MODELLING FOR SENTIMENT ANALYSIS IN ENTERPRISE SOCIAL NETWORKING APPLICATION

BITS ZG628T: Dissertation

by

\<Your Name\>

\<Your ID\>

Dissertation work carried out at

\<Company, Place\>

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)**

November 2017

MODELLING FOR SENTIMENT ANALYSIS IN ENTERPRISE SOCIAL NETWORKING APPLICATION

BITS ZG628T: Dissertation

by

<Your Name>

<Your ID>

Dissertation work carried out at

<Company, Place>

Submitted in partial fulfillment of
M.Tech. Software Systems degree programme

Under the Supervision of
<Supervisor Name>
<Company, Place>

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)**

November, 2017

CERTIFICATE

This is to certify that the Dissertation entitled **Modelling For Sentiment Analysis In Enterprise Social Networking Application** and submitted by **Rakesh M Nair** having ID-No. **2015HT13540** for the partial fulfillment of the requirements of **M.Tech. Software Systems degree of BITS**, embodies the bonafide work done by him under my supervision.

_____

Signature of the Supervisor

Place:

Date :

_____

Name, Designation & Organization &Location

Abstract

Sentiment Analysis in any social network or media is a way to determine the opinion polarity on the topic. Generally, they classify the opinion as POSITIVE or NEGATIVE. An Enterprise Social Network (ESN) systems are one deployed in any corporate for their employees to collaborate and take decisions quicker.

Socialcast™ , a VMware ESN product used by many other companies. This product is sold as a cloud based service as well as an On-premise installation model. I have used cloud based product to build the sentiment analysis solution.

Given the nature of any corporate structure, many of the opinions get biased due to the authority of the person. Hence simple Sentiment analysis on any of the discussion shall not yield correct results many time. In this work, I have brought in the "influence factor" in the sentiments to decide on the overall sentiment score of the discussion. To calculate the Influence factor, I used the history of the user's activities in the system along the line of the similar actions and their authority level.

Acknowledgements

<WRITE YOUR OWN>

List of Abbreviations and Acronyms

CPU	Central Processing Unit
DB	Database
ESN	Enterprise Social Network
NLP	Natural Language Processing
PHP	Hypertext Preprocessor (earlier called, Personal Home Page) A Scripting programming language
RUBY	A scripting programming language
SO	Social Object
UI	User Interface

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Overview

Enterprise Social Networking applications are gaining popularity within the corporate these days. This platform allows employees to collaborate through many forms like starting a discussion on the work, sharing information and documents with co-workers and also chatting with others either in group chat or One-to-One chats. These discussions are made available through the news feed, which appear in the respective members', who are part of the discussion, home stream page and corresponding group page. Given the nature of such open platform, employees could use it anyway they want and hence it is mandatory to analyze the content to understand usage of this platform within the organization. Since the volume of content generated is large, analyzing message by message will be time consuming and CPU intensive. Use of Sentiment analysis will be a better option to build a near real time system.

Socialcast™ is an in-house social networking product. This platform has the capabilities for the employees to create a focused group and start the discussion in either micro-blogging style or as chat. It also provides capabilities to share the files through such groups. Employees could seek attention of other employees in the group by "at mentioning" them in the discussion.

Socialcast platform allows enterprise provides an open culture within the company for the employees to collaborate openly. In nut shell, Socialcast has the features to start a Chat group or Discussion Group or one-to-one chats. The groups can be of Public access level or Private. One-to-One are by default at Private level non-discoverable. The Public groups are discoverable (by build in search feature) and one could join the group either freely or through requests. Private groups are non-discoverable and one can join only through request or by Admin of the group adding them directly.

The chat groups and one-to-one chats have chat like interface and the purpose of such are to do Ad-hoc communications like quickly getting onto a specific topic and discuss with another set of members in the group. Given ad-hoc nature of the discussion like any chat application, understanding the flow discussion is really difficult as one could an answer someone's query or statement at different time period by when the discussions would have moved on.

The Discussion Groups are meant for focused discussions on a specific subtopic the group intended for. For example, Architects group could have discussion on multiple subtopics including technology trend for a general problem, and could also have discussion the architecture discussion on integrating third party system. These subtopics in Socialcast term is called a Social Object. A Social Object contains a POST (original discussion points) and COMMENTS (opinion or responses from the team members). At present Socialcast supports comments only at one level. It also contains user reactions like Liking the POST or COMMENT
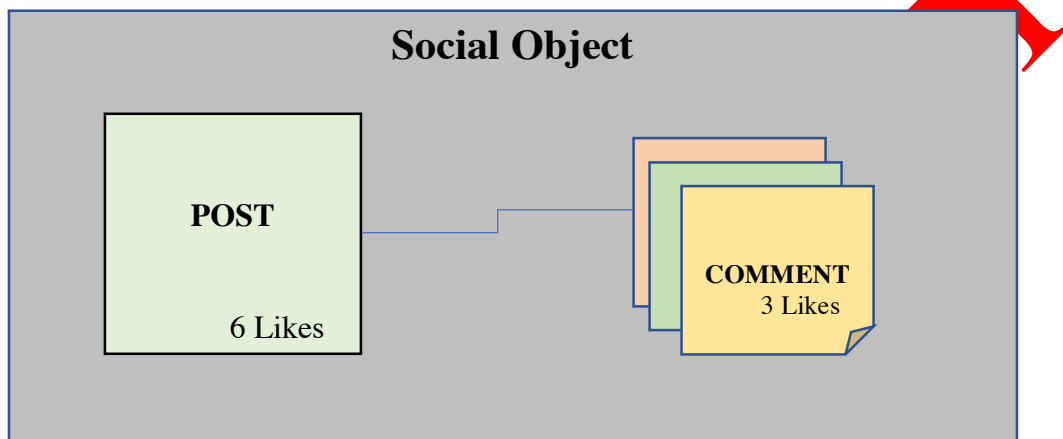
Figure 1 - SOCIAL OBJECT is a combination of a POST, its associated COMMENETs and LIKES associated with POST & COMMENTs

1.2 Problem Statement

A User of the Enterprise Social Network platform like Socialcast create Groups by giving name to the group and adding relevant users to the group for specific purpose like say Design reviews, Architecture Reviews or could be for organizing a fun event. This group could contain multiple discussions around the same topic. Discussion threads go really large depending on the members' active participation. In such a case, where we have discussion thread going really long, getting gist of the opinion of such discussion will be really difficult unless someone really go through the thread.

Proposal here is to build an algorithm to go over all the messages in the thread and analyse its sentiment value and calculate a weighted average of these. Since the discussion are posted by individuals the weight to be applied to the sentiment score should be based on their influence level in the group for example one of the member could be a Vice President of the company and hence his opinion must weigh higher than that of another member who could either be a junior or there could be a member who is not so high in his authority but could be a person all members who look

forward to due to his decisive and pragmatic messages over a period. Such people's messages should also be giver higher weightage.

1.3 Scope of work
The Scope of this work include
- Build a mathematical model to calculate the sentiment score of the discussion using the user's influence factor
- Build a new or identify and enhance (if required) an existing NLP based sentiment analysis library
- Build the background job to evaluate the overall sentiment score for all messages in each of the groups
- Provide a simple user interface to view the scores.

1.4 Plan & Status of work

Tasks	Status	Remarks
Evaluate & customize open source sentiment analysis libraries that is based on NLP	Done	Identified a RUBY Library and a PHP Library
Study the Enterprise Social Network backend DB structure	Done	Evaluated the existing DB Structure and also identified the Tables required to implement the solution
Build a Mathematical model to calculate the Sentiment Score, User Influence Level and overall Sentiment Score	Done	Described as part of this document
Build the backend job to evaluate the messages in each discussion groups and build a score	Done	The scores have been calculated whenever a COMMENT or a POST is committed to DB
Evaluate the Model output	Done	Manually validated the score in all cases
Build a simple UI to display the content to the group admin	Done	Completed and Screenshots are attached

Table 1- Status Report of the Dissertation work

Chapter 2 – Model

2.1 Sentiment Score

Sentiment Analysis is a way of extracting subjective information from a text, to determine "polarity" of the text. It is especially useful for identifying trends of public opinion in the social media, for the purpose of marketing. There are many Natural Language based libraries are available, which classify the text to either **Negative, Neutral or Positive.** In this model, we shall evaluate a library which is popular and uses the same technology that of Socialcast for Sentiment score calculation.

We define a function $\Xi(\textbf{text})$ to determine the Sentiment Scores of the text which is real number in the range of **[0, 1]**. This function shall use the NLP based Sentiment Analysis libraries to determine the value.

2.2 Influence Score

In an Enterprise, every individual come with an influence factor based many factors. In this work, we have considered the following features

2.2.1 Authority Level

Authority Level of the individual represent as $AL(U)$, which is a real number in the range of (0, 1) depending on U's hierarchical position in the company. This could be calculated using formula as follows:

$$AL(U) = 1 - \left(\frac{Level(U)}{\# \ of \ levels \ in \ Org} \right)$$

where $Level(U)$ is the level of U from head of the Organization. In general organization optimizes the number of levels to 5 or 6 depending on the size of the company. This is for better people management and organization Management.

2.2.2 Engagement Level

Engagement Level of the individual in a specific group is represented as P_U^{Engage}, which is user's engagement level in the system (i.e. how often one post messages or comments or like posts or comments in a post in the groups they are member.

$$P_U^{Engage} = P(post \cup comment \cup like)$$
$$= P(post) + P(comment) + P(like)$$

4

2.2.3 Enticing Level

Enticing Level of the individual represented as P_U^{Entice}, is the score given for how much of **U**'s actions bring in reactions from other users.

$$P_U^{Entice} = \frac{\sum(\text{\# of unique users reacted in Action}_i \text{ of } U)}{\text{\# of users}}$$

possible reactions supported against each of the actions in Socialcast are:

	Reactions	
	Comment-able	*Like*-able
POST	Yes	Yes
COMMENT	No	Yes

Table 2 - Reactions possible on each Objects (POST & COMMENT)

2.2.3 User Influence Score

The influence score of the individual, U, is calculated as

$$\Psi_U = AL(U) + P_U^{Engage} * P_U^{Entice}$$

It is clear from the above equation that the Authority Level of a user is more or less a constant for a period of time unless changed due to change in position in the organization (promotions). Variable part in this equation is the based on P_U^{Engage} and P_U^{Entice}. The more the engagement and more the reactions from others on his actions, increases one's overall score. We have used the product value to be sure many of the spam posts or noises in the system. It is learned from the system that spam POSTs are not reacted upon by other user majority of the time.

2.3 Influence-Weighted Sentiment Score
2.3.1 Influence-Weighted Sentiment Score of SocialObject

Now we define the *Influence-weighted Sentiment Score* of the Social Object as $\Xi^\Psi(SO)$ and is defined by

$$\Xi^\Psi(SO) = \left(\frac{(\Xi(\text{post}) * \Psi_U) + \sum_i(\Xi(\text{comment}_{u_i}) * \Psi_{u_i})}{(\text{\# of comments} + 1)} \right)$$

We are taking the mean of the influence-weighted Sentiment scores of all the POSTs and COMMENTs associated with the SocialObject.

2.3.2 Influence-Weighted Sentiment Score of Group

Since Groups are made of SocialObjects, further to calculate the Influence-weighted Sentiment score of a group, we take the mean of Influence-weighted Sentiment score of all its SocialObjects in the group

$$\Xi^{\Psi}(\textbf{GROUP}) = \left(\frac{\sum_i (\Xi^{\Psi}(\textbf{SO}_i))}{(\# \textit{ of SOs})} \right)$$

Chapter 3 - Approach

In this chapter, we shall evaluate the database/table schema of the existing Socialcast Application, which are relevant to for this work. We shall also review additional Tables required to store the data for this model. Secondly, we shall layout the flow diagram of the program(s)/Modules added for this work.

3.1 Database Tables Relevant for the Work from Socialcast

Following are the tables and relevant columns in Socialcast used for this work.

Table Name	Fields	Type
Comments	id	Integer
	user_id	Integer
	created_at	Date
	comment	String
	tenant_id	Integer
	commentable_id	Integer
Favorites	user_id	Integer
Groups	id	Integer
	name	String
	created_at	Date
	user_id	Integer
Group_memberships	user_id	Integer
	tenant_id	Integer
	group_id	Integer
Social_Objects	id	Integer
	title	String
	body	String
	user_id	Integer
	created_at	Date
	tenant_id	Integer
	group_id	Integer
Tenants	id	Integer
	name	String
Users	id	Integer
	first_name	String
	last_name	String
	preferences	String
	tenant_id	Integer

Table 3 - Tables and Columns from existing database

3.2 Additional Database Tables Required

Following table captures the details of additional tables required for this works.

Table Name	Fields	Type
Sentiment_Reference	id	Integer
	tenant_id	Integer
	reference_id	Integer
	reference_type	String
	score	Float
	weighted_score	Float
SC_Sentiment	id	Integer
	tenant_id	Integer
	range_min	Float
	range_max	Float
	name	String
	description	String
Social_Object_Sentiments	id	Integer
	tenant_id	Integer
	social_object_id	Integer

Table 4 - Structure of tables additional required

3.2 Modules and Logic Flow

Socialcast is built on **Ruby on Rails** technology and uses MySQL as the database. To simplify the work, we shall be developing the modules using Ruby and to access the Database we shall use the Rails features. We have identified following modules to help us solve the problem in hand.

3.2.1 Migration Module

Objective of this module to update the influence score of each of the user in the system, calculate the raw sentiment score and the weighted sentiment score of all POSTs and associated COMMENTS currently existing in the system. This module shall be run once in the beginning or can be executed whenever the NLP based Sentiment Analysis Libraries are updated/changes. Built this module using the Ruby on Rails framework. Since there were corresponding model classes in Ruby available in the system, updated these models to include sentiment related information.

Following flow diagrams depict the functioning of the migration module.

8

3.2.1.1 Calculate the Influence score of all existing users

Purpose of this module is to calculate influence score of all users in the system. We developed this using PHP programming language and executed in the environment to display the User Influence score. These scores are included in the migration script to populate the Users Table
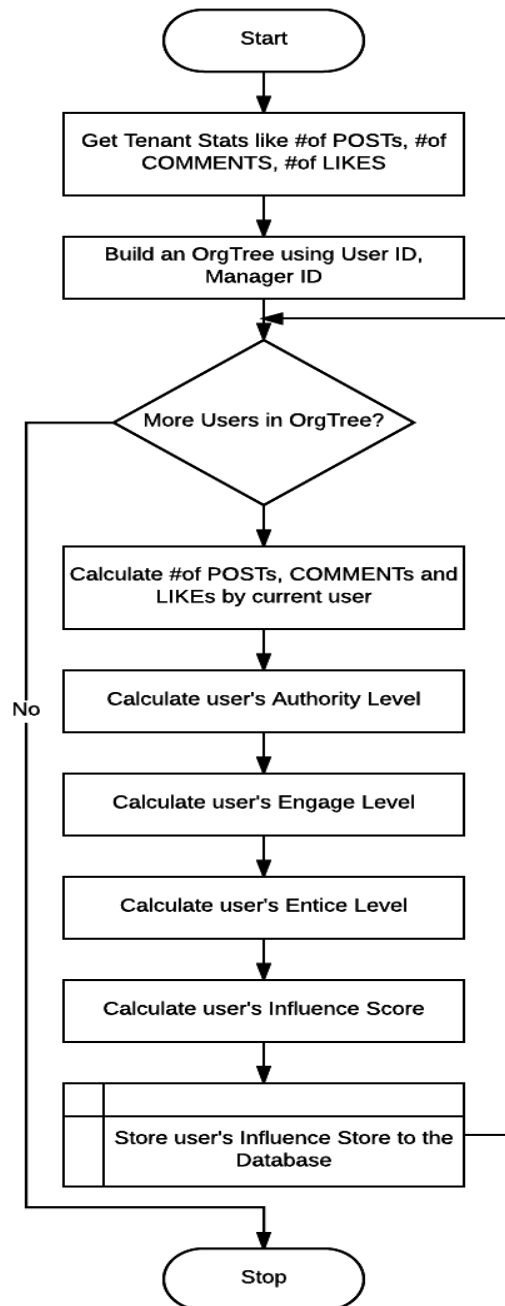

Figure 2 - Steps to calculate existing users' Influence Score

3.2.1.2 Calculate the Sentiment Score of Existing COMMENTs, POSTs, SocialObjects and Groups

This migration step is to calculate the Sentiment Score and Influence Weighted Sentiment Scores of all the existing POSTs & COMMENTs in the system and in turn to calculate their respective SocialObjects and Group's Influence Weighted Sentiment Score. This is developed in Ruby language and executed as one time job.
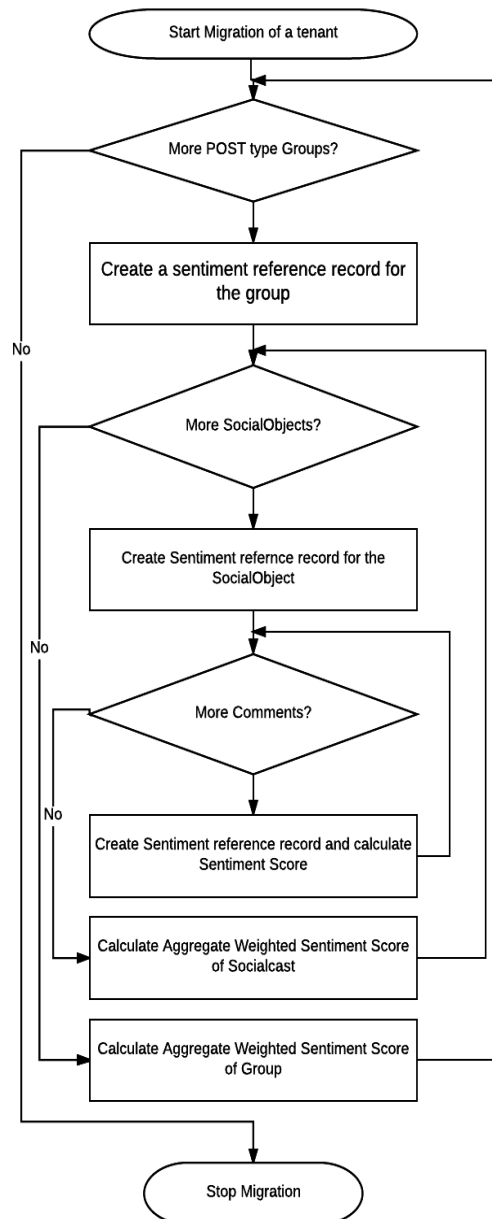

Figure 3 - Migration Steps to calculate the Sentiment Score and Influence-weighted Sentiment score of all COMMENTs, POSTs, SocialObjects and Groups

3.2.2 Hooks for Database Commit

Objective of this module is to write hooks using Ruby to update/insert recodes in to the additional tables with the raw sentiment score when user add a new POST or a COMMENT or edit existing POST or COMMENT.

3.2.2.1 Hook while committing a POST and SocialObject
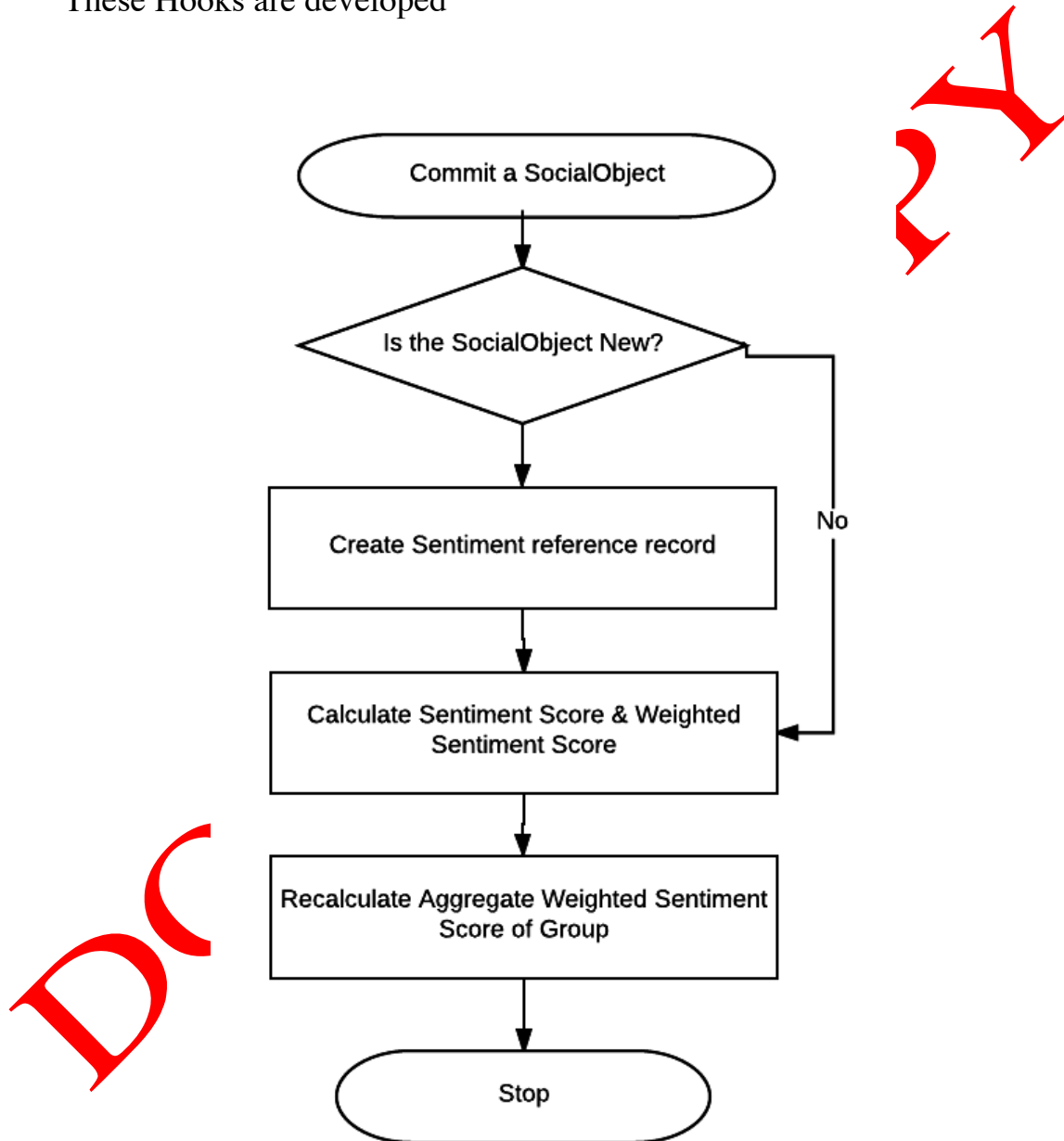
These Hooks are developed

Figure 4 - Steps to add/update sentiment score when a SocialObject is committed

3.2.2.2 Hook while committing a COMMENT
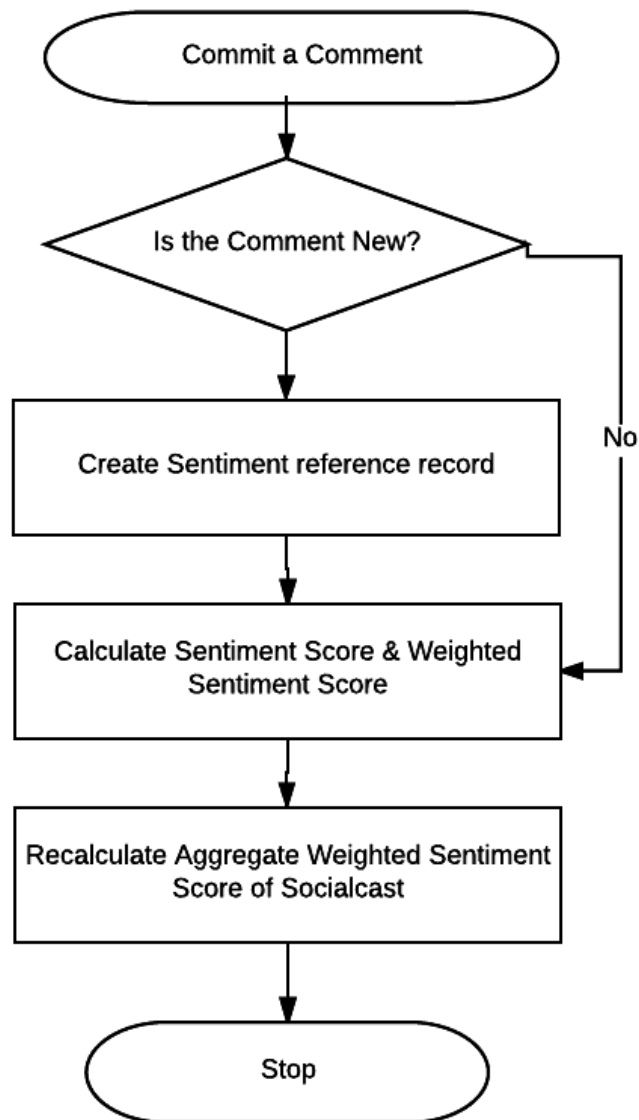
Figure 5 - Steps to add/update Sentiment Score when committing a COMMENT

3.3 User Interface to Display Final Sentiment Score

Objective is to provide User interface links along each of the POST, COMMENT and also along the group. The standard UI templates used in the system for POST type group has been updated to include the Weighted Score of the Group along the Group title, Added Weighted Sentiment Score of the SocialObject in the

main POST and have included the Sentiment Score of the COMMENT along the meta-information of the COMMENT.

3.3.1 User Interface Displaying Sentiment Score of a COMMENT
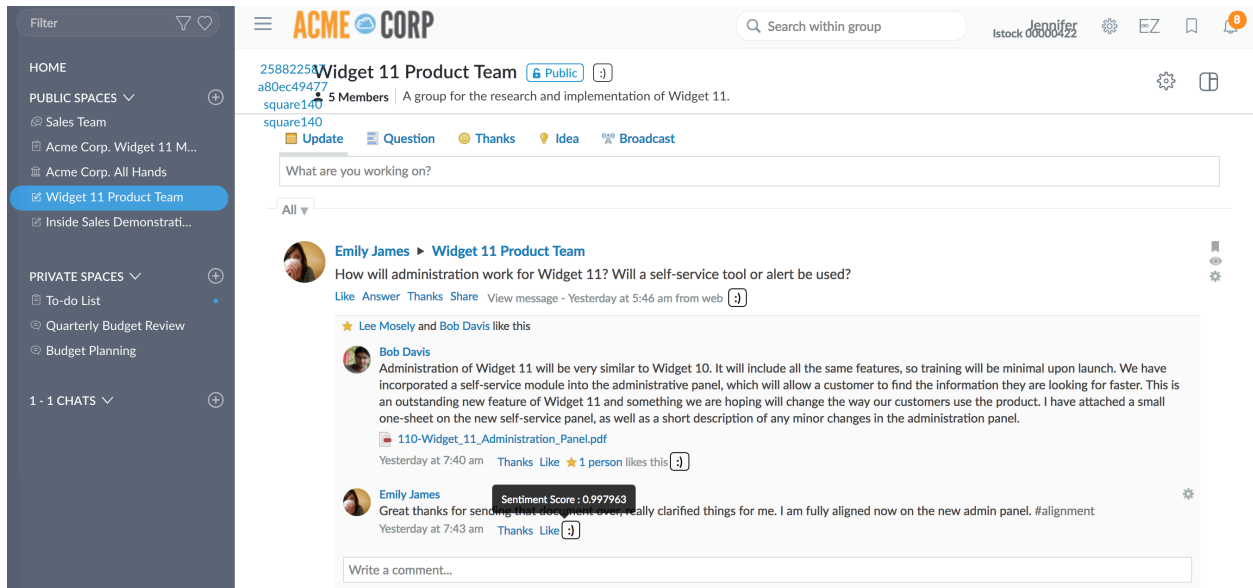

Figure 6 - Screenshot of the Application with Sentiment Score of a Comment

3.3.2 User Interface Displaying Weighted Sentiment Score of a SocialObject
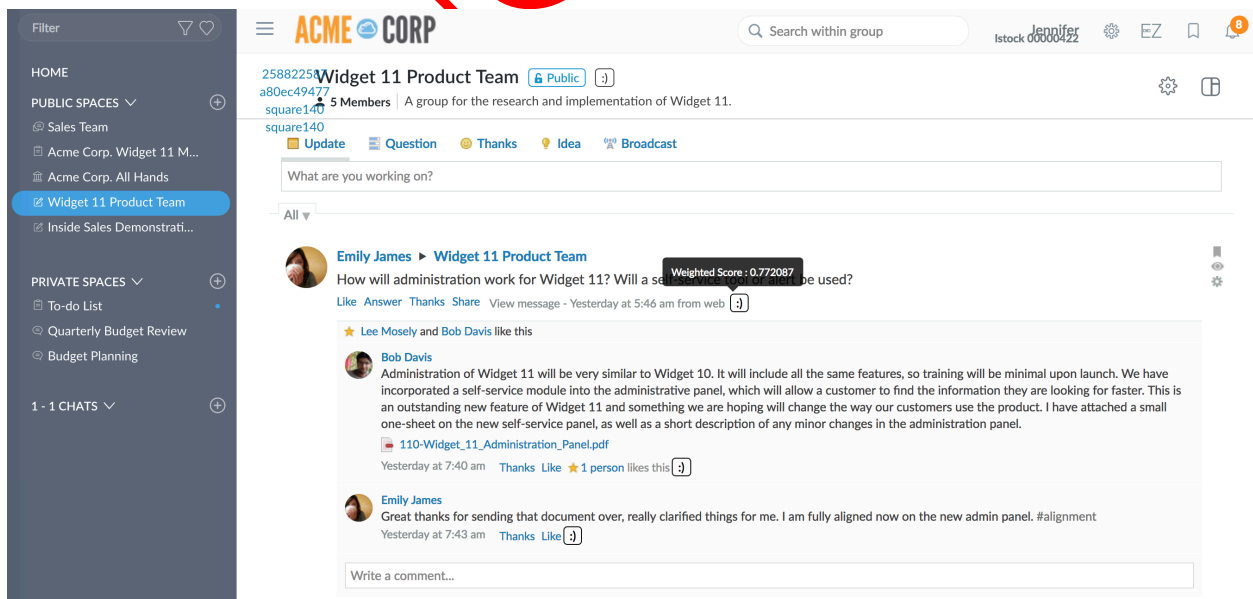

Figure 7 - Screenshot of Application with Weighted Sentiment Score of SocialObject

13

3.3.3 User Interface Displaying Weighted Sentiment Score of a Group
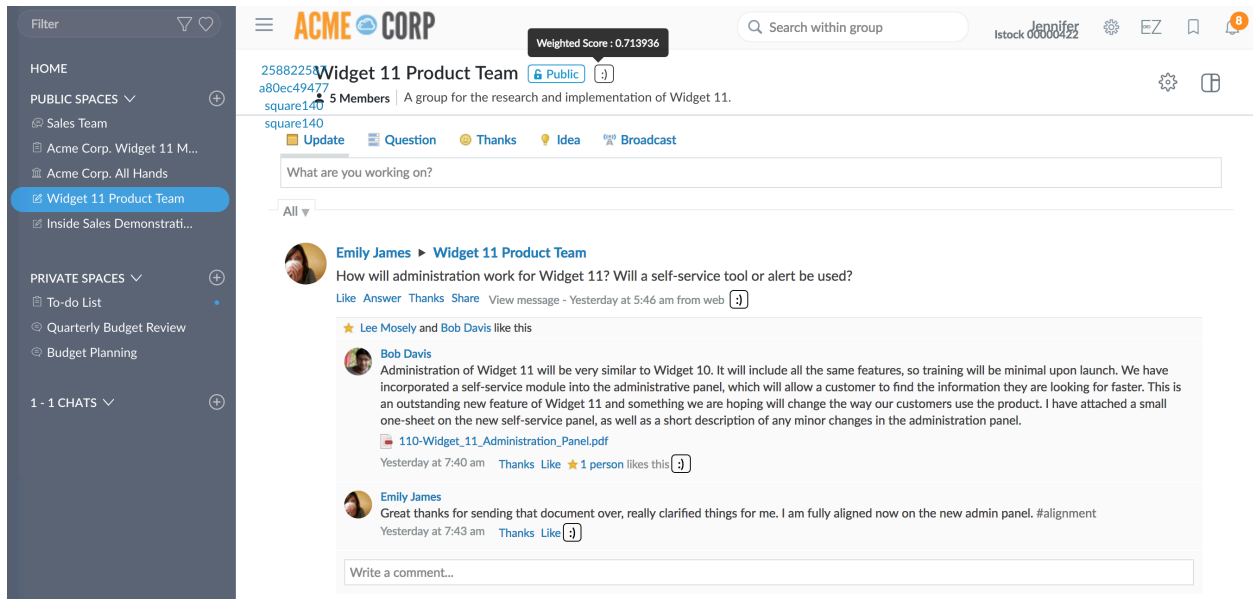
Figure 8 - Screenshot of Application with Weighted Sentiment Score of a Group

3.4 Sample Result of The Model

Following is the sample result that we gathered for a specific SocialObject and its components namely POSTs & COMMENTs

Type	Influence Score	Sentiment Score	Influence-Weighted Sentiment Score
SocialObject			0.590274
POST	0.753	0.999692	0.752768
COMMENT	0.927	0.775992	0.719345
COMMENT	0.753	0.998724	0.752039
COMMENT	0.921	0.998788	0.919884
COMMENT	0.753	0.982685	0.739962
COMMENT	0.415	0.122195	0.0507107
COMMENT	0.845	0.216681	0.183095

Table 5 - Result of a SocialObject and its associated components namely POSTs and COMMENTs to display effect of Influence Score in the model

The following graph gives the visual representation of the above Data
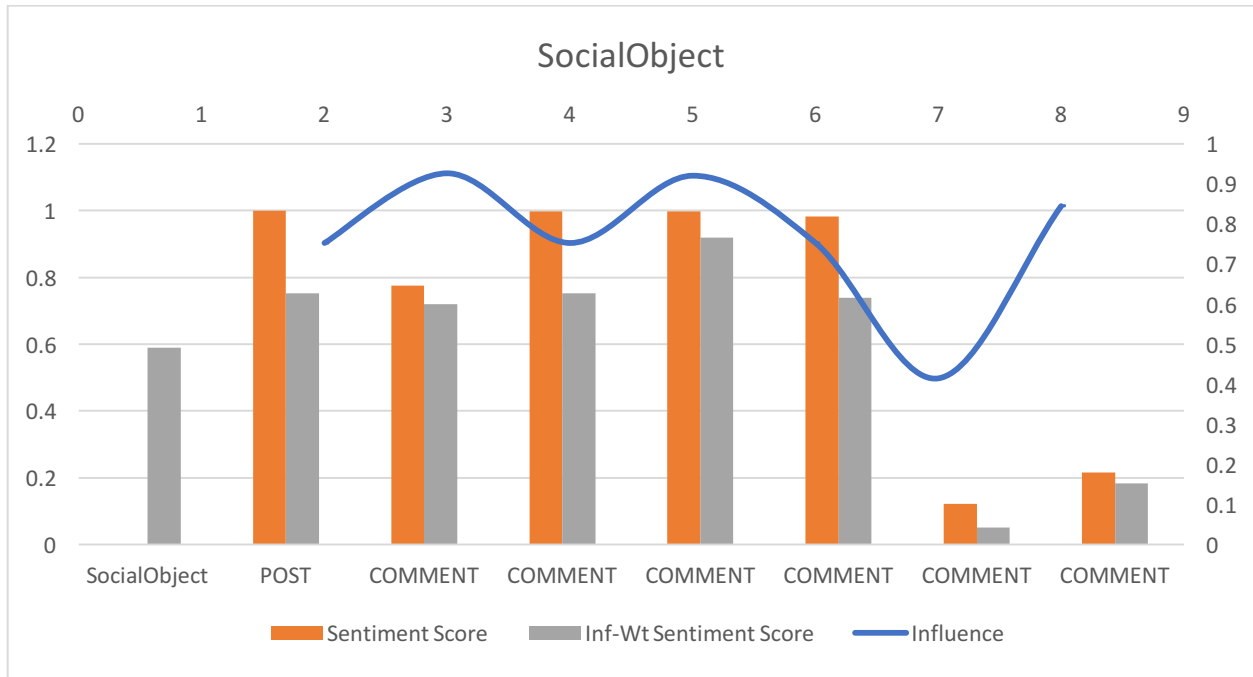
Figure 9 - Graphically representation of the result of a SocialObject and its components namely POSTs and COMMENTs to display effect of Influence Score in the model

It is evident from the data and the graph that the model handles the all cases (high sentiment score and low sentiment score) very well and the resultant value (Influence-Weighted Sentiment Score) stays within the range of values defined for the Sentiment Score in the table SC_Sentiment.

A SocialObject does not show sentiment score as it is calculated based on the Influence-Weighted Sentiment Scores.

Chapter 4 – Summary

Sentiment Analysis has been a feature that was in radar of the product team to include in the application. During the initial discussions, we realized that getting a simple sentiment score would not be the right path, and it would be game changer if the people's influence level is brought into the equation.

As part of this project we have modelled the user influence with less complexity given the timeline and the need of the quicker result. Bringing in the Authority Level of the user helped to seed the initial values and using the user's Engagement and Enticing capability score helped in improve their overall influence score. Authority Level score is more of a constant depending on position of the user at as it does not vary frequently. The Engagement and Enticing capability scores depends on user's activity on in the system. These values are at present calculated at Tenant level. We feel that in the future we could constraint this at group level for better accuracy.

In this model, we considered to have the Authority Level used as seed value of the Influence score. However, to gain more influence user should be highly engaged in the system and should have good amount of reactions from the other users as well. This to some extend eliminate spammers in the system as chances of other users reacting to such actions are minimal.

Library selected for calculating the Sentiment Score of POST and COMMENTS, uses the NLP and uses same technology that of Socialcast. The library has given good results on all the cases. The library has a good amount of bag of words in all categories namely Negative, Neutral and Positive. However, there is a scope for improving the dictionary to include more Enterprise Jargons and also sentences in the mentioned categories.

Overall it has been a great experience and learning working on the model for calculating the User Influence and working on the implementation.

16

Chapter 5 - Conclusions and Recommendations

In this project, we defined a framework to model the sentiment of a Group, its POSTs and COMMENTS. In an enterprise, which is organized hierarchically, every user has their own authority level, in this model we have used a tree data structure to calculate one's level in the organization and thus calculating the Authority Level. We have also used, in the context of Socialcast™, a framework to calculate the user's engagement level and entice factor which along with their Authority Level are used to calculate the overall Influence Score. This influence score is used measure the weighted sentiment of each POST and COMMENT by the user.

This model will surely be applicable in ticketing systems especially involving customer interactions, where the original issue posted could be considered as a POST and the comments as COMMENTS. In case if there are any specific grouping of issues available we could also measure the overall weighted sentiment score of such groups as well

Chapter 6 - Directions for future work

In this project, we have measured user influence score at the community level. We feel that measuring user influence level at individual group could be more precise as there is no hard and fast rule the users would be part of all groups and would be active in the groups. We like to pursue to alter the model to measure influence at more granular level

It would be more valuable to have the influence score calculated dynamically at the time of their submission of POST or COMMENT, which means that my previous post in last one hour would also be accounted for new submission.

It would be more useful to use temporal data of weighted sentimental score to visually represent how the discussion have moved over a period of time. This would require more database tables and corresponding associations with other existing tables.

Like to apply this model in the ticketing systems which follows similar Social Object format to see how customer involved discussions are.

Bibliography and References

BOOKS

1. Pang-Ning, Michael and Vipin, <u>Introduction to Data Mining</u>, Delhi: Pearson, 2016
2. Jiawei, Micheline and Jian, <u>Data Mining Concepts and Techniques</u>, Waltham, MA, USA: Morgan Kaufmann Publishers, 2012
3. Christopher, Prabhakar and Hinrich, <u>Introduction to Information Retrieval</u>: Delhi, Cambridge University Press, 2008

SCHOLARLY JOURNAL ARTICLES

4. Kushal Dave; Rushi Bhatt; and Vasudeva Varma, "<u>Identifying Influencers in Social Networks</u>", Association for the Advancement of Artificial Intelligence (www.aaai.org): 1-9

CONFERENCE PROCEEDINGS

5. Peter Hui and Michelle Gregory, "Quantifying Sentiment and Influence in Blogspaces", 1st Workshop on Social Media Analytics (SOMA '10) July 25, 2010; Washington, DC, Washington, DC: ACM; 2010
6. Kempe, D.; Kleinberg, J.; and Tardos, "Maximizing the spread of influence through a social network." In KDD'03, ACM; 2003.
7. Aral, S.; Muchnik, L.; and Sundararajan, A., "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks." Proceedings of the National Academy of Sciences 106(51); 2009

Checklist for the items in the report
Checklist of items for the Final Dissertation Report
This checklist is to be attached as the last page of the report.
This checklist is to be duly completed, verified and signed by the student.

1.	**Is the final report neatly formatted with all the elements required for a technical Report?**	Yes / No
2.	Is the Cover page in proper format as given in Annexure A?	Yes / No
3.	Is the Title page (Inner cover page) in proper format?	Yes / No
4.	(a) Is the Certificate from the Supervisor in proper format?	Yes / No
	(b) Has it been signed by the Supervisor?	Yes / No
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	Yes / No Yes / No
6.	Is the title of your report appropriate? **The title should be adequately descriptive, precise and must reflect scope of the actual work done.** Uncommon abbreviations / Acronyms should not be used in the title	Yes / No
7.	Have you included the List of abbreviations / Acronyms?	Yes / No
8.	Does the Report contain a summary of the literature survey?	Yes / No
9.	Does the Table of Contents include page numbers?	Yes / No
	(i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1)	Yes / No
	(ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures)	Yes / No
	(iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables)	Yes / No
	(iv). Are the Captions for the Figures and Tables proper?	Yes / No
	(v). Are the Appendices numbered properly? Are their titles appropriate	Yes / No
10.	Is the conclusion of the Report based on discussion of the work?	Yes / No
11.	Are References or Bibliography given at the end of the Report?	Yes / No
	Have the References been cited properly inside the text of the Report?	Yes / No
	Are all the references cited in the body of the report	Yes / No
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a Power Point Presentation, or a user manual. Source code of software need not be included in the report.	Yes / No

Declaration by Student:
I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

Place: _____

Date: _____ _____

_____
Signature of the Student
Name:_____
ID No.:_____