SOCIAL NETWORK SYSTEMS

EE 599

Prof. Keith Chugg

PROJECT 2

Personal Social Media Analysis(PERSMA) Tool

Gulsheen Ahuja

Shrikant Modi

Vipul Garg

Contents

List of Figures

Chapter 1

Introduction

Social Netwotking WWebsites provide a way for users to interact with each other in real time. In the Social Networking Websites, the user joins a network and makes “connections” by creating links with the other users in the network and interacts with them. The widespread popularity and rapid growth of these online social networks encourages us to study its properties and impact. Study of currently available online social networks is useful to design future online social network based systems considering its effect on the Internet. Also, by superlative analysis of social networks, we get the knowledge about information propagation, new directions for information search and concentration of information.

In Project 1 while analyzing different tools for social media we realized that no free of cost tools provide comprehensive analysis from multiple social networks to give more insight into one’s social network personality. Also analysis is done in a static manner and no sophisticated predictions are made on the observed data. This presented a need for a tool like PERSMA and that was the primary goal of our project henceforth.

As part of Project 1 we developed a prototype for a tool “Personal Social Media Analysis Tool” alias “PERSMA”, which aimed to addresse the shortcomings of the available online analysis tools and provide more innovative analysis of social networks. We allowed users to Login with Twitter and analyze their graph of followers who are mutually following each other.

The tool has been enhanced in Project 2 to provide login with Facebook along with Twitter and incorporated many network graphs which produces better analysis of the user’s Facebook and Twitter network data. The network data from two social networking websites are also inegrated in one network graph to provide insight beyond one social network. We also added a simple prediction service that would help a user predict her average retweet count for the near future from her Twitter account.

Chapter 2

PERSMA : Overview

2.1 Motivation

Based on the initial survey of the existing tools we have found the following features that are missing from them.

- As seen so far, the analysis of Social Media is focused on business clients and most of the services are paid. For students and people interested in analyzing their data on a small scale these services are not feasible.
- There is no comprehensive analysis tool which uses data from multiple social networks to give more insight into one's social network personality.
- The analysis is done in a statistical manner and graphs are generated based on the data observed over a period of time. There are no sophisticated predictions made on the observed data to provide better service.

This gave us the motivation to build a new tool for Personal Social Media Analysis that would be used by general public and students like us interested in social network analysis for free of cost. This tool in the scope of this semester would provide integration of data from two popular social networking sites Facebook and Twitter. It would also provide simple prediction analysis of the observed data.

2.2 Architecture

Some of the components in the architecture diagram are explained as follows:

- **Amazon Web Services(AWS)**: AWS provides one of the best cloud computing platforms. We use an EC2 instance in AWS as the host server for both the front end and the back end of the tool.

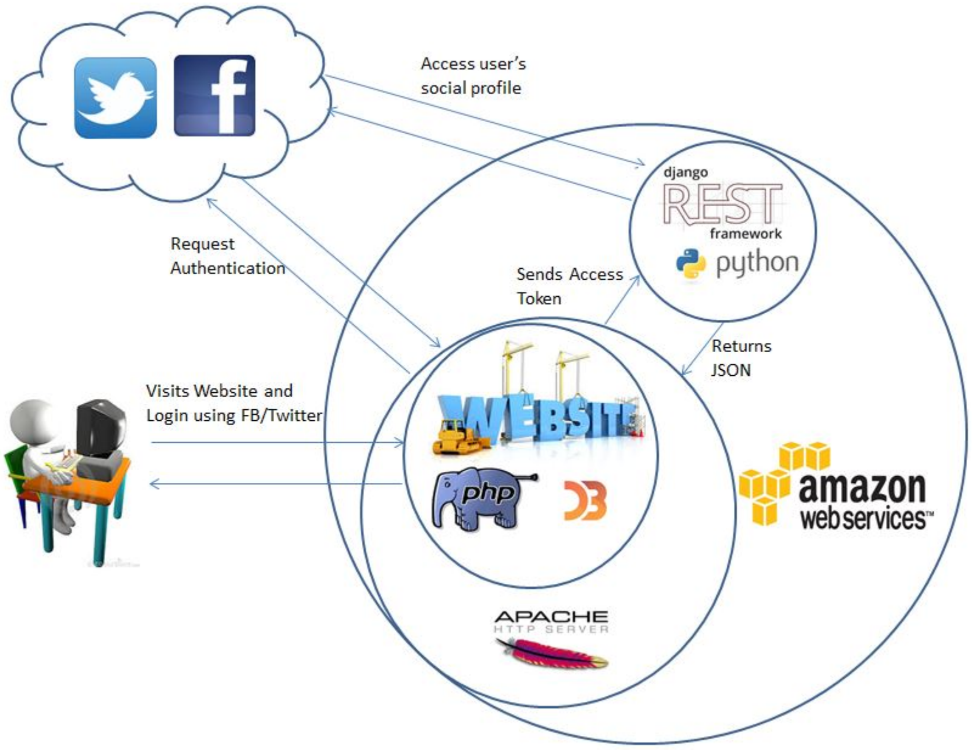


Figure 2.1: Architecture Diagram for PERSMA

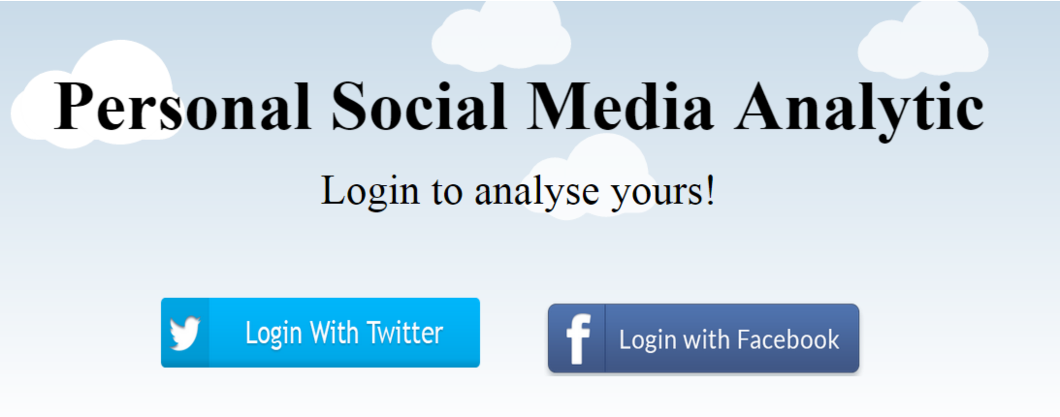
- **Apache Web Server**: Apache provides an implementation of HTTP(Web) server which is used to host the front end of the tool.
- **PHP**: PHP is a scripting and programming language widely used in the web development worldwide. We are using PHP to create the front end of the tool.
- **Data Driven Documents(D3)**: D3.js is a Javascript library used to provide creation and control of charts which are highly interactive and graphical. This is used as the charting library at the front end of the tool.
- **Django Rest Framework and Python**: Python is a really powerful programming and scripting language which has low latency for complex computations. Since, the plan of our project is to integrate modeling and prediction analysis in our tool, Python provides a lot of tools to accomplish the same. Hence our backend of the tool is written in Python. Since, the backend of the tool must make and receive REST calls, we needed to provide REST interface over Python. Django is one of the best frameworks written

over Python to provide the REST functionality. Hence we have made use of it.

2.3 Website Flow

Here is the user flow and some details about the PERSMA tool:

- **PERSMA Homepage:** Persma homepage looks as shown in figure 2.2. The webpage[2] contains two buttons to login with twitter and login with facebook. Once the user clicks on "Login With Twitter" or "Login With Facebook" button she is redirected to twitter or facebook login page respectively. After which, the user needs to provide the credentials and authorize the app. figure 2.3 and figure 2.4 shows the screenshot of authorization of twitter and facebook. Also, The user has been given freedom to start logging in with any of the above social networks. If the user has already logged in to the social networks i.e. facebook or twitter in the browser, after clicking on "Login With Twitter" or "Login With Facebook" button the webpage will auto login to respective social network.

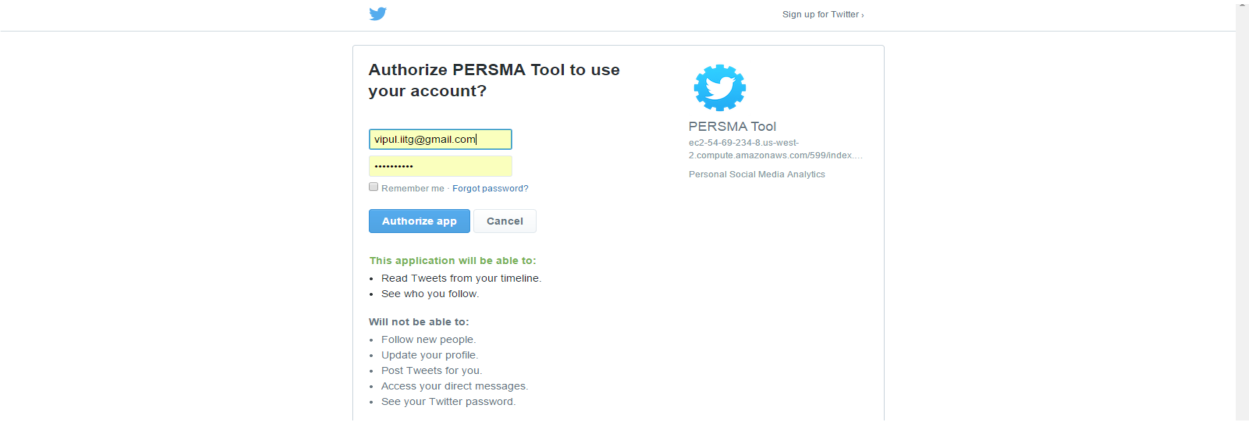


Personal Social Media Analytic

Login to analyse yours!

Figure 2.2: Persma Homepage

- **After Logging in with Twitter:** Once the user logs in with his twitter credentials, he is redirected to the twitter analysis page as shown in figure 2.5. The user can see the profile picture of his twitter social network along with welcome message including user's name. The important feature included over this page is that the user can login to Facebook by clicking on "Login with Facebook" button. This feature allows user



Sign up for Twitter ›

Authorize PERSMA Tool to use your account?

vipul.iitg@gmail.com

••••••••••

□ Remember me · Forgot password?

Authorize app | Cancel

This application will be able to:

- Read Tweets from your timeline.
- See who you follow.

Will not be able to:

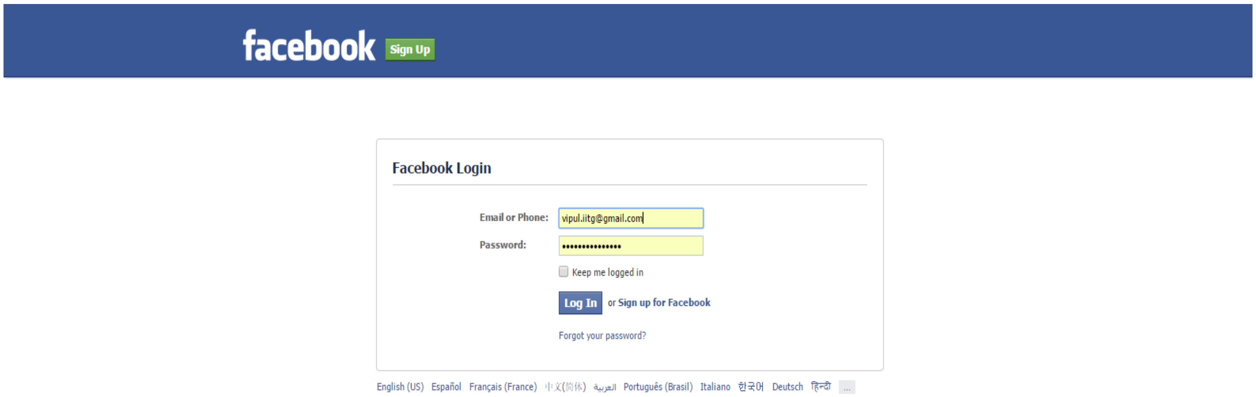
- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your Twitter password.

PERSMA Tool

ec2-54-69-234-8.us-west-2.compute.amazonaws.com/599/index....

Personal Social Media Analytics

Figure 2.3: Twitter Authorization



facebook Sign Up

Facebook Login

Email or Phone: vipul.iitg@gmail.com

Password: •••••••••••••••

□ Keep me logged in

Log In or Sign up for Facebook

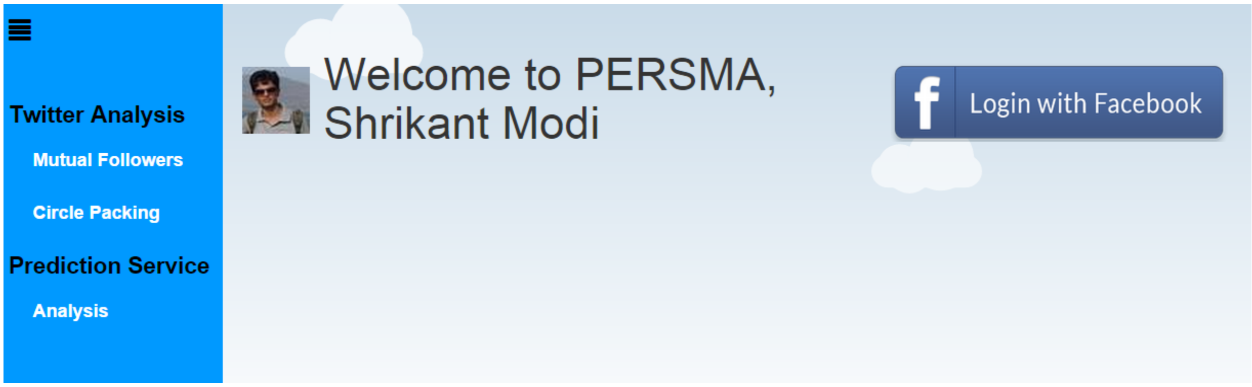
Forgot your password?

English (US) Español Français (France) 中文(简体) العربية Português (Brasil) Italiano 한국어 Deutsch हिन्दी ...

Figure 2.4: Facebook Authorization

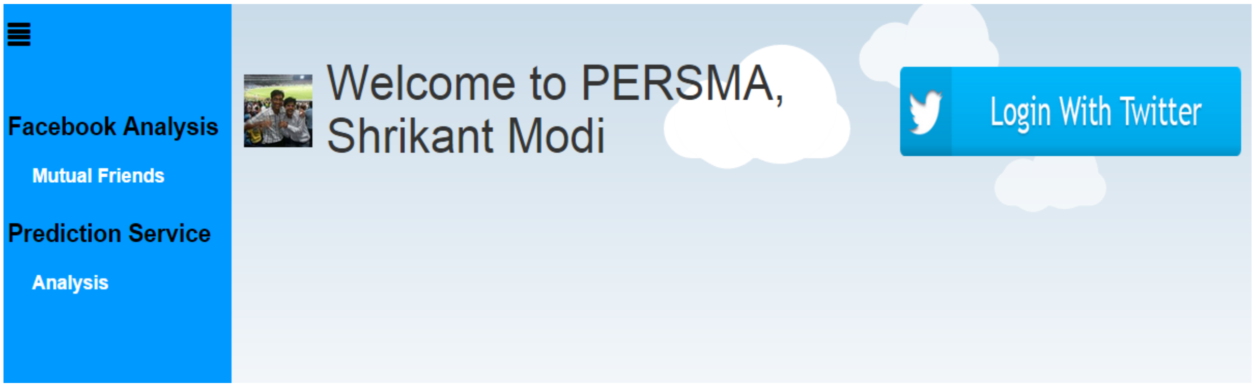
to give details of both social network data to PERSMA. Also, On the left menu bar, features related to twitter analysis are included. By clicking on these features, the user will get graphical and chart based analysis about his twitter data.

- **After Logging in with Facebook:** Once the user logs in with his Facebook credentials, he is redirected to the facebook analysis page as shown in figure 2.6. The user can see the profile picture of his facebook social network along with welcome message including user's name. The important feature included over this page is that the user can login to Twitter by clicking on ”Login with Twitter” button. Also, On the left menu bar, features related to facebook analysis are included. By clicking on these features, the user will get graphical and chart based analysis about his facebook data.



Welcome to PERSMA, Shrikant Modi

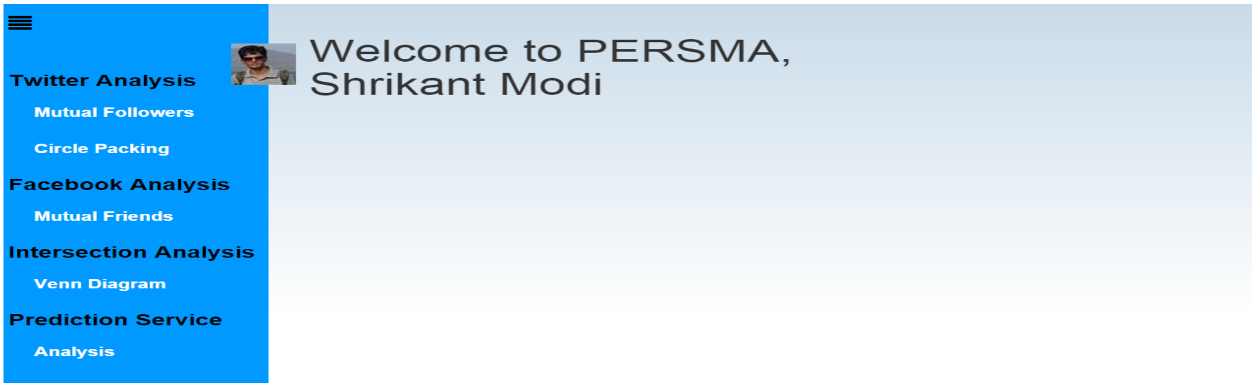
Figure 2.5: Login With Twitter



Welcome to PERSMA, Shrikant Modi

Figure 2.6: Login With Facebook

- **After Logging in with Twitter & Facebook:** Once the user logs in with his Twitter and Facebook credentials, the user sees the page as shown in figure 2.7. The user can see the profile picture the social network he logged in later along with welcome message including user's name. As the user is logged in with both social networks, there is no more additional login button on page. Also, On the left menu bar, the combined features of both social networks are included. These features cover individual social network analysis (Twitter- Mutual Followers, Circle Packing and Facebook- Mutual friends), combined social network analysis (Intersection Analysis venn diagram) and prediction service. By clicking on these features, the user will get graphical and chart based analysis about his facebook data.



Welcome to PERSMA,
Shrikant Modi

Figure 2.7: Login With Twitter and Facebook

Chapter 3

PERSMA : Network Graphs

3.1 Facebook Analysis - Mutual Friends

Mutual friends is one of the important feature related to user's facebook data. This feature gives user more insight about her facebook friend network and helps visualizing the relationship between friend network with the help of circular rotating graph.

- **Similar to Twitter Followers:** The mutual friends graph is similar to the twitter's follower graph which was demonstrated in project 1. In the twitter's analysis, the graph used to give details about relationship between user's followers such as who is following whom, whether two persons are following each other etc.

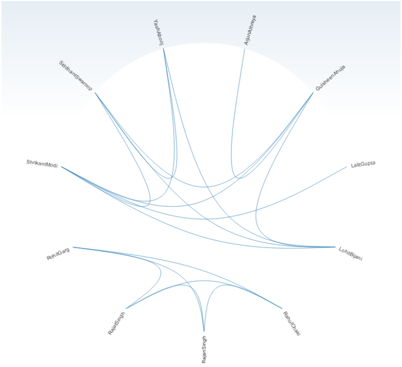


Figure 3.1: Mutual Friends

- **Connectivity between friends:** Similar to Twitter Followers graph, mutual friends graph gives the user idea about the connectivity between the user's friend network. The graph is shown in figure 3.1. The user's friends are listed on periphery of circle. The lines connecting various friends shows the friendship relationship in between them. Also, the user has been given facility to rotate graph 360 degree to see friend names in case they are inverted or at the bottom of the circle. Thus the user gets an idea about who is friends with whom within her facebook friend network.

- **Friends Authorization Limitation:** One of the limitation with facebook analysis is its requirement of authorizing the app. The user has access of her friends data only if her friends have authorized PERSMA tool. The list of friends we could see in graph is due to the fact that only those many people have used PERSMA to provide their facebook authorization.

3.2 Facebook and Twitter Intersection Analysis - Venn Diagram

Intersection analysis gives the user details about the total number of common friends in her facebook and twitter social network. The graph is shown in figure 3.2. The venn diagram covers the visualization of user's total number of facebook friends, total number of twitter followers and intersection count. In this graph, the diameter of the circle is proportional to the number of facebook friends or twitter followers. We can see in figure 3.2 , number of facebook friends are less than number of twitter followers, resulting into small circle for facebook and large circle for twitter. The number of facebook friends we can see here are the user's friend who have authorized access to their facebook data. As very few people have accessed PERSMA, we could get less number of facebook friend count compared to twitter followers count.

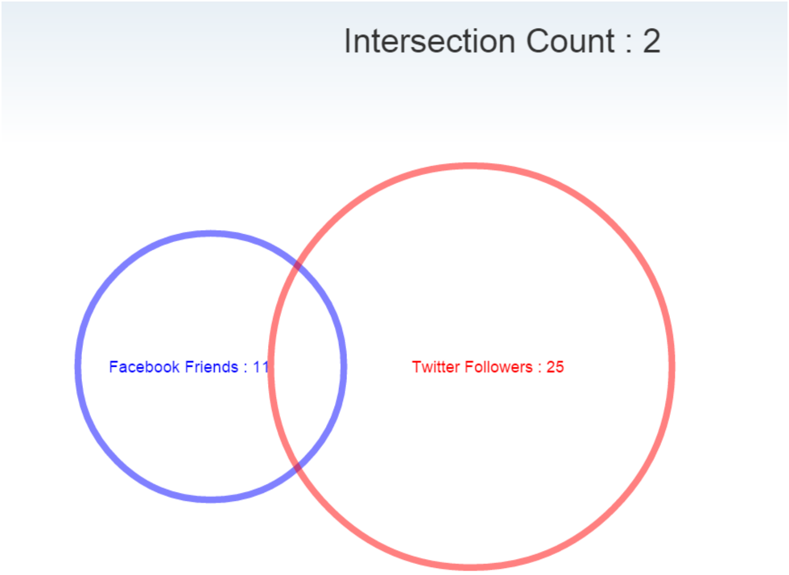


Figure 3.2: Intersection Analysis

3.3 Circle Packing

It is a hierarchical represenattion of twitter follower and sub follower community and provides analysis of the influential people in the same. Each circle represnts a user and the circles within it are its followers. Each circle is of different radius which is propotional to the number of followers each user has.

All followers of a user are fetched from the Twitter API and then sorted in descending order using a comparision function which determine its influence in the network. The function currently simply calculates a users influence by the number of followers it has. This can be further enhanced in the future by taking into consideration their avergae retweet count and other parameters. Thus the ratio of size of the circles represnts the ratio of the number of followers of those users.

The above two images are zoomed in on two different followers shown in 3.3 and you can observe that the circles in the 3.4 have a good follower count ratio compared to the 3.5. This can also help in identifying bots in you follower community. Circles that are immensely larger than the other circles in a particluar user then they have a high probablity of being a bot.

This analysis of follower and sub follower community could be extended to level nth sub

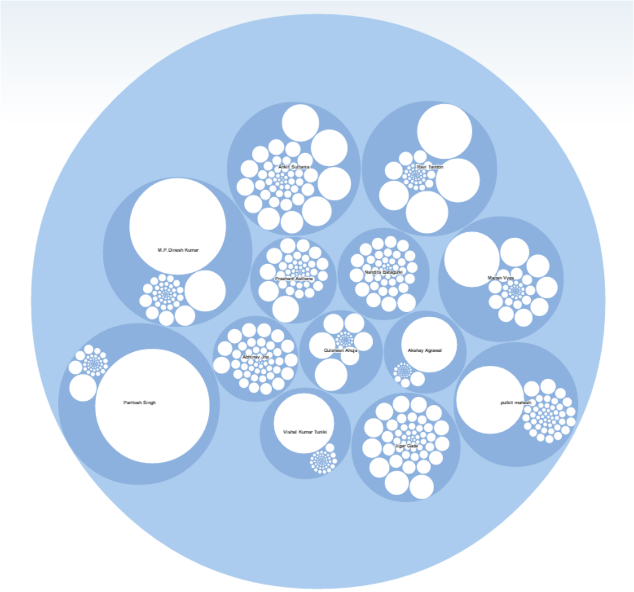


Figure 3.3: Twitter Circle Packing Graph

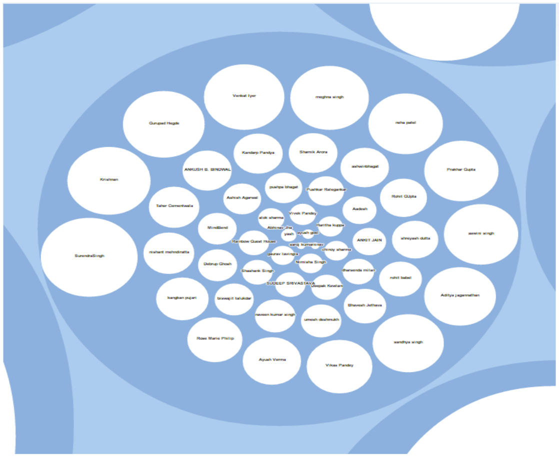


Figure 3.4: Even Circle Sizes

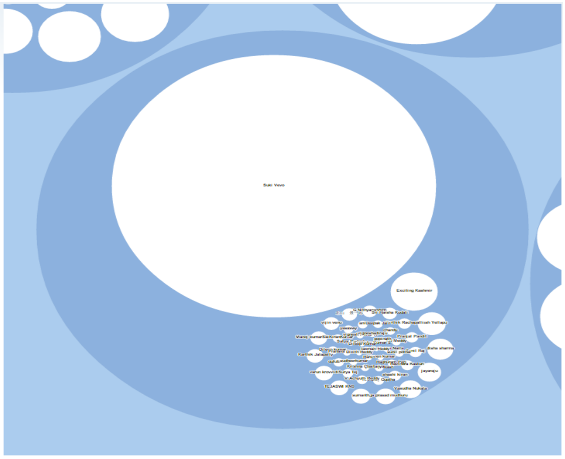


Figure 3.5: Uneven Circle Sizes

followers but generally we are interested in our immediate neighbours than users who are far away in the hierarchical representation. Thus we have restricted the graph to 2nd level which is followers of our followers.

Chapter 4

PERSMA : Prediction Service

4.1 Motivation

As pointed out in the previous reports and the earlier sections, majority of the online social media analysis tools do not have any prediction service built into them for personal analysis of a user. This type of service could be quite beneficial from the user's perspective in the sense that such service can be used to predict the number of retweets a user might get in the near future or predict the number of facebook likes or comments as well. As part of our initial foray into the domain, we tried to incorporate the modeling and prediction of average number of retweets for a given Twitter account.The next sections covers the research, experiments and results for the same.

4.2 Modeling

4.2.1 Linear Regression

It is an approach for modeling the relationship between a dependent variable y and one or more explanatory variables denoted x. In linear regression[1], data are modeled using linear predictor functions, and unknown model parameters are estimated from the observed data. Given a data set of n statistical units $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, a linear regression model assumes that the relationship between the dependent variable yi and the p-vector of regressors xi is linear.

$$y = ax_1 + bx_2 + c$$

Figure 4.1: Linear Regression Equation

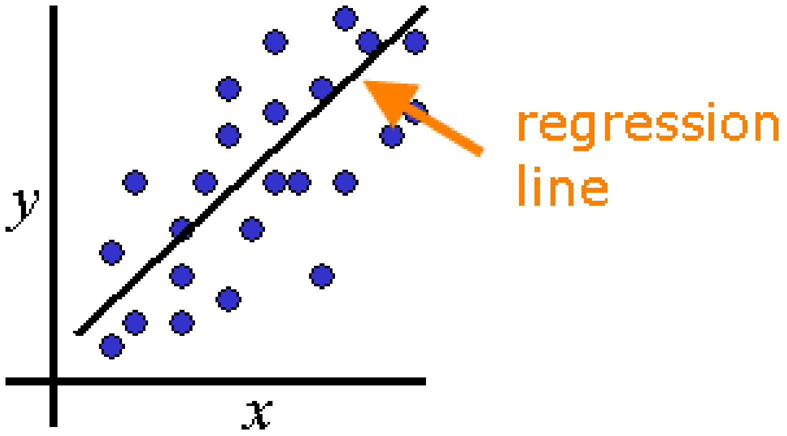


Figure 4.2: Linear Regression Line

4.2.2 Polynomial Regression

It is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial. Although polynomial regression[3] fits a nonlinear model to the data, regression function is linear in the unknown parameters that are estimated from the data.

The higher the degree of the polynomial the better the regression line fits the data but this gives rise to problem of overfitting. Fitting the data very accurately might give rise to steep ends to the regression line and thus small change in the explanatory variable x can give rise to large change in dependent variable y. Thus degree of polynomial regression is restricted by such limitations.

$$y = ax_1 + bx_2 + c + dx_1^2 + ex_2^2 + fx_1x_2$$

Figure 4.3: Binomial Regression Equation

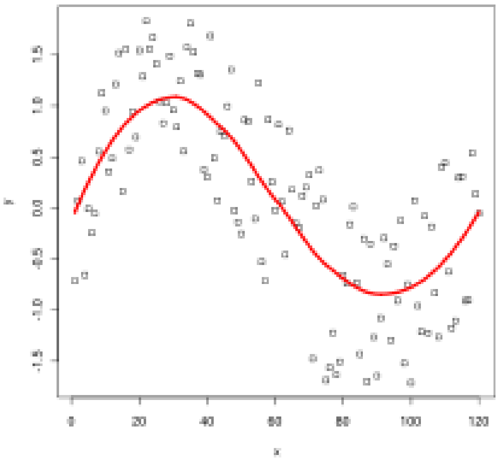
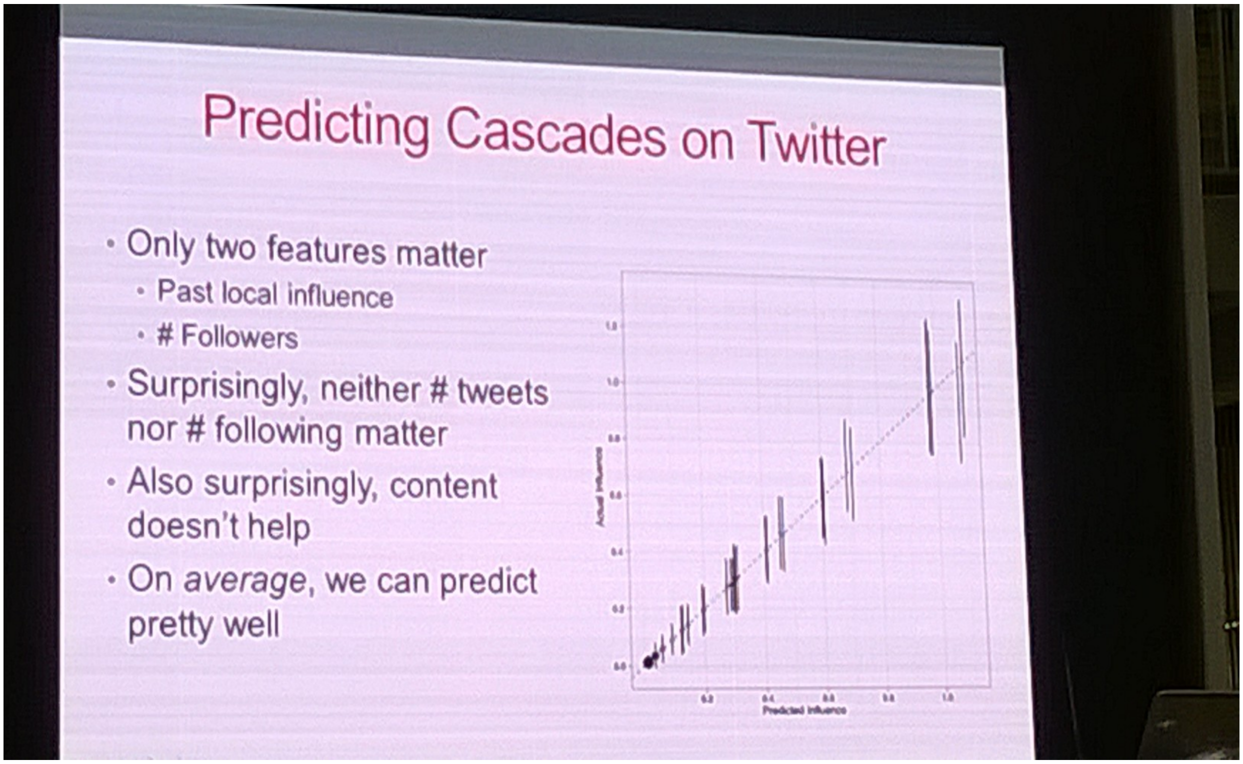


Figure 4.4: Polynomial Regression Line

4.3 Background Research

As part of our initial research in the field, we tried to find any available online tools which have incorporated prediction service into their tool, but could not anything available for free. The major source of inspiration in our research was the talk presented by Duncan Watts on "Social Influence in Markets and Networks" here at USC Annenburg Hall. His talk was mainly based on identifying virality in the social media and networks and try to find the structure of these viral contents in the social network framework. One of the points during his presentation was about predicting cascades on twitter which in turn will lead to know whether a tweet will go viral or not. Following is the screenshot of the slide in which he talked about the same: As we can see in the figure, he clearly pointed out that generally two features are the most important in the prediction analysis of Twitter, those are the Past Local Influence, and the number of followers. It was quite surprising to know that the number of tweets or the number of the people the user is following did not matter at all. Also, the type of the content i.e. whether the tweet contains an url, a video or an image does not affect the prediction model as well. We have included his ideas in our project to create a prediction service for a Twitter user. We used the two main features suggested by Duncan watts for modeling and prediction. The Past Local Influence is taken to be as the past average retweet count of the user. We also tried to incorporate the day



Predicting Cascades on Twitter

- Only two features matter
  - Past local influence
  - # Followers
- Surprisingly, neither # tweets nor # following matter
- Also surprisingly, content doesn't help
- On *average*, we can predict pretty well

Figure 4.5: Prediction Cascades on Twitter

night information of the time of tweet in our model as we thought that will incorporate the demographic information of the follower base which might be an important parameter for the model.

4.4 Data Collection

As mentioned above, the three features we were planning to use for modeling are past average retweet count, historical follower count and timestamp of each tweet.

Twitter API directly provides api's to get the whole twitter feed of any user. The Twitter feed contains a large set of information about the tweet itself, including the content of the tweet, the timestamp, its retweet count, etc. Since, we directly get the retweet count of each tweet for any user and its corresponding count as well, we were able to calculate the past average retweet count by creating different timebins separated by a chosen timestep and putting the retweet count of each tweet inside the corresponding bin based on the timestamp.

Once all the retweet counts of each tweet are binned, we avearge the retweet count inside each bin to get the average retweet count for each timestep.

Same analysis is done to find the average ratio of in terms of day-night tweets for each timestep.

Twitter, unfortunately does not provide API's to get the historical count of the number of followers for each user. It only provides the number of current followers of a given Twitter account. Hence, we could not really integrate the service for each user. Though, there are a few websites which have followed a lot of celebrity twitter accounts from the very beginning and have their historical twitter follower counts. But they provide the last 6 months Twitter follower history for free. Hence, we had to make do with the same. We manually took down the twitter follower count for 5 famous celebrities from different walks of life namely Katy Perry, Barack Obama, Cristiano Ronaldo, Bill Gates and Jimmy Fallon.

4.5 Experiments

We ran a lot of experiments to figure out the final feature set of our model.

- **Tweet Count**: The normal actions that a user can do on Twitter is tweet an original tweet, retweet a tweet by another user or reply to a tweet. Twitter Api provides an option to include/exclude retweets or replies themselves while providing the whole twitter feed for a user. We found out that the retweet count of the retweets done by the user is misleading as it includes the retweets done before the user has retweeted and also the retweets which might have happened not due to the said user retweeting it. Hence we have excluded the retweets done by the user while counting the average rewteet count. We found that the replies of these celebrities also got a lot of retweets from their followers, hence we have included them in our final model.

- **Timestamp**: Initially we were thinking of having a parameter which would have the ratio of day-night tweets of a user to capture the demographic influence of their followers. As explained earlier, we used the timestamp of each tweet to figure out whether a particular tweet occured during the day or night. The timings for the duration which we identified as day were also varied. We tried with day timings as 6am to 6pm and 12 am to 12 pm. We observed that both the combinations did not really have much of a difference in the final model accuracy. Infact, the feature of day-night tweet ratio did not influence the model much as well. Hence, we have excluded this feature from the model entirely.

- **Follower Count**: As told earlier, we were able to manually obtain the historical follower count of five celebrities and use them in our prediction. But we see that the follower count increases on a daily basis but the average retweet count increases or decreases in different timesteps. Hence directly using the follower count did not result in a good model. After this, we used the change in percentage in the current follower count compared to the follower count in the previous timestep as a parameter, but again since the follower count was increasing, the change also always was a positive parameter and did not have much of an effect. After this, we tried instead the rate of change in the follower count in percentage, which means that if the change in follower count in previous timestep was 2% and change in the follower count in the current timestep is 3%, the rate of change of the follower count in percentage for the current timestep would be 50%. We found out that this parameter provided highly accurate model with high prediction accuracy.

- **Timestep**: We also experimented with the timestep of our model as well. We started with a timestep of 1 day, then increased it to 3 days, then to 1 week and then finally to 2 weeks. We did not increase the timestep any further as we had data only for past 6 months and a larger timestep reduced the datapoints for the model itself. We found out that the timestep of 2 weeks gave the maximum accuracy amongs all the experimented timesteps. This might have been due to the reason that the timestep of 2 weeks might have incorporated the cyclical nature of the twitter habits of the users.

To summarize, the features we used to predict the average retweet count of the user in a given timestep are the average retweet count of the user in the previous timestep and the rate of change in the follower growth in the current timestep.

4.6 Results and Integration

The following figure illustrates the the results after we finalized our model. The above figure shows the result of the experiments run on Katy Pery, the one who has the most number

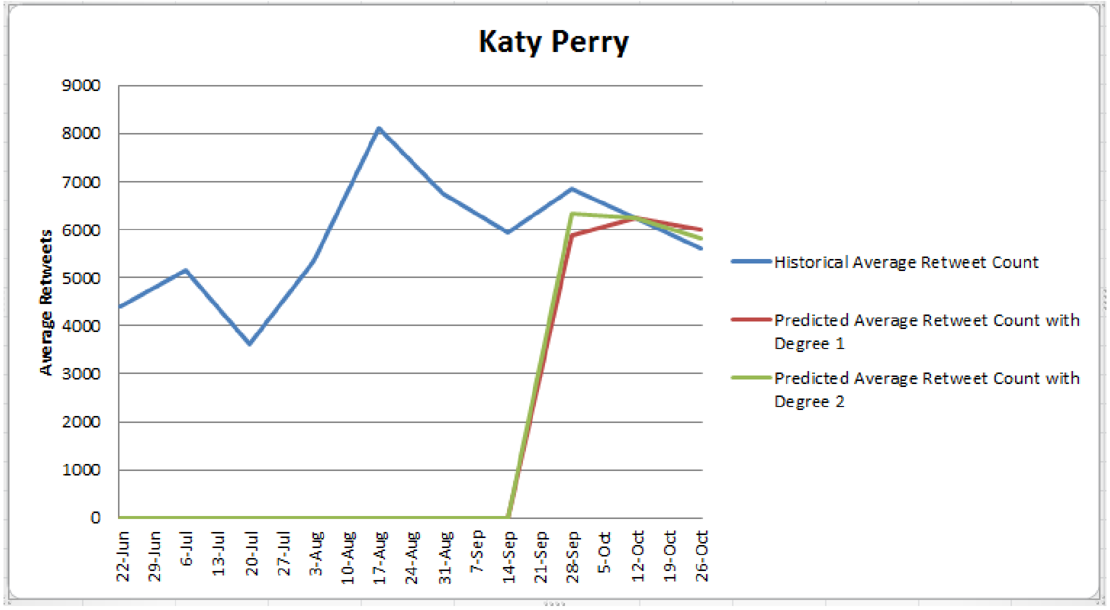
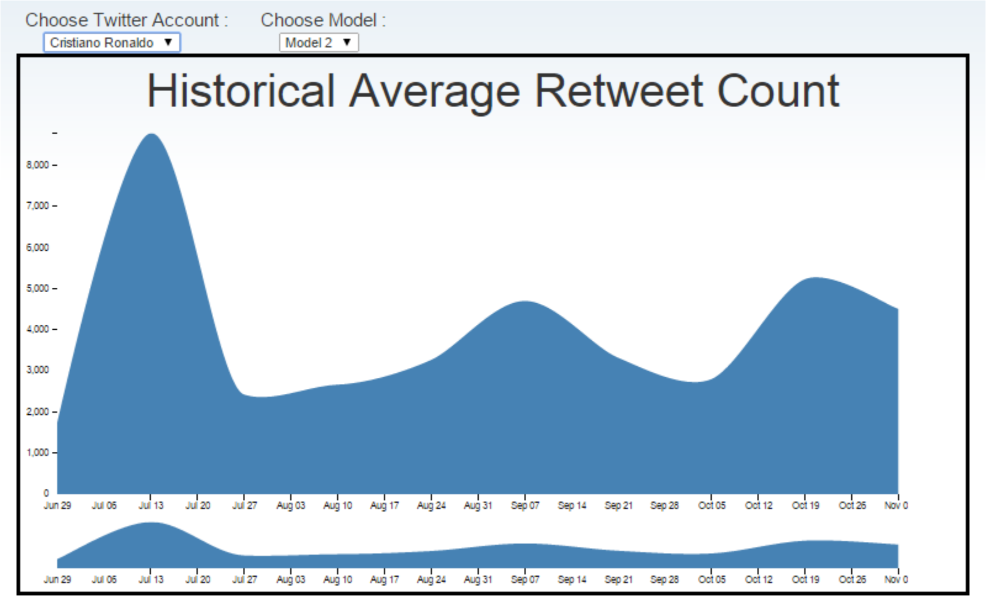


Figure 4.6: Results

of twitter followers. The y axis corresponds to the average retweet count and the x axis corresponds to the date with the minor ticks spaced 2 weeks apart which is the timestep for our model. The blue line shows the historical average rewteet count consisting of 10 datapoints. The red and green line shows the predicted average rewteet count while using the model of degree 1 and 2 respectively. we used the first 7 data points to train the model and then used the last 3 datapoints for verification. We see that both of the lines are quite near to the blue line which suggest a high level of accuracy. Also we see that the degree 2 line is more closer to degree 1 line which suggests that degree 2 models provide a higher accurate model. This is generally the case with regression as generally the higher degree model tends to give a more accurate model but we need to be aware of the overfitting problem which it generates as discussed previously.

Since we were not able to get the historical follower count of each user, we were not able to integrate the prediction service into our website for each user. However, we did upload the data related to the models we generated for our celebrities onto the website for the logged in user to see. When the user clicks the Analysis tab inside the Prediction Service, he is shown three graphs along with the option to choose the analysis for one of the celebrities and select the model. There are two models to choose from, model 1 and model 2 which denotes model of degree 1 and 2 respectively. Figure4.7 shows the Historical Average Retweet Count of

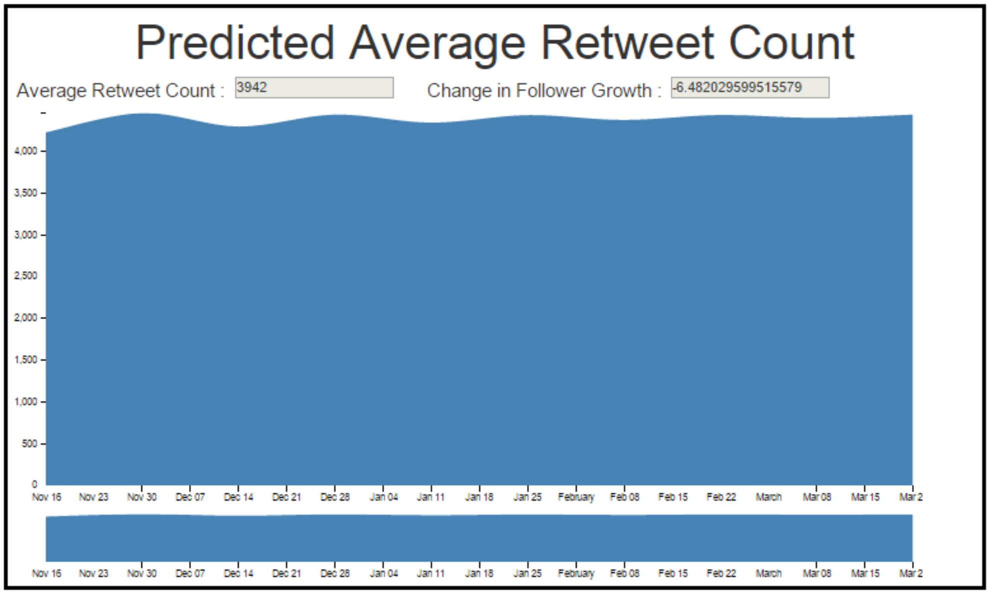


Choose Twitter Account : Cristiano Ronaldo ▼

Choose Model : Model 2 ▼

Figure 4.7: Historical Average Retweet Count

Cristiano Ronaldo. The data consists of only 10 datapoints but the area graph library of the d3 chart engine interpolates the data in between and shows the data as a nice smooth curve. We see that average retweet count of Cristiano is quite high generally and is never lower than 2000. We also see a lot of spiky nature in the graph. We see that he received the most number of rewteets in mid of July this year which coincided with the final week of the Soccer World Cup Finals. This makes sense as he was tweeting on the results and analysis of the most watched sports event and his tweets were generally retweet quite a lot. Figure4.9 shows the Predicted Average Retweet Count of Cristiano Ronaldo using Model 2. It also shows two values one of which is the Average Retweet Count which is the average retweet count of all the tweets of Cristiano Ronaldo over the past 6 months. Similarly, Change in Follower growth is the average of the change in follower growth over every datapoints. We then use the last timestep's average rewteet count and take the last week's rate of change in follower growth as the values to input into the model and then use it to predict the current timestep's average retweet count. We then use the current timestep's retweet count and update the last week's rate of change by adding to it the average rate of change and then use both of these values to generate the next timestep's retweet count. This process is repeated in total 10 times to generate the next 10 predicted values. We see from the graph that the values are quite similar to the average retweet count values.

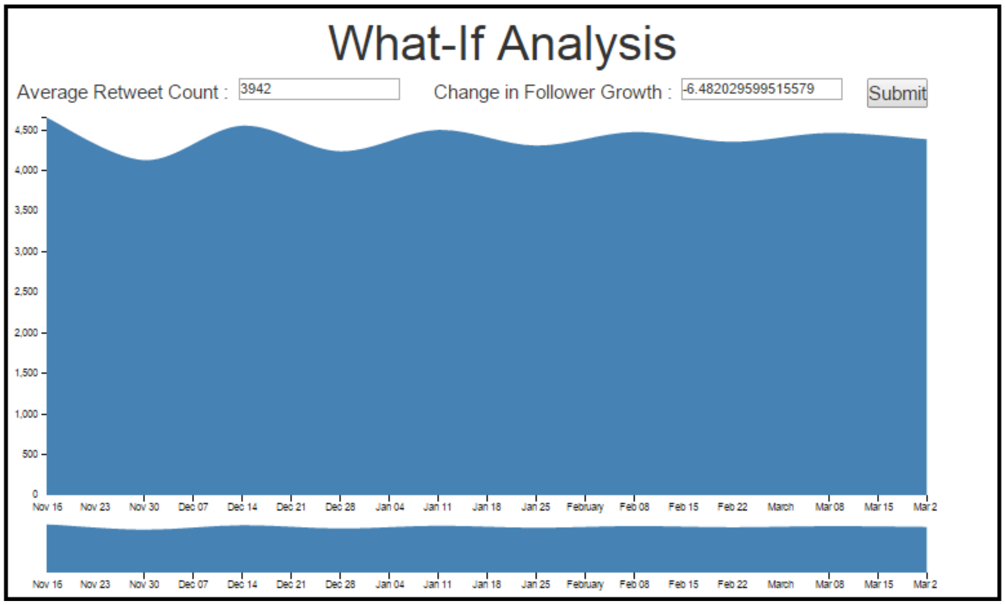


Predicted Average Retweet Count

Average Retweet Count : 3942

Change in Follower Growth : -6.482029599515579

Figure 4.8: Predicted Average Retweet Count



What-If Analysis

Average Retweet Count : 3942

Change in Follower Growth : -6.482029599515579

Figure 4.9: Predicted Average Retweet Count

Figure4.9 shows the Predicted Average Retweet Count of Cristiano Ronaldo using Model 2. It also shows two values one of which is the Average Retweet Count which is the average retweet count of all the tweets of Cristiano Ronaldo over the past 6 months. Similarly, Change in Follower growth is the average of the change in follower growth over every datapoints. It differs from the previous graph in the sense that the user now has to option to change the input to the model by modifying these values and then clicking the Submit button. These changes values are given as the starting input to the model and then rest of the graph is calculated in the same way as the previous graph.

Chapter 5

Summary

5.1 Challenges

Apart from the implementation challenges for the tool, there were lot of issues regarding accessing the Facebook and Twitter Api's via REST, which are relevant for data mining. These are described as follows:

- **Restricted API's**: Even after providing access to the tool via twitter or Facebook api's, the tool might not get the user's friend's data because it might be possible that her friends have blocked their private data. This doesn't give the full analysis of user's social network.

- **Rate Limits**: Twitter puts rate limit for most of its api's to restrict the access. These limits are really very low in the range of 15 queries/15 min. These low rate limits restricts the real time analysis of the network.

- **Privacy Issues**: Due to the growing concern regarding the privacy of its users, Facebook has blocked its apps to fetch the whole friend list of the authenticated user. The apps will only get the list of those friends who also have authorized the app themselves. This, as we can image, really restricts the analysis of the user's network.

- **Lack of Data Point**: For generating perdiction models we need a lot of observed data points in past for number of follower and the average retweet count. The lack of historical count of twitter followers made it difficult for generating good models for predictions.

5.2 Conclusion

This project involved understanding different types of two social network structures, Twitter and Facebook, and extracting user data from these networks using their APIs. We also learnt about interesting graphs available in the D3 charting library and how they can be used for personal analysis of one's social network profile.

As part of this project, we came up with a website for the tool which allowed the user to login via Twitter and Facebook to view graphs about mutual relationship among her friends and followers and other inetresting charts combining followers. We also tried to predict user's data based on her past data from Twitter which gave encoruaging results.

The future work for the tool includes enhancing the tool's analysis capabilities by generating reports offline to overcome the API rate limitations. We shall enhance the analysis of integrated data from Facebook and Twitter. We shall also enhance the predictive analysis of retweet count by extending it to every user by tracking the change in follower count for the registered users.

Bibliography

[1] Linear Regression. http://en.wikipedia.org/wiki/Linear_regression.

[2] Persma Website. http://tinyurl.com/persma.

[3] Polynomial Regression. http://en.wikipedia.org/wiki/Polynomial_regression.