Analysis of Social Media Data to determine Positive and Negative Influential Nodes in the Network

Report Submitted in partial fulfillment
Of the requirements for the degree of
Integrated Master of Science
In
Mathematics & Computing
By

Shubhanshu Mishra
(07MA2023)

Under the supervision of

Prof. Gloria Ng
Prof. Pawan Kumar

Department of Mathematics
Indian Institute of Technology, Kharagpur
West Bengal, India - 721302
May 2012

DECLARATION BY STUDENT

I certify that

a. The work contained in this report has been done by me under the guidance of my supervisor(s).
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
d. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date:__________________

____

Student Signature

CERTIFICATE BY SPERVISOR

This is to certify that the project report entitled **Analysis of Social Media Data to determine Positive and Negative Influential Nodes in the Network**, submitted by **Shubhanshu Mishra,** in partial fulfillment of the requirements for the degree of Integrated Master of Science in Mathematics & Computing to Indian Institute of Technology, Kharagpur, is a record of bona fide project work carried out by him under my supervision.

Date : ___________________________ **Superviser : __________________________**

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Gloria Ng & Prof. Pawan Kumar, for being my advisors in this project and letting me work on a topic of my interest. I am indebted to their support during the course of the completion of my project in the Department of Mathematics at IIT Kharagpur.

I would also like to thank my team from Institute of Systems Science, National University of Singapore comprising of Choo Zhi Min and Ding Yi for helping me with the development of the idea of my project work and also providing me with necessary research material and discussion during the course of my project.

I would also like to extent my sincere gratitude to my family members and friends who have been endless source of encouragement for me.

Last but not the least I would like to thank the authors of all the references I have cited and also the developers of services I have used in the development of my project.

ABSTRACT

Human behaviors have always been of great interest to researchers and scientists. Inferring positive and negative sentiments from conversation has been something which mankind has always had interest from the past. With the advent of social media channels the amount of information we are getting about individuals and their behaviors is huge. Social media has brought the conversations of people as text which can be processed and inferences can be made. Currently there exists a lot of work in the field of sentiment analysis and there have been a few web services which utilize sentiment analysis for influence ranking and brand reception calculation. In my project, however, I have tried to utilize the conversations people create on social media to define their positive influence in their network not only regarding brands and topics but also on an emotional level. I have developed an scheme to determine to overall positive negative influence index of a social media identity and use it determine their ranking amongst the most influential infinities in their network under both positive and negative aspects.

TABLE OF CONTENTS

1. Introduction:

Social Media is a really active field these days with the invention of platforms like Blogs, Facebook, Twitter and the amount of data people are sharing every day. Users of these platforms share their personal, professional lives on the social media channels. This humongous data gives us access to huge amount of insight into how people interact on social media channels. The interactions can be classified into positive and negative.

We define positive interactions as interactions where people spread positive ideas denoting happiness, motivation, interesting facts. These interactions can be predicted by the kind of comments people leave. Words like “Cool”, “Awesome”, “Great”, etc. denote a really positive interaction. Positive interactions also invite a lot of traffic from people who usually don’t know you because a positive thought always affect everyone and act in a magnetic manner.

On the other hand negative interactions are those where people usually are involved in the abuse, demotivation, arguments etc. These interactions can be predicted by a huge amount of negative traffic. Words which involve abuse, disgust, racism, regionalism, hate etc. are very prominent in these interactions. Negative interactions usually drive comments from a specific sector of the network of the person, who has an attachment to the topic.

This project aims to define the various types of positive and negative interactions happening in the social media network of a person and segregate the people who are involved in positive or negative interactions. This division will help us identify the various level of positivity/negativity in the discussions. This will also help us position people on the basis of their impact in terms of positivity/negativity in the social network of the people they are attached with.

The second objective of this project will be to define the overall influence of a person in the network we are observing. This will help us find out who are the most influential people in the network both in the positive and the negative sector. The information can be very useful to marketing companies or advertisers who need to target these influential set to get a greater market share through word of mouth publicity in the network of the most influential people in the network.

The project will involve an algorithmic approach towards mining data, analyzing it and the ranking of people on the basis of the impact of their shared content. It will depend on the co-creation model of information sharing among the people through which we will invite people to take part in the study by submitting their Social Media News with us. We will also scour the open data available in Blogs, Twitter etc. A person will be denoted by all the social media channels he/she is connected to and we will do a collective analysis on that person based on their impact on various social media communities and their overall influence level.

Not much research has been done specifically in this topic. However there have been a few works in the field of Social Network analysis, Social Media Data Mining and Relationship analysis on Social Media.

2. **Problem Definition:**

Given a social network of individuals we try to find out 3 quantities about the influence of the individual **[a]** in the network. These the quantities are the following:

1. Total Positive influence – $\mathbf{P_a}$
2. Total Negative influence – $\mathbf{N_a}$
3. Overall influence – $\mathbf{I_a}$

The influence will be calculated to the conversations created by the individuals between time $\mathbf{t_0}$ to time $\mathbf{t}$ which for our study will not be used to affect the influence rankings and the overall influence values.

For this problem we have simply considered the twitter timeline of users who are participating.

The idea is to create an absolute influence value and a relative influence value of individuals in the network as social media is a relative judgment place. Every individual is given a rank in the system and also given an absolute and a relative influence values. The influences are classified on the basis of positive and negative opinions as used in the sentences. Every sentence used by the individual in their social media channels will be classified as either positive or negative. At the end of each activity the influence value of the individual will change accordingly.

The ranking will be based on the network of the individual and there will also be a system wide ranking however since the algorithm is very robust hence the network based ranking is simply extrapolated to gain the system wide ranking.

3. **Sentiment Analysis:**

Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication.

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

Example:

Fact: - "The painting was more expensive than a Monet"
Opinion/Sentiment: - "I honestly don't like Monet, Pollock is the better artist"

There are several reasons for doing sentiment analysis:

It allows business to track:

- Flame detection (bad rants)
- New product perception
- Brand perception
- Reputation management

It allows individuals to get: An opinion on something (reviews) on a global scale.

4. Opinions:

An opinion is defined as "A personal belief or judgment that is not founded on proof or certainty" (WordNet). But, “The fact that an opinion has been widely held is no evidence whatever that it is not utterly absurd.” (Bertrand Russell).

Opinion for a machine:

It is a "quintuple", an object made up of 5 different things:

$$\mathbf{(O_j, f_{jk}, SO_{ijkl}, h_i, T_i)}$$

$\boldsymbol{O_j}$ = The thing in question (i.e product)
$\mathbf{f_{jk}}$ = a feature of Oj
$\mathbf{SO_{ijkl}}$ = the sentiment value of the opinion of the opinion holder hi on feature $\mathbf{f_{jk}}$ of object $\boldsymbol{O_j}$ at time t_l

These 5 elements have to be identified by the machine
{Defined by Bing Liu in the NLP handbook}

$\mathbf{O_j}$ = Named Entity Extraction
$\mathbf{f_{jk}}$ = Information extraction
$\mathbf{SO_{ijkl}}$ = Sentiment analysis
$\mathbf{h_i}$ = Information extraction
$\mathbf{T_i}$ = Data extraction

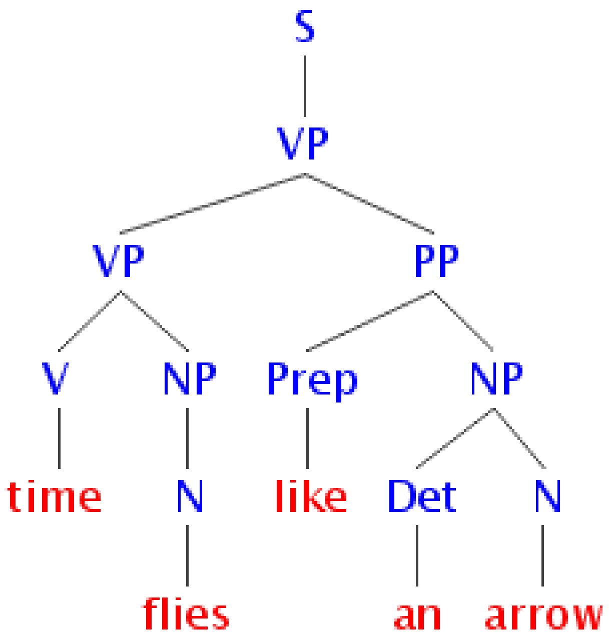
Language is ambiguous and hence we need to be careful while classifying statements for positive and negative and sometimes incorrect. Consider:

- "The watch isn't water resistant" - In a product review this could be negative.
- "As much use as a trapdoor on a lifeboat" - negative but not obvious to the machine.
- "The canon camera is better than the Fisher Price one" - comparisons are hard to classify.
- "imo the ice cream is luuurrrrrrvely" - slang and the way we communicate in general needs to be processed.

5. Sentiment Analysis Process:

1. Part-of-speech tagging (but also position and more):

The word in the text (or the sentence) at tagged using a POS-tagger so that it assigns a label to each word, allowing the machine to do something with it. It looks something like this:



S = subject
VP = Verb Phrase
V = Verb
N = Noun
NP = Noun Phrase
PP = Preposition
Det = Determiner

Then we extract defined patterns like [Det] + [NN] for example

2. Sentiment Orientation:

We look at sentiment orientation (SO) of the patterns we extracted. For example we may have extracted:

Amazing + Phone

This is:

[JJ] + [NN] (or adjective followed by noun in human)

The opposite might be "Terrible" for example. In this stage, the machine tries to situate the words on an emotive scale (so to speak).

3. Average Sentiment Orientation:

The average Sentiment orientation of all the phrases we gathered is computed. This allows the machine to say something like:

"Generally people like the new iphone"
=>They recommend it

OR

"Generally people hate the new iphone"
=> They don't recommend it

4. Classifying Sentiments:

This is very difficult but experiments have been done using:

- **Naive Bayes** (probabilistic classifier using Bayes theorem)
- **Maximum Entropy** (Uses probability distributions on the basis of partial knowledge)
- **Support vector machine** (Data is set as 2 vectors in an n-dimensional space) Pang et al. found the SVM to be the most accurate classifier (around 80%).

There are other methods being explored as well.

6. Methodology:

In order to achieve the objectives of the project, I have identified the various parameters to be considered for defining the influence of an individual and given weightage to them. The weights keep changing as the application grows in the number of factors being considered.

The feeds which we have considered for the 1st phase of testing of the application are of the twitter profile of the individuals. The reason for choosing twitter profiles is that they are open and easily accessible and give us a really simple data set to be analyzed.

First of all we start by defining the influence group for the user. For this we have considered the following circles of influence:

- Followers of the User
- Profiles the User is following
- Individuals in the follower list in the same city as the user
- Individuals mentioned by the user in tweets

Influence of a profile in a given circle is defined by the function f_i (for twitter profiles) following equation:

$$f_i: (m+ rt)/n$$

Where:

f_i: *influence of user in the circle*
m: *Total mentions of the user*
rt: *total re-tweets of the user's content by people*
n: *total number of individuals in the circle*

For deciding the overall influence we have taken the weights of the possible parameters.

The mathematical model for defining influence is through 2 arrays

- Array containing the influence value in each circle defined: $\boldsymbol{A_i}$
- Array containing weights of each influence circle: $\boldsymbol{A_w}$
- Array containing value of n for each circle: $\boldsymbol{A_n}$

The above data helps us in defining the absolute influence of a user on our network. However this data is not very beneficial as influence is a relative aspect and we should be considering total influence as a relative quantity between the most influential and least influential user in the network.

The total absolute influence ($\boldsymbol{I_a}$) of an individual throughout our network can be found using:

$$I_a = \frac{\sum_{j=1}^{N} A_i[j] * A_n[j] * A_w[j]}{\sum_{j=1}^{N} A_n[j] * A_w[j]}$$

This total absolute influence will be used in calculating the Overall Influence of user on the network ($\boldsymbol{I_o}$) which will be the quantity mostly beneficial for the users. In order to define the overall influence we have to consider 2 more quantities:

$\boldsymbol{I_{a,\ max}}$**:** Maximum absolute influence in the network

$\boldsymbol{I_{a,\ min}}$**:** Minimum absolute influence in the network

This will help us in determining the Overall Influence of user on the network ($\boldsymbol{I_o}$) which will be given by the formula:

$$\boldsymbol{I_o = \frac{I_a - I_{a,min}}{I_{a,max} - I_{a,min}}}$$

7. Sample Data Sets:

In solving our problem we have considered the sample data set which consists of the following social media sites and their data:

Site	Users
Twitter	380 million
Facebook	800 million

From these we can get the various interactions happening on the social media sites.

For our project we have used the data from Twitter over Facebook majorly for the following reasons:

1. Twitter data is has fewer properties and lesser interaction points. Meaning ever tweet can either be replied to or re-tweeted as opposed to Facebook interactions where each interaction has properties like comments, sharing, likes and likes on comments, which make the data a bit difficult to use.
2. Twitter data has a fixed character limit which helps in designing easy algorithms and judging influence better as everyone is judged on how they are influencing per 140 characters.
3. Twitter dada is always text and hence can easily be processed to determine sentiments using simple Language Processing techniques, Facebook data on the other hand can be text, pictures, videos, application posts etc. which are difficult to process as per the limitations of our algorithm.
4. Overall daily postings on twitter by an individual are higher as compared to that on Facebook which makes the inferences drawn each day to be better.

Also for classification we have considered the following data set of positive and negative keywords as base along with all their synonyms and slangs:

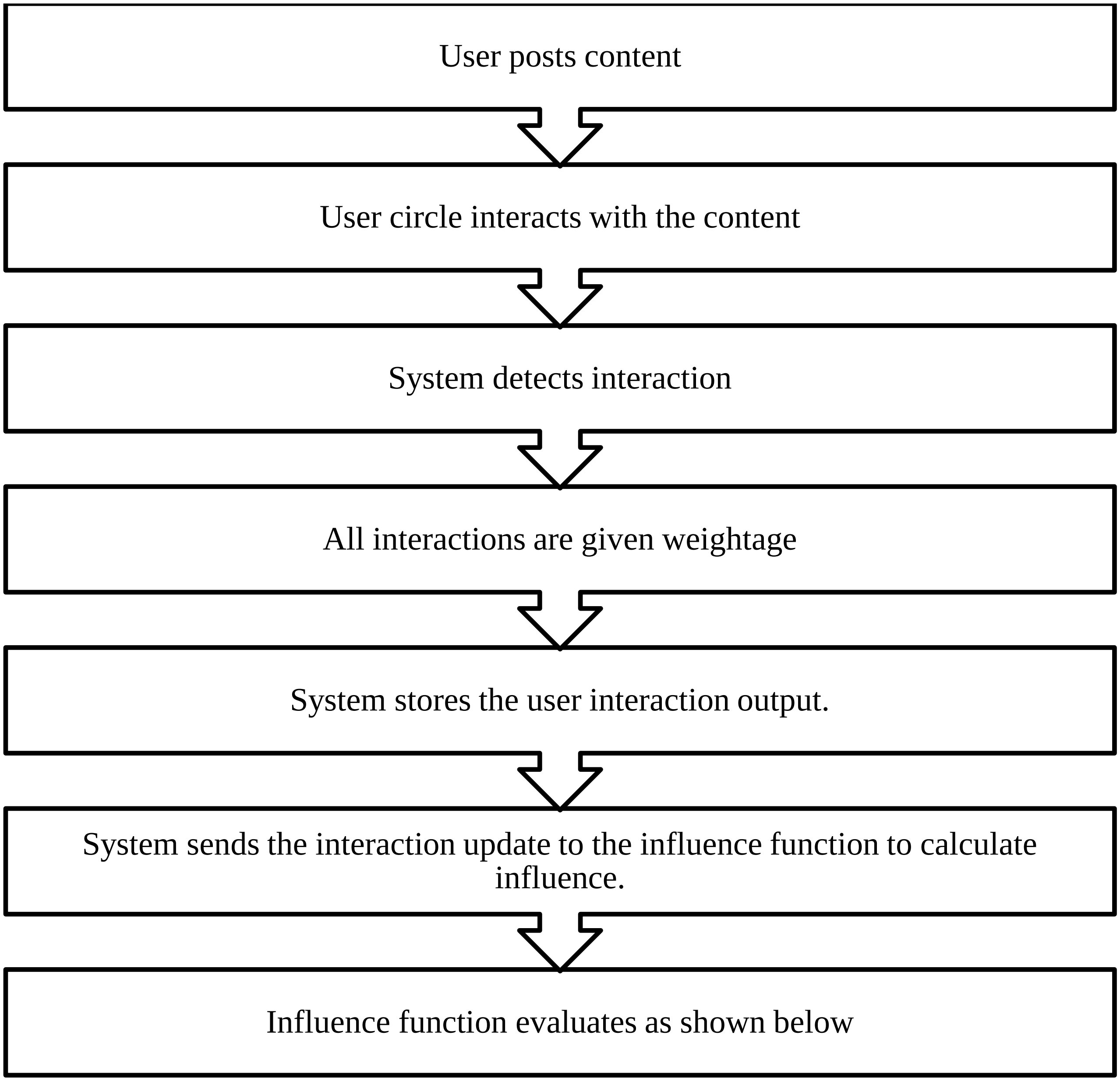
Positive Words	Negative Words
agreeable, alert, alluring, ambitious, amused, boundless, brave, bright, calm, capable, charming, cheerful, coherent, comfortable, confident, cooperative,	angry, annoyed, anxious, arrogant, ashamed, awful, bad, bewildered, black, blue, bored, clumsy, combative, condemned, confused, crazy, flipped-

courageous, credible, cultured, dashing, dazzling, debonair, decisive, decorous, delightful, detailed, determined, diligent, discreet, dynamic, eager, efficient, elated, eminent, enchanting, encouraging, endurable, energetic, entertaining, enthusiastic, excellent, excited, exclusive, exuberant, fabulous, fair, faithful, fantastic, fearless, fine, frank, friendly, funny, generous, gentle, glorious, good, happy, harmonious, helpful, hilarious, honorable, impartial, industrious, instinctive, jolly, joyous, kind, kind-hearted, knowledgeable, level, likeable, lively, lovely, loving, lucky, mature, modern, nice, obedient, painstaking, peaceful, perfect, placid, plausible, pleasant, plucky, productive, protective, proud, punctual, quiet, receptive, reflective, relieved, resolute, responsible, rhetorical, righteous, romantic, sedate, seemly, selective, self-assured, sensitive, shrewd, silly, sincere, skillful, smiling, splendid, steadfast, stimulating, successful, succinct, talented, thoughtful, thrifty, tough, trustworthy, unbiased, unusual, upbeat, vigorous, vivacious, warm, willing, wise, witty, wonderful	out, creepy, cruel, dangerous, defeated, defiant, depressed, disgusted, disturbed, dizzy, dull, embarrassed, envious, evil, fierce, foolish, frantic, frightened, grieving, grumpy, helpless, homeless, hungry, hurt, ill, itchy, jealous, jittery, lazy, lonely, mysterious, nasty , naughty, nervous, nutty, obnoxious, outrageous, panicky, repulsive, scary, selfish, sore, tense, terrible, testy, thoughtless, tired, troubled, upset, uptight, weary, wicked, worried

8. User Interaction:

The initial user interactions will be sending to our system which will do some pre analysis on the data and finally send the data to the Influence function.

This strategy is based on recording each user interactions.



User posts content

User circle interacts with the content

System detects interaction

All interactions are given weightage

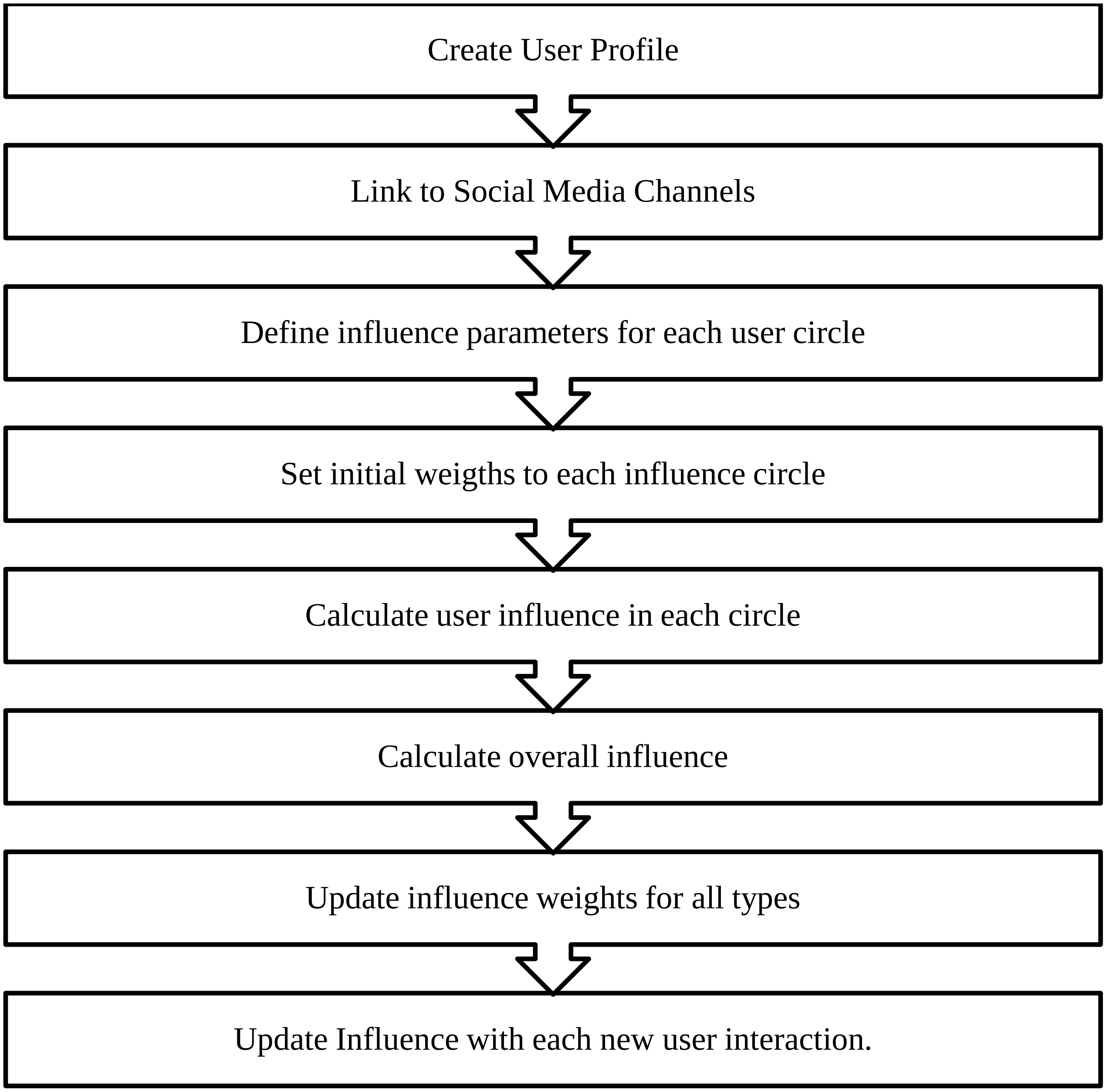
System stores the user interaction output.

System sends the interaction update to the influence function to calculate influence.

Influence function evaluates as shown below

9. Process Diagram:

A web application was developed which followed the following flowchart in calculating the influence of each user profile.


Create User Profile
Link to Social Media Channels
Define influence parameters for each user circle
Set initial weigths to each influence circle
Calculate user influence in each circle
Calculate overall influence
Update influence weights for all types
Update Influence with each new user interaction.

10. Applications:

There can be following very good uses of the algorithm:

1. Identifying top consumers of a product for direct marketing, here in every company can find out who are the top influencers in the market sector they want to tap. Using this, the company can just try selling their products to the top influencers and once that individual recommends the company product using their social media channels then the overall acceptance and appreciation of the product increases. It will also help in identifying which consumers not to engage while publicizing, usually for the negative influencing participants in the network who are top negative influencers.

2. Deriving inferences of public sentiments regarding certain news and event as for every news the people will react in a certain manner. The important thing is top influencers in a region have a great power to drive the sentiments in positive or negative direction strongly. A good way of using this service will be that you identify the top positive influencers and request them to handle the situations and also get to know about the top negative influencers and prepare accordingly for actions they can take.

3. Improving gamification platforms using social media data can be another very great way of utilizing this service. Gamification platforms depend on the strategy of reward and promotion and hence user interaction can be clubbed to reward the genuine candidates in a better manner and make the process fair as many times cheating and forgery is hard to catch, but gaining information from social media channels about people can be very helpful.

4. Human Resources and recruitment tools for organizations can utilize this service for prescreening candidates on how their behavior is in their circle and how fit are their communication style for the company.

11. Possible Improvements

1. Add functionality for Facebook and other social media channels like blogs, Flickr, YouTube as this will give a more comprehensive influence value of the individuals participating in a circle.
2. Define weightage functions for improving weights as different types of posts can have different influence like test status updates have less influence than videos which have less influence than pictures.
3. Add functionality for considering more impressions and interactions like comments, likes sharing on social media sites in making influence value more useful as this will also determine the mutual relationship between an individual and the people they influence.
4. Data can be analyzed in terms of specific terms or values like: trending topics, recent happenings and for each type there can be an influence metric.
5. Overall influence for each type of parameter can be also listed down like keywords, location, and age group.

12. Conclusion

Social media is becoming a very important part of communication for people and it’s important that we develop our systems to utilize this mode of communication for better learning and great value for the people. Marketing is growing at a rapid pace and word of mouth still remains the most trusted mode of marketing. Hence through this project we have tried to identify the opportunities for people to utilize the social media activity of their clients and relevant individuals and utilize them to give them better service and also improve their own processes.

The project has helped in developing an influence calculation algorithm and a ranking algorithm to find out who the most influential positive and negative influential nodes are in a given network. The idea has a wide area of applications in marketing and human resource industry and politics.

With more and more social media channels coming and mode diverse amount of information getting shared there is a huge opportunity to tap into the data of people and develop better systems which will be able to predict and determine human behaviors in a better manner in future.

13. References:

- Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews - Peter D. Turney, Institute for Information Technology
- Sentiment analysis in Text (SFS) -http://www.scienceforseo.com/opinion-mining/sentiment-analysis-in-text/
- Opinion mining and sentiment analysis (Bo Pang, Lillian Lee) - http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html
- Opinion Extraction, Summarization and Tracking in News and Blog Corpora (Ku, Liang, Chen) - http://research.microsoft.com/apps/pubs/default.aspx?id=65490
- Sentiment analysis and subjectivity (Bing Liu) - http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf
- International sentiment analysis for news and blogs (Bautin) - http://www.cs.sunysb.edu/~mbautin/pdf/int_senti_analysis.pdf
- Sentiment analysis: does coreference matter? (Nikolov) - http://www.francosalvetti.com/Sentiment-and-Coreference.pdf
- CIKM Workshop on sentiment analysis - http://sites.google.com/site/tsa2009workshop/
- National Research Council of Canada, Ottawa, Ontario, Canada, K1A 0R6
- Graph mining applications to social network analysis - Lei Tang and Huan Liu (http://www.public.asu.edu/~ltang9/papers/graph_mining.pdf)
- Social Network Analysis and Mining for Business Applications - Francesco Bonchi, Carlos Castillo, Aristides Gionis, And Alejandro Jaimes, Yahoo! Research Barcelona
- Information Propagation and Network Evolution on the Web - Jure Leskovec, Mary mcglohon, Christos Faloutsos, Natalie Glance, Matthew Hurst (http://www.ml.cmu.edu/research/dap-papers/mcglohonkdd.pdf)
- Facebook Data taken from: https://www.facebook.com/press/info.php?statistics
- Twitter Data taken from: http://en.wikipedia.org/wiki/Twitter

14. Resources for Analysis:

1. SentiWordNet - http://sentiwordnet.isti.cnr.it/
2. LingPipe sentiment analysis - http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html
3. Long list of tools at CodeSpeak - http://lordpimpington.com/codespeaks/drupal-5.1/?q=node/5
4. The Toolkit for Advanced Discriminative Modeling (TADM) - http://tadm.sourceforge.net/
5. RapidMiner - http://rapid-i.com/content/view/55/85/