than one, it conveys that user is sending more no of messages than he is receiving, spammers send more number of messages irrespective of getting response for their previous messages. This can be used to identify a spam profile. If this ratio is close to 1, it signifies user actively participate in Facebook either by initiating a conversation or giving response to messages he received.

3.3 TYPE OF MESSAGES:

This factor analyses the messages received to determine whether the message contains spam. If the messages are recognized as spam, then user account can be considered as spam.

Bayesian Filter: The system is designed to handle the training data i.e. message that is annotated or labeled as either ham or spam, and hence it is designed to handle the words and their occurrence counts, rather than handling a text message directly. The labeled messages when passed to the classifier will tokenize the message for extracting the features and their occurrence counts in the message, which are both domain specific and normal text. This tokenized data is fed to the classifier for the classification, which in turn calculates the posterior probability for the features independently. The calculated probabilities will define the classes for those features, in this case two classes spam or legitimate.

Now, when the test data of the same pattern of training data as mentioned above, is fed to the Naïve Bayes classifier, it will calculate the probabilities and declare the message as spam or not. The accuracy is calculated on the basis of number of messages the classifier was able to classify correctly i.e. message labeled as spam need to be classified as spam and the same for ham.

Let us consider an example to see how classification is done. Consider the specific instance x, and the assignment of values will give us feature vector X={x1, x2, x3, x4, …, xn}. The instance values could be any list of word occurrence in the message, such as 'banking', 'dollars', '$$$', 'lottery', etc. In our case we have only two classes "SPAM" & "HAM".

By Bayesian theorem:

$$P(C=ck \mid X = x) = P(X=x \mid C = ck) * P (C = ck) / P (X = x)$$

Equivalently,

$$\text{Posterior Probability} = (\text{likelihood} * \text{Prior Probability}) / \text{evidence}$$

By Naïve Bayesian classifier which assumes each feature as independent and hence:

$$P(X = x \mid C=ck) = \prod_i P (Xi = xi \mid C = ck)$$

Examples by applying Naïve Bayesian to the spam filtering:

$$P(\text{SPAM} \mid \text{"\$\$\$ Won Lottery"}) =$$

$$P(\text{"\$\$\$"} \mid \text{SPAM}) * P (\text{"Won"} \mid \text{SPAM}) * P(\text{"LOTTERY"}|\text{SPAM})* P(\text{SPAM}) /$$

$$P (\text{"\$\$\$"}) * P(\text{"Won"}) * P(\text{"LOTTERY"})$$

i.e. $P(\text{SPAM} \mid x1, x2, ...xn) = P(x1, x2, x3, ...xn \mid \text{SPAM}) * P (\text{SPAM}) / P (x1, x2, x3, ...xn)$

Posterior probability for spam is conditional probability of features such as $$$, Won & Lottery in either the spam or legitimate class.

posterior probability is □i n=1 P (Xi | SPAM)

But, in our case as the number of occurrences are only provided, we need to calculate the probability of a feature given either spam or ham is as follows:

P (Xi/SPAM) = (Count_Xi / (C_All_Spam + C_Unique) …..(a)

Count_Xi = Count of all features Xi in the Spam Messages.

μ = Smoothing Factor.

C_All_Spam = Count of All Words in Spam Messages

C_Unique = Count of Unique Words In Spam and Non-Spam Messages.

Px - Probability of the work Xi being spam in total word count.

4. ALGORITHM USED IN CLASSIFIER

Classifier is designed based on the following properties in Facebook. Number of friends, Number of friend requests sent, Number of friend requests received, number of messages sent, number of messages received, number of tags, number of URLs in message sent/received, number of posts in a week, content of messages received from a user is collected for all user accounts in an ego network and given to the classifier.

If a user account has more number of friends and ratio of friend requests to the friend requests send is close to one, then the user has a very good social network. If a user is more into posting on his/her wall and tagging friends or himself/herself frequently and ratio of number of messages received to no of messages send is approximately equal to 1 then user is treated as active user, that is user participates very actively in social interaction. To do further analysis on spam, in an individual ego network, spam filters are applied on the messages received from all the users. Spam filtering is done by using Bayesian naïve filter approach.

Network, Activity on Facebook, Type of messages are the three characteristics we considered. Every user account is initially provided with legitfactor1, legitfactor2, legitfactor3 respectively. Initially, considering all the user accounts in ego network as legitimate accounts, legitfactor1 and legitfactor2 were given a value of 1000 and legitfactor3 as 0.

In Network, the average value for no of friends in an ego network is calculated. A user with no of friends lesser than the average value gets his legitfactor1 decreased by value of 200. Average value of ratio of no of friend requests received to the no of friend requests send is computed by taking all the user accounts in ego network. The absolute difference between average ratio and ratio of a user account is multiplied with 1000 and is subtracted from 500 as this ratio is given 50% weightage in cluster 1. The result is subtracted from legitfactor1. If the final legitfactor1 is greater than 400 then user account is considered to have a good social network.

In Activity on Facebook, the average value for no of posts is calculated. A user with no of posts lesser than the average value gets his legitfactor2 decreased by value of 100. The average value

for no of tags is calculated, a user having lesser than average value gets his legitfactor2 decreased by value of 100.The average value of ratio of no of messages received to no of messages send is computed. For a user account, the absolute difference between average ratio and ratio of a user account multiplied with 1000 and is subtracted from 500 as this ratio is given 50% weightage in this category. . If the final legitfactor1 is greater than 500 then user is a highly active user on Facebook.

In type of messages, the total no of messages received from the user are given to Bayesian filter to know the spamness in the messages, then ratio of total no of spam messages to legitimate messages is considered. If this ratio is less than 0.5 then legitfactor3 is set to 1 otherwise it is set to 0.

Based on legitfactor1, legitfactor2, legitfactor3 values, users are classified into various clusters

5. CLUSTERS FORMATION AND ANALYSIS:

Network, Activity on Facebook and Type of messages are the three characteristics considered. Using these the below mentioned clusters are created, every user account goes into one of these clusters.

Spammers in Initial Stage: A user with a network less in size and very less active in Facebook, sends spam messages to others is considered as spam account in initial stage. For a user account, if its legitfactor1 is less than 400 and legitfactor2 is less than 500 and legitfactor3 equal to 1, then user account falls into this cluster.

Spammers with Good Network: A user with a network more in size and very less active in Facebook, sends spam messages to others is considered as spammer with good network. For a user account, if its legitfactor1 is greater than 400 and legitfactor2 is less than 500 and legitfactor3 equal to 1, then user account falls into this cluster. There is a potential harm with these kind of spammers as they are managing to not getting detected and expanding their network.

Spammers with High activity: A user with a network less in size and very less active in Facebook, sends spam messages to others is considered as highly active spammer in Facebook. For a user account, if its legitfactor1 is less than 400 and legitfactor2 is greater than 500 and legitfactor3 equal to 1, then user account falls into this cluster. These spammers are being identified as spammers to certain extent because of which they are unable to expand their network, but they are very active in Facebook in terms of posting and sending messages.

Spammers with Good Network and Activity: A user with a network more in size and highly active in Facebook and sends spam messages to others is considered as highly active spammer in Facebook. For a user account, if its legitfactor1 is greater than 400 and legitfactor2 is greater than 500 and legitfactor3 equal to 1, then user account falls into this cluster. These spammers might not have been detected in initial stages and expanded their network and are very active in Facebook.

Legitimate users in Initial Stage: A user with a network less in size, very less active in Facebook, and who doesn't send spam messages to others is considered as legitimate user in initial phase in Facebook. For a user account, if its legitfactor1 is less than 400 and legitfactor2 is lesser than 500 and legitfactor3 equal to 0, then user account falls into this cluster.

Legitimate users with good network: A user with a network more in size, very less active in Facebook, and who doesn't send spam messages to others is considered as legitimate user with good network phase in Facebook. For a user account, if its legitfactor1 is more than 400 and legitfactor2 is lesser than 500 and legitfactor3 equal to 0, then user account falls into this cluster. These kind of users do not actively participate in tagging. Posting but tries to maintain a good network.

Users with High Activity: A user with a network less in size, highly active in Facebook, and who doesn't send spam messages to others is considered as highly active legitimate user of Facebook. For a user account, if its legitfactor1 is less than 400 and legitfactor2 is greater than 500 and legitfactor3 equal to 0, then user account falls into this cluster.

Further analysis is done in this cluster in order to differentiate between spammer and legitimate account. These kind of users do actively participate in tagging, posting but doesn't have good network. If spam content in the messages received by this user is zero or approximately equal to zero, then user is considered as legitimate user.

The user might be a spammer because of which his friend request was not accepted by many other users. If no of friend requests send are really large in number than no of friend requests received, and number of friends, then user is considered as spammer in initial stage, although he is not sending spam messages.

Legitimate users with High Activity and Good Network: A user with a network more in size, highly active in Facebook, and who doesn't send spam messages to others is considered as highly active legitimate user of Facebook. For a user account, if its legitfactor1 is greater than 400 and legitfactor2 is greater than 500 and legitfactor3 equal to 0, then user account falls into this cluster. These kind of users do actively participate in tagging, posting and also concentrates on expanding their network.

6. DATA COLLECTION:

The main challenge in the project is gathering the data, as we were not able to gather data that is relevant for us to carry out our implementation, we need to collect the data of our own. We have collected data from the users by conducting the survey and collecting the data from them. We are fortunate enough to gather data from more than 300+ Facebook users of our ego network. We have provided them either general APIs available for Facebook for collecting this data. Most of the cases they are required to manually extract the details to share with us. The data is collected for more than four weeks of time. We have taken enough care to see that the data is not tampered and

The data that we collected has helped us in framing the features that are required to categorize the spam nature of the profile and the messages they exchange. We asked the participants to provide us the following set of data inputs with respect to their Facebook usage and data. We are also fortunate enough to gather the messages from the users of our Facebook network, which included our project partner's messages they exchanged with their friends and other friends who exchanged their recent messages.

The data and posts that are collected are analyzed for the spamness in the content. Both the types

of data i.e. messages and posts are analyzed for two aspects, one being the spamness of the content and the other being categorizing the URLs on the basis of their domain category. We have collected this messages so as to carryout spam analysis on this data and correlate with our messages and posts. This has helped us analyze the spam-ness of the data in the same ego network. We developed a spam naive Bayes algorithm to identify the spam nature and correlate with our data. We have utilized some of the data that we have obtained for our Facebook friends for both training and test data. We have taken enough care to see the data is not biased and will not show the biased results for the test data. Hence, while distributing the data for training we have uniformly distributed the spam and legitimate data into the training set.

On top of it for collecting the data we have created around 12 spam profiles, which catered us in number of ways. We were able to replicate these profiles to spam profile. To our surprise we started receiving the friend requests on these profiles, which we created to replicate the spam profile. To our surprise, these requests are from actual spammers. In this way these profiles has helped us both in collecting the real spammers and their data and has helped us replicate the spam profiles for our experiments. The details about these spam profiles and their data is also attached in the report. We ran the scripts that could run on this data and carryout the needed analysis and be able to properly categorize the test data as spam and properly categorize the legitimate profiles. We have taken care to see that, there is not much false positives and false negatives.

The Survey details that we collected as follows:

1. Number of friends
2. Approximate Number of friend requests sent in a week
3. Approximate Number of friend requests received in a week
4. Approximate number of messages sent in a week
5. Approximate number of messages received in a week
6. Approximate number of tags in a week
7. Approximate number of URLs in message sent/received in a week
8. Approximate number of posts in a week

7. TESTING AND RESULTS

Data is collected for single ego network, as this data comprises only legitimate user's information, we created 12 spam accounts out of which 3 were identified by Facebook as spam accounts in the process of managing those accounts, and the rest 9 accounts data is combined with the legitimate user's data. Then the data is processed through the tool, all the 9 accounts were detected as spam accounts, in addition to these few real users were identified as spammers. The last 9 records in the data file corresponds to the spam accounts we created.

Every user is categorized into a cluster. This was compared to ground truth data for every user. Then efficiency of tool was calculated as total no of comparisons that came out to be true divided by total no of comparisons. Total no of comparisons that came out to be true for the ego network we considered are 359 and total number of comparisons are 366 so the efficiency is 98.8%

Total no of users count in each cluster as per our classifier are

Spammers in initial stage: 0

Spammers with good network: 1

Spammers with high activity: 15

Spammers with good network and high activity: 12

Legitimate users: 122

Legitimate users with good network: 0

Legitimate users with high activity: 0

Legitimate users with high activity and good network: 191

8. CHALLENGES

The challenges that we faced were

- While collecting data, Facebook users are not ready to share their information, then we targeted 20 users per day and personally requested them and made sure that they submit the survey.
- Identification of features to design the algorithm was a big challenge for which we have gone through many articles and papers on social networks.
- As Facebook is also taking certain steps to stop the creation of profiles if profiles are created from same IP, with great difficulty, we created 12 spam profiles.
- We monitored these 12 profiles every now and then, in terms of sending friend requests, messages, posting on walls, but Facebook identified 3 out of those 12 profiles and asked for a proof of identity.

9. CONCLUSION

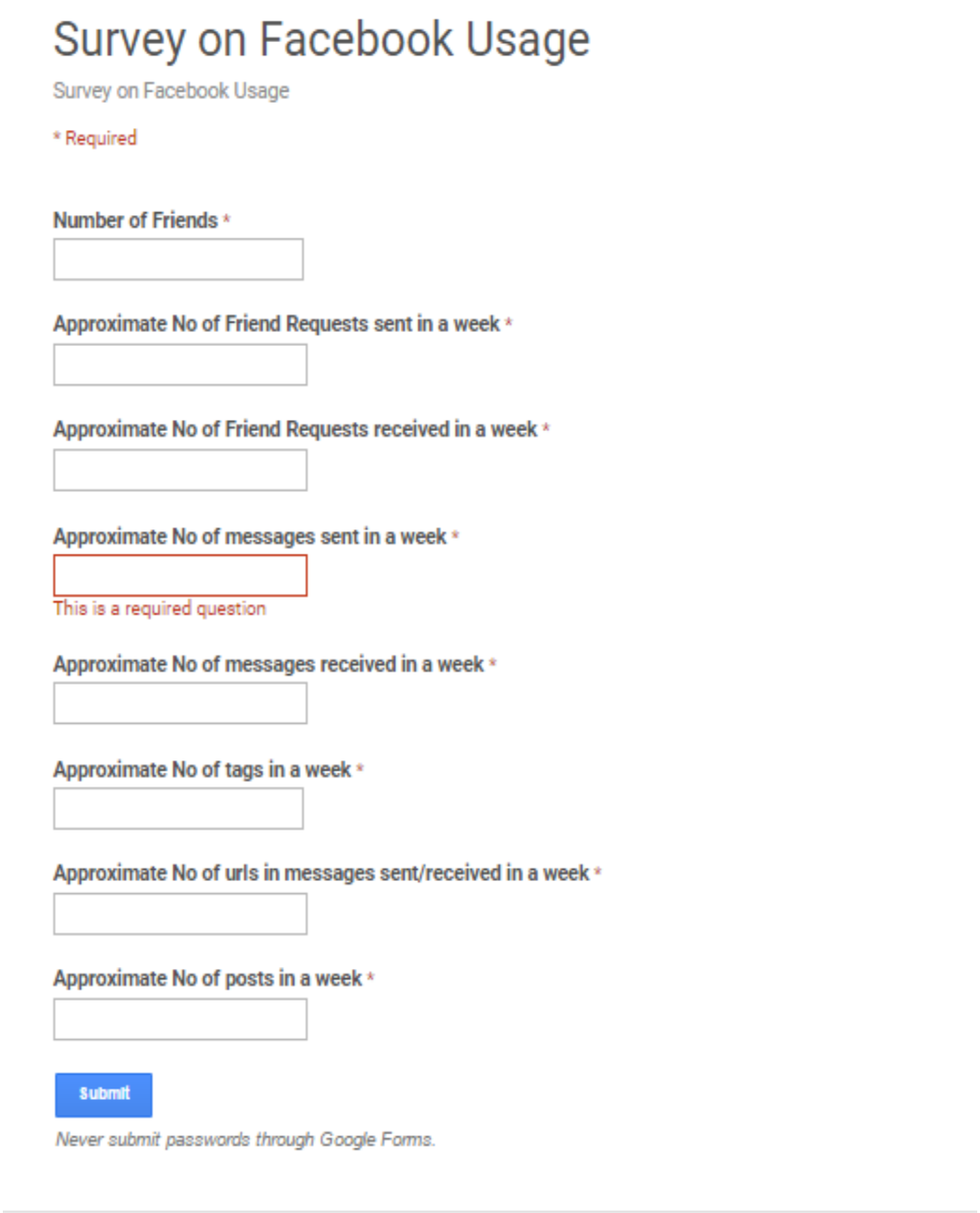
This project is a tool used for identification of spam profiles in an ego network. If Facebook provides this tool to every user, then user will be able to identify spam accounts that are linked to his account. The efficiency of tool was 98.8%. The classifier was developed by considering characteristics of Facebook- Network, Activity in Facebook and type of messages. Each characteristic is again estimated based on the factors such as no of friends, ratio of no of friend request received /no of friend requests send, no of posts, no of tags , ratio of no of messages send/ no of messages received, content of the messages.

10. REFERENCES:

http://en.wikipedia.org/wiki/Social_networking_service

APPENDIX 1

DATA COLLECTION SURVEY FORM:



Survey on Facebook Usage

Survey on Facebook Usage

* Required

Number of Friends *

Approximate No of Friend Requests sent in a week *

Approximate No of Friend Requests received in a week *

Approximate No of messages sent in a week *

This is a required question

Approximate No of messages received in a week *

Approximate No of tags in a week *

Approximate No of urls in messages sent/received in a week *

Approximate No of posts in a week *

Submit

Never submit passwords through Google Forms.

Data collected is in the format:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Index	Number o	Approxim	Approxim	Approxim	Approxim	Approxim	Approxim	Approxim	message											
2	0	448	2	2	50	70	2	10	1	If you don't have any referrals in Microsoft then apply for it during career fair in September;Hey, are you still going to SSI											
3	1	438	10	15	100	150	1	10	2	Write notes in your own words											
4	2	700	1	1	10	20	5	5	1	Actually I'm not going on monday;I'm going on tuesday;I think Sowmy is going on monday											
5	3	337	1	1	20	20	0	0	0	How are you;oh, he was okay with Tuesday too;he was telling me that you might come too											
6	4	1044	1	1	500	500	1	1	3	Use the dictionary to check the meaning of any unfamiliar words;and we three will go;Im waiting for authorization letter											
7	5	523	1	2	50	50	5	10	0	all the best;so that i can join u guys on tuesday;alright man											
8	6	1100	1	2	15	17	0	0	0	kick start;r u going to ssn office tomorrow;I want too											
9	7	523	1	1	25	25	5	5	3	have a nice day;I don't have a ride yet;i also want to go tomorrow											
10	8	621	3	5	230	230	8	5	2	Open up the discussion;There is some guy who drives people right;ya but he has some meeting it seems tomorrow											
11	9	791	5	10	600	600	10	10	3	I love you;Ah unsure emoticon;he can drive us tomorrow afternoon											
12	10	613	1	1	10	10	2	5	1	all is well;That's good for me;but we should be 3											
13	11	586	1	1	2	2	2	5	1	please vote for me;Do you know Akhilesh;He wanted to go too											
14	12	483	1	2	10	10	1	1	1	mumbai is clean city;ya he went now to collect i20;he will confirm once he gets it											
15	13	500	1	1	10	10	1	1	30	india wins the match;ill ask gaurav once akhilesh confirms;Alright. Sounds good!											
16	14	700	5	4	500	500	5	100	4	have you moved to new place;akhilesh will be joining us;i checked with gaurav											
17	15	500	1	1	25	25	5	1	1	sun rises in the east;we will start at 11;is that ok for you;cool.. Ill confirm to him											
18	16	800	5	2	12	8	2	3	15	I want to a job in google;hey man, I have a class till 2;from 1 to 2:30 actually											
19	17	400	1	1	25	15	10	5	30	weather is very bad in newyork;I am not sure if there is going to be a quiz in the class unsure emoticon;we can't get back											
20	18	395	1	1	20	20	1	8	1	game of thrones seaason 2 episode 1 is awesome											
21	19	500	1	2	20	20	5	5	10	when will your exams start;ohh.. not sure;it takes min 3 hrs											
22	20	300	1	3	5	5	1	1	1	fan is not working properly;ah unsure emoticon;can we go after 2:30											
23	21	740	2	2	20	20	2	5	1	please sit on the chair;hmm.. office closes at 4;or earlier in the day, may be 9 or something											