Surfacing Personas from Enterprise Social Media
to Enhance Engagement Visibility

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in
the Graduate School of The Ohio State University

By

Ramiya Venkatachalam

Graduate Program in Computer Science and Engineering

The Ohio State University

2013

Master's Examination Committee:

Dr. Jay Ramanathan, Thesis Advisor

Dr. Rajiv Ramnath
Copyright by
Ramiya Venkatachalam
2013
Abstract

Enterprise social media potentially plays a pivotal role in capturing useful insights about employee behavior. This thesis explores how, with the availability of social media analytics tools and methods, this potential is realized in effective decision making within the business processes of an enterprise. Specifically it proposes a generic analytics framework that provides visibility into employee engagement which is a measure of organizational effectiveness in the enterprise through data extracted from internal enterprise social media and other traditional data sources. This extracted data is called an "Engagement Persona". Here the standard employee profiles (factual information from human resource systems) are enhanced with behavioral insights derived from social media and captured interesting relationships that serve as rich information assets. The thesis shows that the Engagement Persona provides visibility that could ultimately reinforce engagement. Sample Engagement Personas are shown by implementing a working model of the proposed framework within a large insurance company that uses Yammer - a micro blogging system to update colleagues on company events, ask work-related questions, and broadcast problem situations. Further, the implementation showed large classes of messages revealing user intent. These were aggregated into a form
reflecting the user's underlying Engagement Persona. This holistic view of each employee was shown to be valuable for business roles like management, communications and the Human Resources. The value is in providing visibility into aspects of employee engagement thus leading to smarter management decision-making. Feedback from these roles shows that they are satisfied with the improved visibility and ability to respond on a timelier basis to engagement changes. Thus the thesis shows the framework helps in boosting engagement activities and also fosters better knowledge sharing across the different business units of the enterprise.
Dedication

*This thesis is dedicated to my parents and my loving sister, Gowree.*
Acknowledgments

I would like to take this opportunity to thank all the people who have been instrumental in many ways in making this research a worthwhile experience.

I would like to convey my deepest gratitude to Dr. Jay Ramanathan for being my advisor and helping me think out of the box. She has helped me become a better professional. Her active nature and zest for life has inspired me in many ways and will continue to inspire me. I would also like to thank Dr. Rajiv Ramnath for guiding me in making important decisions about my career and also in my research. I am indebted to Dr. Jay & Dr. Rajiv for considering me worthy of the opportunity to work on such a cool project like social media analytics. They both have helped me chase my dream. I would like to thank all my mentors at Nationwide. I would like to thank Bruce Craig for being the coolest mentor I could have ever asked for. He has literally been a godfather and has guided me beautifully. He has motivated me to do better every single time and has been a true support in making this project a successful one. I would also like to thank Tara Paider, who has been my US mom all the way. I want to thank her for hiring me onto her team for working on such a cool project. I also want to thank Alex Hay for being an
indispensable support and a magician during the last leg of my thesis project. I want to thank Chris Plescia, Jeff Schumann & Niki Burke for providing the Yammer data without which the validation of this thesis would have been impossible. I also would like to thank Dr. Arnab Nandi & Dr. Srinivasan Parthasarathy for their constant support and for helping me learn more about the data space.

I do not have enough words to thank my family. I want to thank my dad - Late Mr K.S Venkatachalam - as I know he always wanted the best for me. He always wanted me to get a Masters degree in Computer Science. I am glad that I have reached that milestone today. I want to thank my mother for being a truly inspiring single mom. I also want to thank my loving sister, Gowree for motivating me throughout my stint here. I am truly blessed to have such a wonderful family.

Special thanks to all my friends at OSU, for making this an enjoyable experience. I would like to thank Chaitanya Shivade for being a great mentor. I want to thank Alhad, Mitesh & Karthik for guiding me through this journey and helping me keep my cool. I want to thank my roommates, Arts & Shweta, who have been there through thick and thin. I want to thank Rohan, Chatty, Prasanth, Dharma, Niranjan, Ziqi, Farhang, Vineet, Manas for making me laugh during tense moments. I would also like to thank Aishu, for baking my birthday cake and I wish her all the best as she takes on my role at Nationwide.
Vita

2003 to 2005 ............................................S.P. College, Pune, Maharashtra, India

2005 to 2009 ............................................B.E, Computer Science, University of Pune, India

July 2009 to July 2011 .........................Senior Software Engineer, Persistent Systems Ltd., Pune, India

April 2012 to May 2013 ......................Analytics Intern, Data & Analytics, Nationwide Insurance, OH, USA

September 2012 to present ......................Graduate Research Associate, Department of Computer Science & Engineering, The Ohio State University, OH, USA

Fields of Study

Major Field: Computer Science and Engineering
# Table of Contents

Abstract ................................................................................................................................. ii

Acknowledgements .............................................................................................................. v

Vita ........................................................................................................................................ vii

Fields of Study ...................................................................................................................... vii

Table of Contents ................................................................................................................ viii

List of Tables ....................................................................................................................... xi

List of Figures ....................................................................................................................... xii

Chapter 1: Introduction ....................................................................................................... 1

1.1 Terminology ..................................................................................................................... 3

1.1.1 Named Entity .............................................................................................................. 3

1.1.1.1 Generic Named Entities ....................................................................................... 3

1.1.2 Rule-based Text Analytics ......................................................................................... 4

1.1.3 Triplestore .................................................................................................................. 4

1.1.4 Engagement Metric ................................................................................................... 5

1.1.5 Persona ....................................................................................................................... 5

1.1.5.1 Engagement Persona ............................................................................................ 6

1.1.5.2 Engagement Visibility Level ................................................................................ 6

viii
1.2 Contributions ........................................................................................................ 7

1.3 Organization .......................................................................................................... 8

Chapter 2: Related Work .............................................................................................. 9

Chapter 3: Problems & Challenges - An Insurance Company Use Case .................. 16

3.1 Current Engagement Visibility Approach ......................................................... 17

3.2 Problems ............................................................................................................... 20

3.3 Motivation .............................................................................................................. 21

Chapter 4: The Engagement Visibility Framework (EVF) ......................................... 22

4.1 Conceptual View .................................................................................................. 23

4.1.1 Entity Analytics Layer .................................................................................. 23

4.1.2 Summarization Layer ..................................................................................... 26

4.1.3 Business Process Layer .................................................................................. 27

4.2 The Engagement Persona .................................................................................... 30

4.2.1 The Engagement Visibility Level ..................................................................... 33

4.2.2 The Engagement Persona Data Model ........................................................... 36

4.3 Value Proposition ................................................................................................ 37

Chapter 5: Implementation of EVF - Insurance Industry Use Case ....................... 39

5.1 Datasets - Insurance Company .............................................................................. 40

5.1.1 Yammer Data .................................................................................................. 40

ix
List of Tables

Table 1: Classification of the EVF as a Solution Pattern [25] .......................... 14
Table 2: A sample Engagement Persona Data Model ........................................ 36
Table 3: Important attributes of the Yammer Data Model (messages.csv) .............. 41
Table 4: Important attributes of the Organization's Call Directory Employee Profile 41
Table 5: The Engagement Persona Data Model for the Insurance Company use case 48
Table 6: Dictionary entries for identifying success stories .................................... 82
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gallup Survey questions mapped to the Engagement hierarchy [30]</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>Graph showing the Enterprise level Engagement Score (2008-2012)</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Conceptual View of the EVF</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>Entity Analytics - Classification Hierarchy</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>Interaction Model for EVF</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>Conceptual View of The Engagement Persona</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Semantic social graph showing Engagement Personas of multiple employees</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>Semantic Social Graph with Engagement Visibility Levels</td>
<td>33</td>
</tr>
<tr>
<td>9</td>
<td>Semantic Social Graph at Engagement Visibility Level 1</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>Semantic Social Graph at Engagement Visibility Level 2</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>The Value of the Engagement Persona</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>Technologies used for implementing EVF</td>
<td>42</td>
</tr>
<tr>
<td>13</td>
<td>High Level Object View of the EVF Implementation</td>
<td>43</td>
</tr>
<tr>
<td>14</td>
<td>Real Time (Production) Scenario of the EVF Implementation</td>
<td>44</td>
</tr>
<tr>
<td>15</td>
<td>Interaction Diagram of the EVF Implementation</td>
<td>46</td>
</tr>
<tr>
<td>16</td>
<td>Engagement Persona Data Model populated for a sample Yammer message</td>
<td>49</td>
</tr>
<tr>
<td>17</td>
<td>IBM InfoSphere Streams Application Graph - Entity Analytics Layer</td>
<td>51</td>
</tr>
<tr>
<td>18</td>
<td>The iterative business process for engagement visibility</td>
<td>53</td>
</tr>
<tr>
<td>19</td>
<td>Yammer activity for the period 2008-2012 across the insurance enterprise</td>
<td>56</td>
</tr>
<tr>
<td>20</td>
<td>Classes of Yammer messages identified by the Entity Analytics Layer</td>
<td>58</td>
</tr>
<tr>
<td>21</td>
<td>Histogram showing the count of Yams in each identified class</td>
<td>59</td>
</tr>
</tbody>
</table>
Figure 22: #Yams in 2012 per topic related to the IT Service Desk (ITSD) .......................... 60
Figure 23: #Yams mentioning each phone number .......................... 61
Figure 24: #Yams mentioning each city in the United States .......................... 62
Figure 25: Semantic Social Graph with "Office" = "Corporate" .......................... 63
Figure 26: Semantic social graph showing cross collaboration at the "Office" level...... 64
Figure 27: Semantic social graph showing cross collaboration at the "Dept." level. ..... 65
Figure 28: Semantic social graph - Sparse in nature .......................... 66
Figure 29: Semantic social graph - Dense in nature at EVL 3 .......................... 67
Figure 30: Semantic social graph aggregated on "Office" attribute .......................... 68
Figure 31: User Survey Questions - Part 1 .......................... 69
Figure 32: User Survey Questions - Part 2 .......................... 70
Figure 33: Results of the User Survey - Part 1 .......................... 71
Figure 34: Results of the User Survey - Part 2 .......................... 72
Figure 35: Results of the User Survey - Part 3 .......................... 73
Figure 36: Key Components or building blocks of IBM SystemT [26] Constructs ....... 79
Chapter 1: Introduction

Employee engagement is a well-known business management concept that is targeted towards a collaborative and a healthy workplace [2]. According to Scarlett Surveys [3], "Employee Engagement is a measurable degree of an employee's positive or negative emotional attachment to their job, colleagues and organization that profoundly influences their willingness to learn and perform at work". Thus engagement is distinctively different from employee satisfaction, motivation and organizational culture [2]. With numerous companies competing for the same employee, retention is a crucial issue. Employee engagement was described in the academic literature by Schmidt et al. (1993). A modernized version of job satisfaction, Schmidt et al.'s influential definition of engagement was "an employee's involvement with, commitment to, and satisfaction with work. Employee engagement is a part of employee retention" [2]. It is thus important to rethink the process of engagement and leverage it for making the workplace more productive.

Engagement opportunities have greatly evolved over time. Conversations over coffee amongst employees have now moved on to enterprise social media. The era of social
business has thus made its way into the enterprise. Consequently, many of the Fortune 500 companies have adopted such enterprise collaboration platforms for engagement, in particular to share knowledge and inculcate a feeling of being a close-knit community. And hence, there has been a flurry of activity in the enterprise social media networks as it gives a voice to the employees for asking questions and sharing ideas. **IBM Connections** [4] and **Yammer** [1] are some of the prominent ones. Social media analytics are also available today to analyze raw and rich social media data. It has thus become feasible to use social media as a rich source of data to track employee behavior and provide engagement visibility. The feeds can be voluminous and hence mining the feeds for meaningful information and summarizing them close to real time becomes critical.

The research work thus far in this space has not focused much on the enterprise specifically and, especially, how it meets some of its engagement goals by summarizing the social media streams. The overall objective of this research is thus to contribute to the emerging discipline of behavioral analytics based on social media used within an actual enterprise. One cannot understand a big enterprise community upfront and this understanding evolves over time. Thus a framework approach rather than a ‘big bang’ answer is needed.

Given the increasing need for near real-time analytics and lack of a generic framework for social media analysis to enhance engagement visibility specifically in an enterprise,
the main research contribution of this thesis is a portable analytics framework for mining enterprise social media for engagement visibility.

1.1 Terminology

This section provides the definitions of concepts that are introduced as a part of the proposed EVF and also of the generic or standard terms referred to in this thesis.

1.1.1 Named Entity

In data mining, a Named Entity [5] is a phrase that clearly identifies words or patterns from the input text as a specific type. For example, phone numbers and addresses are named entities. Phone numbers in the US are of the form XXX-XXX-XXXX and hence they can be named as “USPhoneNumber”. Based on the domain, the named entities are classified as generic and custom named entities.

1.1.1.1 Generic Named Entities

Generic Named Entities are well-recognized entities which are common across domains and which can be extracted using the standard heuristics. For example, phone numbers and addresses are generic named entities as they are not domain specific.
1.1.1.2 Custom Named Entities

*Custom Named Entities* are entities that are domain specific and require custom rules for extraction and summarization. Custom Named Entities can be explicit or derived. For example, *helpdesk service* is a derived custom named entity as it requires special domain specific rules. Such special rules would relate "Lotus Notes" and "Login failure" (both custom entities) to the derived custom entity *helpdesk service*.

1.1.2 Rule-based Text Analytics

In data mining, the process of extracting named entities and relationships between them for analyzing unstructured text for decision making is called as text analytics. Rule-based text analytics relies on gigantic dictionaries and sophisticated linguistic rules. Analysts can improve the out-of-the-box translation quality by adding their own terminology into the translation process. They create user-defined dictionaries, which augment the system's default dictionaries [6]. For example, if all the cities in the US have to be extracted as an entity, 'USCity' from the input text, then the analyst must build a custom dictionary containing the names of all the U.S. cities and then apply pattern-matching rules to extract the U.S. cities from the input text using dictionary matches and name them as 'USCity'.

1.1.3 Triplestore

A triplestore is a purpose-built database for the storage and retrieval of triples, [31] a triple being a data entity composed of subject-predicate-object, like "Bob is 35" or "Bob
knows Fred”. Triples are useful in demonstrating and visualizing simple relationships in graphical forms. In this thesis, a triplestore is the data model proposed for representing the named entities and relationships extracted from the internal social media data. Predicates serve as relationship names and named entity instances serve the role of both subjects and objects depending on the context. This data model is explained in detail in Section 4.2.2.

1.1.4 Engagement Metric

In the context of an enterprise, an ‘Engagement Metric’ is any measure that contributes to an understanding of organizational effectiveness. For example, in the context of enterprise social media, the number of posts can be an engagement metric that gives an idea about how active and, thus, how engaged an employee is. One might also measure frequency of user posting, diversity of group membership a user displays and other quantitative measures. In the case of an important campaign involving many employees, like a blood donation campaign in a company, the number of posts using a hashtag related to the campaign would be an important engagement metric to be tracked.

1.1.5 Persona

The dictionary meaning of Persona, is the aspect of someone's character that is presented to or perceived by others. Persona in the context of this thesis is represented by a set of named entities in aggregate form that surface behavioral insights about a person or a group of people along with relationships between those entities.
1.1.5.1 Engagement Persona

In this thesis, an Engagement Persona is an instance of a Persona in the context of the digital discourse of an enterprise related to engagement. This Persona is derived by aggregating and referencing all the named entities related to specific employees or groups from the enterprise social media data and thus augmenting the existing employee factual records. It is visualized in the form of graphs called as Semantic Social Graphs where the nodes are named entities and the edges between the nodes indicate relationships between the different named entities. At an implementation level, the triplestore structure contains this information as described above.

1.1.5.2 Engagement Visibility Level

The Engagement Persona data for an enterprise is a complex Semantic Social Graph consisting of many instances of named entities and relationships between them. A means to focus on particular named entities is needed. In this thesis, an Engagement Visibility Level (EVL) is defined. This EVL specifies the depth of the enterprise social graph that will help manage the density to convey the Engagement Persona effectively.
1.2 Contributions

This thesis addresses the problem of lack of engagement visibility by mining the enterprise social media for named entities to enhance the employee profiles that exist in the organization's directory database. The key contributions of this thesis are as follows:

- A generic and portable analytics framework for enterprise to provide engagement visibility from internal social media by extracting Engagement Personas. This is called - "The Engagement Visibility Framework" or "EVF".
- An initial identification of the facets in a digital Engagement Persona.
- An EVF data model that supports faceted search and graphical exploration of the data for making the persona more explicit with a future personal visualization interface.
- Engagement Personas that generate rich social graphs which can be managed by specifying Engagement Visibility Levels (Density of the social graph).
- Key elements of the proposed framework for an insurance company that uses Yammer (enterprise microblogging service) are implemented and discussed.
1.3 Organization

The rest of the thesis is organized as follows: Chapter 2 summarizes the related work. Chapter 3 begins with the AS-IS scenario of a large insurance company using enterprise social media like [1]. The challenges and the manual techniques used to track engagement are discussed in detail. This chapter also introduces the need for an analytics framework like EVF. Chapter 4 discusses in detail the various components of EVF and provides a conceptual view of the analytics framework. Chapter 5 presents an implementation of EVF for the insurance company using a rich Big Data technology stack by IBM InfoSphere Streams v3.0 [7] and an open source graph visualization software by Gephi 0.8.2-beta [8]. This chapter also discusses feedback, results & the insights obtained. Finally, Chapter 6 covers the conclusion and future work.
Chapter 2: Related Work

Extensive literature survey shows that there exists a significant body of research work in the social media analytics domain in the context of an enterprise over the past few years. The introduction of a social networking site inside of a large enterprise enables a new method of communication between colleagues, encouraging both personal and professional sharing inside the protected walls of a company intranet. Professional people use internal social networking to build stronger bonds with their weak ties and to reach out to employees they do not know. Their motivations in doing this include connecting on a personal level with coworkers, advancing their career with the company, and campaigning for their projects [9].

The first study in this area was done by Zhao and Rosson [10]. They interviewed 11 Twitter users in a large IT company. Using a conceptual framework of possible beneficial consequences of informal communications, they analyzed the potential impact of micro-blogging at work and provided valuable insight into why and how people micro-blog and how to support micro-blogging in an organizational context.
Preliminary investigation into an internal corporate blogging community called BlogCentral by conducting semi-structured interviews with fourteen active bloggers to investigate the role of blogging and its effects on work processes led to similar findings: how the BlogCentral facilitates access to tacit knowledge and resources vetted by experts, and, most importantly, contributes to the emergence of collaboration across a broad range of communities within the enterprise [11].

Following these studies, there have been many works around similar theories [12], [13] and [14]. In [13], a case study about the early adoption and use of micro-blogging in a Fortune 500 company revealed that users vary in their posting activities, reading behaviors, and perceived benefits. The analysis also identified barriers to adoption, such as the noise-to-value ratio paradox. The findings can help both practitioners and scholars build an initial understanding of how knowledge workers are likely to use micro-blogging in the enterprise.

The work of Michael Brzozowski [15] in the year 2009, describes WaterCooler, a tool that aggregates shared internal social media and cross-references it with an organization’s directory. They also deployed the WaterCooler in a large global enterprise and present the results of a preliminary user study which reveal that - WaterCooler changed users’ perceptions of their workplace, made them feel more connected to each other and the company, and redistributed users’ attention outside their own business groups. In the work of Brzozowski et al [16] in the same year, the authors present the results of a year-
long empirical study of internal social media participation at a large technology company to provide deeper insights about the community. They find that the feedback in the form of posted comments is highly correlated with a user’s subsequent participation. Recent manager and coworker activity were found to relate to users initiating or resuming participation in social media. These findings extend, to an aggregate level, the results from prior interviews about blogging at the company and offer design and policy implications for organizations seeking to encourage social media adoption.

Recently in 2012, a different take on the enterprise social media was described in the works of Muller [17] that focuses on statistical patterns of contributing vs. “lurking.” They examine patterns of participation by employees who are members of multiple online communities in an enterprise community’s service. The majority of contributors (in one or more communities) were also lurkers (in one or more other communities). These results argue against hypotheses derived from common theories of participation and lurking. They propose that contributing and lurking are partially dependent on a trait (a person’s overall engagement), modified by the individual’s disposition toward a particular topic, work task or social group. Some of Muller's ideas are adapted in this thesis and is very relevant to the work done in this thesis.

In the work of Lee et al [18] in early 2012, the authors show that text mining is able to surface employee’s frequency of communication and topics of conversation through posting activities using Yammer. The work done in this thesis closely aligns with this
work and also with some of the recent case studies on Yammer that are specific to a
different kind of enterprise [19] [20]. They investigate emerging communicative work
practices on Yammer within Deloitte & Cap Gemini. The authors perform a genre
analysis of actual communication data captured on the Yammer platform to conclude that
Yammer in the case company has become 1) an information-sharing channel, 2) a space
for crowdsourcing ideas, 3) a place for finding expertise and solving problems and most
importantly 4) a conversation medium for context and relationship building.

The demo by IBM Research recently presented at the SIGMOD 2012 [21] provides a
broader perspective and use-case on the whole stream computing paradigm for social
media. They have built a system that provides a 360 degree view of users on social media
near real time for lead generation and brand management.

This thesis mainly focuses on how real time insights from internal social media can
benefit the enterprise in promoting collaboration and engagement. The portable analytics
framework - EVF is aimed at making employee profiles richer by augmenting them with
insights from user's social media activity to provide engagement visibility. Further, the
framework is validated by implementing a working model for a large Fortune 100
insurance company which is known to use Yammer proactively and is one of the top
companies using the platform.
In the work of Plinio Thomaz Aquino Junior & Lucia Vilela Leite Filgueiras [22], the authors discuss the Personas concept applied to a user centered project, as the best way of employing user information compared to other models that serve this same purpose. Personas are fictitious user representations created in order to embody behaviors and motivations that a group of real users might express, representing them during the project development process. This article describes the Persona as being an effective tool to the users’ descriptive model [22]. This thesis uses the Personas concept in a different context, namely enterprise social media. It considers Persona as a set of facets of user behavior derived from social media representing a person or a group of persons which potentially enhances engagement visibility in the enterprise.

Enterprises organize their complex business and IT systems into reference architectures to help communicate and manage their environments. These reference architectures are kind of an architectural pattern. An architectural pattern is a concept that solves and delineates some essential cohesive elements of software architecture. Countless different architectures may implement the same pattern and thereby share the related characteristics. Furthermore, patterns are often defined as something "strictly described and commonly available" [25]. Mapping the EVF into this scheme, EVF is within the business intelligence and analytics reference architecture [28] with extension in the data reference architectures to include enterprise social media data. Table 1 gives the classification hierarchy of the framework proposed in this thesis.
<table>
<thead>
<tr>
<th>Sub-Domain Area</th>
<th>Architecture Pattern Name</th>
<th>Design Patterns</th>
<th>Solution Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Intelligence</td>
<td>Operational Reporting Analytical Reporting</td>
<td>Analytical Reporting Data Access Analytical Dashboard Data Access Operational Dashboard Data Access Data Mining</td>
<td>Real-time dashboards</td>
</tr>
</tbody>
</table>

Table 1: Classification of the EVF as a Solution Pattern [25]

In the work of Guy et al [23] at IBM Research, they studied personalized item recommendation within an enterprise social media application suite wherein recommendations are based on two of the core elements of social media - people and tags. Relationship information among people, tags, and items, is collected and aggregated across different sources within the enterprise. Based on these aggregated relationships, the system recommends items related to people and tags that are related to the user. This thesis proposes to surface such hashtags, named entities and the relationships between them from enterprise social media posts which can be used for aggregations downstream. A recommendation system based on these newly available named entities and relationships would be a potential future research direction.
In the work of Qureshi S. et al [24], a framework is proposed for activation of knowledge through electronic collaboration for the creation of more engaged work environments. The EVF proposed in this thesis can ultimately be a knowledge activation framework in itself as it surfaces named entities hidden in the posts which can be potential knowledge sources for various teams within the enterprise thus promoting collaboration.
Chapter 3: Problems & Challenges - An Insurance Company Use Case

This chapter presents a comprehensive case study of an insurance enterprise that uses internal social media like Yammer and actively promotes engagement at the enterprise level. The large insurance company under consideration is a Fortune 100 Company and has over 33,000 employees. With the emerging use of social media for engagement purposes, the same insurance enterprise has been listed in the top 10 companies that proactively use Yammer. The company goal is to make employees more productive by boosting collaboration and engagement activities.

Yammer as an enterprise wide micro blogging application targets a relatively smaller user base than Twitter/Facebook. Professionals use Yammer to update colleagues on company events, ask work-related questions, and broadcast problems. Serious questions and issues are thus raised and discussed along with the more casual and informal comments. Yammer posts usually have some syntactic structure (author, date, tags) but the posts themselves are free-form and unstructured text. Yammer posts unlike tweets do not have a character limit. Since it is used in the context of an enterprise, employees are aware of the audience on the network and hence tend to write fuller sentences and grammatically
correct posts. Posts become conversational when replies are posted, but there is adhoc activity as well. During campaign promotions or events in the enterprise, the activity on Yammer goes up as posts tend to beckon replies. Overall, the posts result in considerable variety within the overall company's Yammer network.

The insurance giant felt it crucial that they maintain the scale and influence of a large company but be able to deliver a personalized experience and the feeling of a close-knit community for every policyholder – to think global but act local. Engaged employees are better equipped to deliver this experience. The key benefits identified in using Yammer to achieve this goal included:

- More Transparency: Employees are more connected and open to communication.
- Talent Identification: Conversations and ad hoc teams expose hidden expertise and new talent.
- Enhanced Knowledge Flow: Information is not siloed into specific regions or departments.
- Better Innovation: The best ideas are surfaced and discussed more broadly.

### 3.1 Current Engagement Visibility Approach

An enterprise engagement score is a metric that quantifies how engaged the workforce in an enterprise is. It is designed to echo the loyalty of employees towards the enterprise and
represents the overall company morale. Improving the enterprise engagement score is a key business objective of the company. Currently the company makes use of Gallup Surveys [30] once every fiscal year to establish the engagement metrics for the enterprise as a part of their standard process. Figure 1 shows the Gallup's engagement hierarchy model, which shows how the questions in the survey are mapped to a pyramid for the engagement score across the company. All people leaders with teams larger than 10 receive organizational engagement scores.

Figure 1: Gallup Survey questions mapped to the Engagement hierarchy [30]

Figure 2 provides an overview of how the engagement score at the enterprise level has been consistently improving over the years. The score is on 1.0 - 5.0 scale.
The engagement score has been computed using the results provided in Gallup Surveys by the associates of the company. The exact process behind the computation of this engagement score is not transparent and hence this thesis does not discuss how the score is derived in detail. The score is computed based on the responses to the survey questions, each of which has their own rating scale.

Recently, more attention to social media in the press and with their own experiences outside the enterprise have led some to consider whether they should be leveraging some for the enterprise social data that is available. Around the company, different teams began manually monitoring the Yammer feed in an ad-hoc manner for collecting metrics such as number of posts, replies etc. for engagement purposes. A typical manual approach that is followed:

- Make a list of employees to be tracked
- Look at every employee’s feed manually – stare at the screen for hours
- Periodically check for responses
- Manually infer user behavior and note down the observations.

### 3.2 Problems

On interviewing the employees of the company and asking them about the manual approach, some of the notable issues that came up include:

- Difficulty in sifting through posts
- Finding specific posts and identifying intent/entities from the post
- User behavior hidden across messages
- Manual process is tiring.
- Carrying out the steps consistently

Yammer's characteristic is to supplement current communication media for employees to have employees in meaningful conversations on a daily basis. It is thus obvious to look at the feed for tracking engagement metrics more frequently. This thesis thus addresses the problem of managing enterprise social media feeds by providing a flexible analytics framework that can extract insights out of social media data and thus automate this process of tracking engagement information. This capability would enable stakeholders in the engagement metrics to be more responsive throughout the year.
3.3 Motivation

There is no tool or application currently that analyzes the content of these Yammer messages to gauge the amount and the level of collaboration taking place across the insurance enterprise. Currently available analytical tools for Yammer perform only statistical analysis of the messages like top 5 yammer users based on frequency of posting etc. – i.e. based just on counting. Further, preliminary analysis of the archived data indicated that - 55% of the total number of Yammer messages (~200k) across 4 years (Sept. 2008 - Sept 2012) were posted in the year 2012. This phenomenal increase in social media activity across the enterprise triggered the idea of mining the content of these messages to understand how the associates are making use of Yammer in their daily routine for engagement purposes.

To address this problem of engagement visibility, this thesis proposes a flexible and effective analytics framework that will surface user Engagement Personas from any internal social media of any enterprise. The framework is flexible as it is generic enough to be applied in the context of any enterprise using internal social media for promoting engagement opportunities.

Chapters 4 & 5 describe the proposed analytics framework and the implementation of the framework for this insurance enterprise.
Chapter 4: The Engagement Visibility Framework (EVF)

The AS-IS exploration in the previous chapter postulates that enterprises could benefit from structured information extracted from the unstructured internal social media feeds. This thesis proposes a real time analytics framework to mine social media data, which provides actionable insights thus potentially enhancing engagement visibility in an enterprise. This is called "The Engagement Visibility Framework" (EVF).

The EVF consists of a process to develop text mining dictionaries and extractors that will identify both the generic named entities and allow an enterprise to iterate over their own social media stream and construct enterprise specific or custom named entities. Enterprises are in some ways like communities and share common social norms and discourse in their communication. The custom entities that are extracted are meant to capture this distinction or difference that characterizes an enterprise. It will also allow the enterprise to explore these named entities to generate interesting insights that will potentially provide better engagement visibility.
4.1 Conceptual View

Figure 3 provides a conceptual overview of the software layers of EVF:

![Conceptual View of the EVF](image)

The three main layers are described in the following sections.

4.1.1 Entity Analytics Layer

The first (bottom most) layer or the Entity Analytics Layer involves extraction of meaningful business entities or named entities from social media messages that can be
further aggregated downstream. An entity is an abstraction for a set of meaningful values. (ex. Person and Phone Number). These entities can be further classified into generic and custom entities. Generic named entities are named entities that span across domains i.e. entities like Person, Phone Number, and URL etc. Custom entities are specific to the domain of the problem, in this case an enterprise. Ex. Meeting names, Industry services etc. A special type of custom entity is a derived entity like the user intent or sentiment which can be inferred from the context or other entities extracted from the posts. Figure 4 summarizes the above discussed classification hierarchy at the entity analytics layer.
Thus the layer primarily consists of tools and approach to extract these named entities. Once the entities are extracted, this layer then involves the population of the Engagement Persona data model which captures both the extracted entities of all types and also captures relationship information via a set of rules. This data model becomes the input to the next layer - Summarization Layer. The Engagement Persona data model is described briefly in Section 4.2.2.

**Rationale behind adopting a rule-based approach for entity extraction:** Several alternative methods were considered to carry out the entity extraction process including machine learning techniques. With machine learning, the heuristics for classifying and identifying entities are encoded within a parametric model and the process of altering them is not that transparent. The insurance enterprise supports various rule-based schemes in pricing and underwriting business processes and the technical organizations are more familiar with rule-based approaches. It is expected that the different teams like Human Resources & Communications in the enterprise will maintain this framework and alter it as per the need for tracking engagement. To avoid incurring additional overhead in having the analysts or the different Human Resources personnel to develop new skills, rule-based analytics becomes a logical choice. With dictionaries and regular expressions more easily configurable in a rule-based techniques, it thus is the most straightforward and transparent approach to be incorporated in this framework for an enterprise setting.
4.1.2 Summarization Layer

The second layer of the EVF is used to perform aggregations and analysis on the extracted named entities and relationships that are a rich source of signal. This layer provides a rich summarization of the social media streams and the other data sources. This layer considers two forms of aggregations, Structural and Semantic. The Structural aggregations are simply a "group by" on a particular entity that gives statistics like number of posts containing a particular entity. (ex. Number of Posts Vs. Phone Number Mentions). The semantic aggregations visualized in the form of social graphs, that lead to surfacing the Engagement Personas is an important contribution of this analytics framework. They correlate the external employee profile information with the entities extracted from the employee's social media posts to create these Engagement Personas. This thesis defines an “Engagement Persona” as a set of facets of user behavior derived from social media and other data sources representing a person or a group of persons which potentially enhances engagement visibility in the enterprise. It will evolve over time thus requiring the framework to be flexible. It can either be for a single person or it can be at the level of role or teams within the enterprise. The underlying data model of the Engagement Persona (Section 4.2) becomes important in managing and visualizing the Persona effectively due to its flexible nature and dynamic extensibility. Section 4.2 provides a detailed overview of the Engagement Persona and the underlying Data Model.
4.1.3 Business Process Layer

The topmost layer (third layer) involves introducing the metrics or Persona data into the existing processes. Various data visualizations can be used to present these results to the different teams for tracking the engagement metrics to foster collaboration in the enterprise.

It is important to mention that the bottommost layers i.e. the Entity Analytics Layer & Summarization Layer are iterative and can be refined over time. This thesis demonstrates the efficacy of this framework for an insurance industry use case wherein employees make use of a micro blogging tool like Yammer to engage in meaningful conversations.
Figure 5: Interaction Model for EVF
In Figure 5, the diagram shows a single user interaction with the EVF and how all the different layers of the EVF interact with each other. Any employee using the enterprise social media will post messages on the network. The entity analytics layer of EVF will takes these posts as input and execute analytics. A named entity extraction request is sent to the rule-based text analytics module for executing the analytics. This module uses the rules and the dictionaries for the extraction process. On returning all the named entities, the entity analytics layer will convert them into a triplestore format which is discussed in the next section. Whenever an employee from the Human Resources or the Communication teams is interested in tracking engagement, the employee will send a request to the visualization tools in the summarization layer to view the Engagement Persona visualization. The EVF triples generated at the entity analytics layer are then fed to the visualization tools in the Summarization layer upon this request. The visualizations of the Engagement Persona are rendered to the employee. The Human Resources or the Communication teams then analyze and infer from the Engagement Persona and decide on the necessary action items for boosting engagement. These interactions are a part of the business process layer.
4.2 The Engagement Persona

In this thesis, an Engagement Persona is an instance of a Persona in the context of the digital discourse of an enterprise related to engagement. This Persona is constructed in the Summarization Layer of the framework by using insights from the enterprise social media data and thus augmenting the existing employee profiles. The Persona can be at an individual level or at the role/team level. Also, since it is derived from social media posts that evolve over time, the EVF can expose a temporal aspect to the Persona. Figure 6 provides a conceptual view of The Engagement Persona.
The Engagement Persona is comprised of four kinds of named entities and semantic relationships between the entities. It provides a coherent view of all the named entities within a data model through visualizations in the form of social graphs. The notion of coherent is due to the fact that it is an aggregation or a summarization of all the named entities mentioned across several social media posts cross referenced with employee's profile information. Meaningful views can be created that will be illustrated later. The external entities are the named entities that are readily available in the standard employee profiles (ex. designation and team etc.). The entities from the social media post are the named entities - both generic and custom. The user mentions from the social media post
are explicit mentions, i.e. the social media user handle is used in the post. For example - "I like to talk to @John and @Martha". The derived entities from social media post are the entities which are derived using rules (code) and not explicitly mentioned in the post. For example - "I love Lotus Notes", messageType = "Helpdesk" becomes a derived entity. All these different facets or entities of a specific Engagement Persona provide a quick overview of the user behavior over time. It is thus a coherent view of all the named entities that are extracted in the Entity Analytics Layer.

The visualization of the Engagement Persona through social graphs can get quite complex and dense if all the named entities and relationships for all employees are considered at once. An example of such a dense visualization is shown in Figure 7. The figure provides a visualization of Engagement Personas of multiple employees expressed in the form of a dense semantic social graph.

Figure 7: Semantic social graph showing Engagement Personas of multiple employees
Thus in order to manage, navigate through and study the graphs, this thesis introduces a notion of managing the depth of the graph using the Engagement Visibility Level.

4.2.1 The Engagement Visibility Level

An Engagement Visibility Level (EVL) is the level of depth of the enterprise social graph that will help manage the density of the rich graph to convey the Engagement Persona effectively. For example, an EVL of Level 1 in a semantic social graph will only display the entities or relationships that are directly referenced by the user or the group of user's whose Engagement Persona is being represented by the graph. Whereas Level 2 will provide one level deeper insight and so on and so forth.

Figure 8: Semantic Social Graph with Engagement Visibility Levels
In Figure 8, the semantic social graph is represented using the Engagement Visibility Levels. The X axis reflects the cumulative references to the different entities of the graph as of the time point. It shows the progression of time from \(t_1\) to \(t_2\). The Y axis is the Engagement Visibility Level. It shows how at EVL Level 1 through time \(t_1\), the graph is sparse. And at EVL Level 1 and time \(t_2\), the graph has evolved to become richer with more number of relationships and references to entities. Further the notion of EVL becomes clear as the graph becomes denser from Level 1 to Level 2 at time \(t_1\) as the insights are a level deeper. Such a view is thus more manageable and improves traceability of engagement. It also provides a way to focus on the Engagement Persona of specific individuals or groups of individuals more easily.

Figure 9: Semantic Social Graph at Engagement Visibility Level 1
In the Figure 9, the graph shows the Engagement Persona of an intern marked as the red node at the center at Level 1 and time t1. This intern directly references four kinds of named entities in the social media posts including data analytics and IBM. In the Figure 10, at Level 2 and time t1, the named entities referenced by the intern are expanded to show which other users reference those entities. So an AVP of the company called Tara and a consultant called Bruce mention IBM and data analytics in their posts respectively. Thus a potential insight or an action item as a result of such engagement visibility would be to engage Tara, Bruce and the intern as they all talk about the same topics namely data analytics and IBM on the social media network. Thus in Figure 10, Level 1 and time t2,
the graph could thus become richer as the intern now references (mentionsUser relationship) Tara & Bruce as named entities in the social media posts thus potentially boosting engagement.

4.2.2 The Engagement Persona Data Model

A key artifact that is obtained from the Entity Analytics Layer and passed on to the Summarization Layer is the augmented data model which is called - The Engagement Persona Data Model.

<table>
<thead>
<tr>
<th>messageId</th>
<th>Predicate</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:hasAuthor</td>
<td>&lt;author name&gt;</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic:hasGenericNamedEntity1</td>
<td>&lt;generic&gt;</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:hasCustomNamedEntity1</td>
<td>&lt;custom&gt;</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Derived:hasDerivedNamedEntity1</td>
<td>&lt;derived&gt;</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External:hasExternalEntity1</td>
<td>&lt;external&gt;</td>
</tr>
</tbody>
</table>

Table 2: A sample Engagement Persona Data Model

Table 2 provides an overview of the proposed initial Engagement Persona Data Model. This thesis proposes a data model that has a structure similar to a triplestore containing 3 columns. It is a simplified version of a triplestore and does not have the formality of RDF (Resource Description Framework [32]). First column is the social media post identifier - messageId. The second column is the predicate that describes the named entities that are extracted & related to the posts. The third column is the Value, which is the actual instance of the predicate.
This thesis proposes a format for the data model that is a triplestore. The rationale behind adopting this model is to have a flexible and transparent data model that can be visualized effectively while summarizing various views of the data. This model can easily evolve over time as the engagement profile evolves by adding every identified named entity as a new triple in the triplestore. This model thus becomes key in summarizing and exposing The Engagement Persona as it provides a method to dynamically construct and add more named entities as a triple that evolve over time. For example, if sentiment of a message is extracted as new entity, the triplestore can be simply augmented by adding the sentiment as a derived custom entity triple. The predicate column in the triple helps make the named entities more self explanatory and explicit. Additional entities such as timeframe can be obtained through the messageId linking the original entities from Yammer.

4.3 Value Proposition

Figure 11 describes how The Engagement Persona is a key contribution of this thesis as it transforms unstructured information into a multi-dimensional view of an employee which is the Engagement Persona.
As seen, the current view shows how the Yammer Feed or the social media feed has no structure and has to be looked at manually. Further, the employee HR fact record provides little hint about an employee’s engagement. In the proposed view, the thesis shows how the two sources of information can be correlated to generate a meaningful multi-faceted Persona by adding structure to the social media feed. This model evolves over time as the social media activity feed is updated by employees. The idea is to provide such multi dimensional information asset to the intended user in the form of a faceted browser as a part of future work. Better engagement visibility provided with this information will help drive better and timelier decisions.
Chapter 5: Implementation of EVF - Insurance Industry Use Case

This chapter now describes the implementation details of the EVF for the insurance industry use case described in Chapter 3. To tie back to the use case described earlier, the insurance company uses Yammer internally as an enterprise social media feed for communication and collaboration and some have observed that it could support engagement analysis. Chapter 3 described the various problems associated with engagement visibility from this social media feed within the company. The goal was to apply the EVF in this context and derive The Engagement Persona from this Yammer data and the existing employee profile data of the large insurance company.

It is important to note that not all parts of the framework have been implemented fully for this use case. Given the scope and the time available, the first two layers of the EVF were implemented. The Summarization Layer built is exploratory and there is still some scope for generating the ideal set of graphs for visualizing the Persona effectively. As a part of the future work, a dashboard for faceted browsing of the Engagement Persona would be the culmination of all the results generated. Also user studies for such a dashboard when in production will validate the usefulness of the EVF in providing the enhanced engagement visibility.
The next two sections will describe the datasets used and the underlying implementation details for the insurance company setting.

5.1 Datasets - Insurance Company

There were two kinds of datasets that were available - social media or the Yammer dataset and the existing employee profile data from the organization's corporate directory. The next section describes the structure of both these datasets.

5.1.1 Yammer Data

The investigated data was four years of archived Yammer data used by the insurance company. The time period was from September 2008 (when the company started using Yammer) to February 2013. The data was provided in the form of CSV files. This archived data was considered to build the entity extraction layer which is a potential real time module whenever the system will go into production and will tap in a real time Yammer stream.

The messages.csv file contains all the Yammer messages with details like sender name, sender id, whether it was posted on a group, whether it was a reply comment or not etc including the creation timestamp. There were approximately 200K messages posted by “User”. 60% of these messages are reply comments and 40% are the original Yammer posts. Following is the schema of the CSV file. The important attributes mentioned in the
Table 3 below are the ones that are used as a part of the Engagement Persona data model that is generated as an output of The Entity Analytics Layer.

<table>
<thead>
<tr>
<th>id</th>
<th>Group_id</th>
<th>Group_name</th>
<th>Sender_email</th>
<th>Body</th>
</tr>
</thead>
</table>

Table 3: Important attributes of the Yammer Data Model (messages.csv)

### 5.1.2 Existing Employee Profile Data

This data was obtained from the organization's directory database which contains the basic factual information about every associate or employee of the insurance enterprise. This becomes the set of external entities or the employee profiles which are augmented with the social media entities. The important attributes are listed in Table 4 as follows:

<table>
<thead>
<tr>
<th>employeeEmail</th>
<th>title</th>
<th>office</th>
<th>department</th>
<th>Phone</th>
</tr>
</thead>
</table>

Table 4: Important attributes of the Organization's Call Directory Employee Profile

Each of the attributes is self explanatory. The "title" attribute refers to the designation of the employee in the company. The "office" attribute refers to the major business units of the company. The "department" on the other hand is a much narrower classification wherein employees from the same "office" can belong to different departments.
5.2 Implementation

This section describes the tools and the approach used for building the system. Figure 12 describes the implementation of the EVF for the insurance company use case. Each layer is annotated with information about which technology was used to build that layer and what output each layer rendered, that was used as input in the subsequent layers.

![Diagram of technologies used for implementing EVF](image)

This view provides an overview of the technologies used for implementing the different layers of the EVF. The Engagement Persona data model generated by the Entity Analytics Layer is fed to the Summarization Layer for generating the Engagement Persona. These personas are then used at the Business Process Layer for better
engagement visibility. Sections 5.2.1 to 5.2.3 provide a detailed overview of each of the layers and the technologies used.

![Diagram](image.png)

**Figure 13: High Level Object View of the EVF Implementation**

The Figure 13 provides a high level overview or a blue print of how the EVF will be used at the insurance enterprise for engagement visibility. Employees will use Yammer for posting messages. These messages will be consumed by the system implemented using EVF. The EVF implementation will thus generate the entity triples for the Engagement Persona. These triples will be used by the visualization tools for generating interesting graphs or networks for understanding the Engagement Personas effectively. These visualization tools will be used by the employees and the interested parties in the company with a goal to achieve improved engagement visibility.

Figure 14 provides a basic user scenario where the implementation will be used in production and a real time Yammer stream can be tapped. It gives the details about a single user interaction with the EVF implementation.
An associate (employee) will write a post on Yammer. The EVF Implementation will then consume these posts real time and execute the text analytics on them to extract all the named entity triples as per the Engagement Persona Data Model. The employee will then issue a request for exploring the visualizations and the EVF will render all the aggregations including the Engagement Persona to the end user as and when required.
To understand how the different building blocks of the EVF implementation exactly function together, Figure 15 provides a detailed overview of the interactions that take place between the different components inside the EVF implementation.
Figure 15: Interaction Diagram of the EVF Implementation
Figure 15 shows how the archived Yammer data and the organization directory data is consumed by the entity analytics layer of the EVF Implementation. The named entity extraction requests to the text analytics rule module result in extraction of all the generic and custom named entities from the Yammer post. This information asset is then converted into a triplestore format. This output of the entity analytics layer is then stored into the database. This data is further transformed into a format acceptable by tools like Tableau and Gephi for meaningful visualizations leading to an effective display of the underlying Personas.

In the next few sections, this thesis describes each of the implemented layers in detail. It also provides an overview of the technologies used for the implementation.
5.2.1 Entity Analytics Layer using Rule-based Text Analytics (IBM SystemT)

For building the entity analytics layer, the rule-based text analytics toolkit by IBM Research called *IBM SystemT* [26] was used. This toolkit is a part of a bigger platform called *IBM InfoSphere Streams* [7] which is a big data platform for performing real time analytics on streaming data. This platform lets us tap in a real time stream. Currently the system prototype uses archived Yammer data, but when it goes into production, the Yammer stream can be tapped in real time. The rule-based module uses dictionaries and pattern matching techniques to extract all the named entities from the Yammer posts. The output of this module for the data at hand is as follows:

<table>
<thead>
<tr>
<th>messageId</th>
<th>Predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:hasAuthor</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:inGroup</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasPhoneNumber</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasURL</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasLocation</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasPerson</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasOrganization</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic: hasAddress</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom: hasTopic</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom: hasUserMention</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Derived: hasMessageType</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External: hasTitle</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External: hasOffice</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External: hasDept</td>
</tr>
</tbody>
</table>

Table 5: The Engagement Persona Data Model for the Insurance Company use case

Table 5 is The Engagement Persona Data Model for this use case. An important point to note is that this data model can be made richer over time by extracting more named
entities. Given the scope and the time at hand, this is the data model that could best represent the Engagement Persona.

<table>
<thead>
<tr>
<th>messageId</th>
<th>predicate</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:hasAuthor</td>
<td><a href="mailto:bob@insurance.com">bob@insurance.com</a></td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:inGroup</td>
<td>Sharepoint</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic:hasLocation</td>
<td>Columbus</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic:hasPerson</td>
<td>John Doe</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Generic:hasOrganization</td>
<td>ABC</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Custom:hasUserMention</td>
<td>@john</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Derived:hasMessageType</td>
<td>SuccessStory</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>Derived:hasMessageType</td>
<td>Congratulatory</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External:hasTitle</td>
<td>AVP</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External:hasOffice</td>
<td>Corporate</td>
</tr>
<tr>
<td>&lt;id1&gt;</td>
<td>External:hasDept</td>
<td>IT</td>
</tr>
</tbody>
</table>

Figure 16: Engagement Persona Data Model populated for a sample Yammer message

In Figure 16, a sample Yammer message and the organization directory data of the author of the message is transformed into the Engagement Persona Data Model. This figure provides a better understanding of the underlying model and it can be easily seen how extensible the model for accommodating more named entities.
All the generic named entity triples are extracted using the off the shelf rule libraries available in [26]. The external named entity triples are obtained from the organization directory data. The custom and the derived entities specific to the discourse of the insurance enterprise under consideration are extracted using a custom build annotator consisting of pattern matching rules & respective dictionaries. Some of the rules written as a part of this module are explained in the Appendix A.

Figure 17 shows the overall implementation stack for the entity analytics layer. A Streams Processing Application as a part of the IBM's platform was written to consume the archived datasets and that calls the generic and custom text analytics operators for extraction of the named entities from the Yammer posts. Towards the end of the application, all the output entities are converted into a triplestore format and are written to a local file. The organization directory data is also processed and converted into a triplestore format. This forms the set of external entities that contribute to the Engagement Persona. It is important to note that these 2 entity sets: named entities from the Yammer post and the external entities can be correlated by performing a match on the employee's email in the organization directory and the sender's email for the Yammer post. The triplestore files are then imported into a database for further aggregations as a part of the Summarization Layer.
5.2.2 Summarization Layer - Tableau & Gephi

The Engagement Persona data obtained from the Entity Analytics Layer is then used for generating structural & semantic aggregations as a part of the Summarization Layer. The data is first filtered and reorganized in a database so that it can be consumed by the data visualization tools.
Tableau [27] was used for the structural aggregations, wherein just a *group by* on a particular named entity was performed to find the total number of Yammer messages containing the named entity. This gave an overview of trending topics and users. For building more interesting formalisms between users and entities, Gephi [8], an open source graph visualization software was used. Such visualizations will have users as nodes and edges between the nodes indicated by user mentions. Each user node will have the additional attributes extracted into the Engagement Persona data.

The results obtained are discussed in detail in Section 5.3.

**5.2.3 Business Process Layer - Engagement Tracking Process**

The implemented EVF also enhances the existing business process of the insurance enterprise for tracking engagement visibility with the implementation of the EVF. The direct beneficiaries of this framework at this layer in the insurance company are the Human Resources, Communications Team and the higher management.

The proposed modified business process is shown in Figure 18. Thus it can be noted that it does not eliminate the human in the loop completely and the process of engagement visibility for making decisions is essentially iterative over time as the social media feed evolves periodically. Thus the EVF essentially provides visibility into the engagement process but ultimately the human in the loop or the analyst makes the decisions and decides the action items for boosting the engagement opportunities. Also, the engagement
insights from the social media complement the Gallup survey results leading to a holistic understanding of the Engagement Personas of employees.

Some sample scenarios describing this process are as follows:

**Scenario 1:**

There are various roles in the enterprise like Associate Vice President, Intern, Consultant etc. All of these roles talk about "blood donation" on Yammer which will be extracted as a part of the Engagement Persona and can be visualized using the Gephi software. So as a part of the Act step of the business process, the Human Resources could send an
invitation to the Blood Donation Yammer Group to those employees and ask them to promote the campaign. This will thus lead to a sense of empowerment and thus boost engagement.

*Scenario 2:*-

Suppose in the graphical visualization, it is observed that an employee is talking about the same named entity over time on Yammer. Then the HR or any other management team can infer how the employee can be a best fit for a particular role. For example, if an employee John is talking about Java very frequently in his posts, then his role can be evaluated if required to put him in the best team working on Java.

*Scenario 3:*-

Suppose a manager has to choose between two employees to take on a particular role, the manager can refer to the graphical visualization of the Engagement Personas of these employees to look at different aspects or traits of the employees. The manager may end up choosing an employee who is better at influencing other employees on Yammer and resolving technical doubts even if the other employee might have a better skill set for the role. Thus such insights can potentially lead to better personnel management in an enterprise.
5.3 Results

This section explains the results obtained for the insurance company use case. Some of the results were shared with the various beneficiaries, and their feedback is positive and promising. The results are categorized as per the process adopted in building this framework which was incremental. Initially, some preliminary data analysis was done to get some rough idea about the data. The other layers of the EVF where then built bottom up and the thesis presents results from each layer.

5.3.1 Preliminary Data Analysis

This data analysis was carried out prior to building the entity analytics layer. This analysis was done to explore about the nature and the underlying schema of the data.

There has been significant increase in the number of messages posted on Yammer per day over the past 4 years. 55% of User messages over these 4 years were posted in the year 2012 as seen in Figure 19.
Figure 19: Yammer activity for the period 2008-2012 across the insurance enterprise

Other key initial results:

- The text mining package of R [29], a tool for statistical computing was used for establishing some basic statistics from the data. This helped in finding the top 5 users, top 5 posts etc.
- A key insight was that the top 5 users in terms of the number of posts on Yammer, included the COO of the company and also an employee who used to spend time in just welcoming every new person joining the Yammer network. This highlighted the diversity of context and impact likely to be found in posts.
- Some of the dictionaries were built using common words from the posts that were found using the tool's machine learning package, that were later used for the rule-based entity extraction.

5.3.2 First Layer: Entity Analytics – Results

The rules of the implemented entity analytics layer classified 70% of the original messages into 6 broad categories. The 7th category called "Other" is the 30% of the
original messages which still remain to be explored. For designing these rules, the data was first manually analyzed for deciding the classification scheme. Thus this process uses a supervised classification to decide the different message types. As a part of this process, some custom dictionaries were built to identify each message type or class. For example, "helpdesk" hashtag was one of the keywords for classifying a message as a "helpdesk" message. Some of the dictionaries were obtained from the different teams in the enterprise. And the remaining were obtained by frequency analysis of words across messages using R. On building these dictionaries, the rules were designed to classify or tag every message with a broad messageType entity as a derived custom entity. Once all the original messages classified, the same classification rules can be applied to the reply comments. Further by tweaking the rules and extending the dictionaries, it can be used for identifying all the classes of information specific to reply comments as well. The number of original messages that were considered were 50K. Figure 20 provides an overview of the different classes of messages that were identified.
5.3.3 Second Layer: Summarization Layer - Results

This section describes the different aggregations achieved using the Engagement Persona Data obtained from the Entity Analytics layer. Both the structural and the semantic aggregation results are discussed in this section.

5.3.3.1 Structural Aggregations
These aggregations or histograms were generated by doing a "group by" on some of the named entities extracted from the Yammer posts. These give an idea about the frequency of various named entities across the Yammer posts.

Figure 21: Histogram showing the count of Yams in each identified class

Engagement Insight: In the histogram shown in Figure 21, Long/Short Yams categorization is another aspect which classifies the different message types into long Yams (message composed of >10 words) and short Yams (<=10 words in the message). This histogram shows how "questions" and "ITSD" (helpdesk) messages top the number of Yammer messages. Interesting to note is that "SuccessStories" class which contains
Yams wherein employees talk about customer related success stories since it is an insurance enterprise setting always have longer Yams.

![User Mentions Vs. ITSD Topics](image.png)

Figure 22: #Yams in 2012 per topic related to the IT Service Desk (ITSD)

*Engagement Insight:* In Figure 22, the chart shows how “Sharepoint” is the most talked about topic in 2012. This is backed up by the fact that Microsoft acquired Yammer in 2012 and there have been many talks about integrating Sharepoint and Yammer effectively to increase collaboration.
Engagement Insight: In the Figure 23, the chart shows how "blood donation center number" of the insurance enterprise is the most mentioned phone number across the Yammer messages. Most of the mentions were during the blood drive campaign. This can be valuable for the HR/higher management to understand the Persona of the enterprise as a whole. This result was added as a Yammer post for feedback which invoked a positive response from the employees of the insurance enterprise who saw this Yammer post.
Figure 24: #Yams mentioning each city in the United States

Engagement Insight: In Figure 24, the chart shows how "Columbus" followed by "San Antonio" are the most mentioned cities across Yammer posts. The insurance enterprise is headquartered in Columbus and hence the numbers are not surprising. The buzz for "San Antonio" was due to the new data center that opened up recently in the insurance enterprise. Again a useful insight for the HR to understand the trending topics related to the location of the company offices.

5.3.3.2 Semantic Aggregations

The next set of graphs (Figures 25 - 30) is derived from the named entities forming a relation (predicate) - "User A mentions User B". Thus each node in the graph is a Yammer user (an employee). An edge between 2 nodes indicates that at least one of the users has mentioned the other user in his/her Yammer message by using the Yammer handle. For example, if John is the author of the message, his message will read as-
"@Mark The meeting went well. And we are not headed towards a dead end!". Here
@Mark is thus an explicit user mention. So a graph representing such information will
have 2 user nodes for John and Mark. There will be an edge between these 2 nodes. Each
edge is annotated with an attribute where predicate = "mentionsUser". These semantic
social graphs representing Engagement Personas of specific employees (individuals) are
further annotated with the EVL and the time to tie back to the conceptual view of the
Engagement Persona. It is important to note that the EVL is not implemented yet and is a
part of the future work.

Figure 25: Semantic Social Graph with "Office" = "Corporate"

*Engagement Insight:* In Figure 25, the attributes at each node label are in the form -
<empNum-role-department-office>. This graph shows how employees within an "Office"
but having different roles and in different departments in that Office are actively engaged
in conversations. Nodes are the employees belonging to the same "Office" and are
colored as per the "Office". The time instance here is ti which is the timestamp of set of
Yammer messages considered for the extraction of this Engagement Persona. The EVL here is Level 1, as it shows references to the entities (users in this case) explicitly mentioned by the node 110153 which represents an employee with employee number 110153. This graph has a third dimension called predicate = "mentionsUser", thus limiting the number of edges and the entities.

Figure 26: Semantic social graph showing cross collaboration at the "Office" level.

Engagement Insight: In Figure 26, the attributes at each node label are in the form - <empNum-office>. This graph shows how employees across Offices are collaborating with each other. Nodes are colored as per the "Office". The time instance here is ti which is the timestamp of set of Yammer messages considered for the extraction of this
Engagement Persona. The EVL here is again Level 1, as it shows references to the entities (users in this case) explicitly mentioned by the node 118960 which represents an employee with employee number 118960. This graph also has a third dimension called predicate = "mentionsUser", thus limiting the number of edges and the entities.

Figure 27: Semantic social graph showing cross collaboration at the "Dept." level.

*Engagement Insight:* In Figure 27, the attributes at each node label are in the form - `<empNum-role-department-office>`. This graph shows how employees across different departments are cross collaborating with each other. All of the employees are from "Nationwide PCIO" office. Nodes are colored as per the "Department". The time instance here is ti which is the timestamp of set of Yammer messages considered for the extraction of this Engagement Persona. The EVL here is again Level 1, as it shows references to the
entities (users in this case) explicitly mentioned by the node 109256 which represents an employee with employee number 109256. This graph also has a third dimension called predicate = "mentionsUser", thus limiting the number of edges and the entities.

![Semantic social graph - Sparse in nature](image)

**Figure 28: Semantic social graph - Sparse in nature**

*Engagement Insight:* In Figure 28, the attributes of the node labels are of the form `<empNum-role-department-office>`. It shows how there are small groups of employees engaged in sporadic conversations. This triggers an insight of engaging these employees in larger groups/conversations. Nodes are colored as per the "Office". The time instance here is ti which is the timestamp of set of Yammer messages considered for the extraction of this Engagement Persona. The EVL here is again Level 1.
Figure 29: Semantic social graph - Dense in nature at EVL 3

*Engagement Insight:* In Figure 29, the attributes of the node labels are in the form `<empNum-office>`. Nodes are colored as per the "Office". The time instance here is `ti` which is the timestamp of set of Yammer messages considered for the extraction of this Engagement Persona. The EVL here is Level 3, as it shows three levels deeper references to the entities (users in this case) explicitly mentioned by different employees. This graph also has a third dimension called `predicate = "mentionsUser"`, thus limiting the number of edges and the entities. It shows how there is a possibility of community detection in such a graph. One of the most relevant features of graphs is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges are joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of
a graph, playing a similar role. The nodes 73271 and 111854 are employees who have triggered a larger conversation causing a ripple effect to engage more employees. Such a pattern separates the graph into 2 communities or clusters at these nodes. Another interesting insight, the users "111854" in the right part of the graph is a potential influencer as the user mentions different kinds of users falling under different "Offices" across the company.

![Engagement Insight: Collaboration across offices](image)

Figure 30: Semantic social graph aggregated on "Office" attribute

*Engagement Insight:* In Figure 30, the graph shows how the Nationwide PCIO (Property/Casualty Insurance Operations), Allied and the Corporate offices are the most actively engaged offices in the company. They thus have an influencing Engagement
Persona. It also provides an insight about which of the offices are not actively engaged on Yammer. Hence this can be an action item for future engagement campaigns.

5.3.4 User Survey

As a part of validation of the problem statement, a user survey was carried out asking the importance of engagement and how Yammer can be used for engagement purposes. Some interesting insights validate the purpose of this thesis in building this rich analytics framework - The Engagement Visibility Framework. Figure 31 and Figure 32 shows all the questions of the survey sent to the employees.

<table>
<thead>
<tr>
<th>Nationwide: Process of Tracking Engagement Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is Gallup survey (once every year) enough for tracking engagement?</td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>

| 2. How often do you think engagement should be tracked for Nationwide? |
|-----------------|----------------|
| Yearly | Weekly |
| Quarterly | Daily |
| Monthly | |

| 3. Do you think Yammer is important for tracking engagement? |
|-----------------|----------------|
| Yes | No |

| 4. Do you use Yammer? |
|-------------------|---|
| Active User | Passive User (Lurker) |
| Sometimes | No |

| 5. How easy is it to track metrics on Yammer? |
|-----------------|----------------|
| Easy | Hard |
| Moderate | |

Figure 31: User Survey Questions - Part 1
6. How useful is the Yammer analytics page (app that gives basic statistics like top users, topics etc.)?
- Very useful
- Useful
- Not useful
- Never Used

7. How easy is it to infer a user's behavior manually from his/her Yammer posts?
- Easy
- Medium
- Hard

8. How many man-hours on an average do you spend for tracking metrics using Yammer?
- 1 week
- 1 day
- More than 2 hours
- 1-2 hours
- Instant (matter of few minutes)

9. Do you need additional automated insights/summarizations from Yammer?
- Yes
- No

10. What kind of automated insights do you want from Yammer?

Figure 32: User Survey Questions - Part 2

Survey Results:-

The survey was sent to 20 employees of the insurance company. These 20 employees were selected based on their interest in engagement and collaboration activities across the enterprise. The rationale behind selecting these set of employees was to get constructive feedback and a clear idea about the requirements as they would ultimately be the direct beneficiaries of such a framework. The employees thus spanned across the HR, the Communications Leadership and the data analytics teams. The survey was taken by 14
out of the 20 candidate employees. All of the responses are anonymized. This thesis thus looks at the overall results for each question in the survey.

Figure 33: Results of the User Survey - Part 1

Figure 33 shows responses to the first set of survey questions. Interesting to note how 8 out of 14 of the employees feel Gallup Surveys are not enough for tracking engagement and feel engagement should be tracked frequently instead of the yearly Gallup Surveys. The most important result here is the answer to the second question which validates the motivation of this thesis. 70% of the employees feel Yammer is important for tracking engagement.
Figure 34 shows responses to the second set of survey questions. Majority of the employees feel that tracking metrics on Yammer is moderate to hard. Further inferring user behavior manually also falls under a similar bracket. This justifies the need for an iterative model or a framework like the proposed EVF in this thesis for addressing this challenge of providing a better engagement tracking method. Also most of the employees require more than 1-2 hours to track metrics. This problem is addressed in this thesis by providing a near real time framework that will allow shorter response times for tracking such metrics.
Figure 35: Results of the User Survey - Part 3

Figure 35 shows responses to the final two survey questions. 72% of the employees believe they require additional automated insights from Yammer. Further the question that involved free form comments for describing the various automated insights they would like got some interesting responses as well. Some of the common insights mentioned were finding hot topics, insights by team etc.
Chapter 6: Conclusion & Future Work

Standard employee profiles when augmented with Engagement Persona data, derived from enterprise social media, potentially enhance several levels of engagement visibility. Through a real world use case, this thesis shows that the Gallup surveys used in an insurance company for tracking engagement are complemented by user Personas from internal social media. Through iteration, the evolving Engagement Persona Data Model becomes useful in describing the overall user behavior effectively. The generic and flexible nature of the EVF makes it applicable across any enterprise using internal social media.

As future work at the implementation level, the Engagement Persona Data Model can be made richer by combining additional sources of employee information including the social media tool’s User Profile data. Exploring the different Engagement Visibility Levels through faceted browsers and other reporting schemes to help manage the density of the rich enterprise social media graph will be an important next step. This will aid the Human Resources or other teams in assessing (making visible) some of the enterprise engagement. As a part of next steps at the enterprise level, it would be interesting to see how the overall engagement score gets affected when these new behavioral insights
derived from social media using EVF are used in the process of computing the Engagement Score. Preparing user scenarios illustrating how the beneficiaries will interact with the social graphs and act upon the engagement insights obtained from the graphs will be an important next step. It will also be crucial to showcase and study the correlation of the EVF data with the actual Gallup Engagement Survey scores at the various organizational levels using the semantic social graphs which convey the Engagement Personas.
References


Appendix A: Rule-based Text Analytics

In this section, this thesis provides an overview of the Rule-based text analytics implementation stack that was used for the insurance enterprise use case.

![Diagram of Rule Based Text Analytics](image)

Figure 36: Key Components or building blocks of IBM SystemT [26] Constructs

Figure 36 provides a brief overview of all the different building blocks of the IBM SystemT rules. Using standard and custom dictionaries is a traditional approach. These
can be made configurable to the end user thus making the extraction process more manageable. For example, for a blood donation campaign, if the company is tracking posts for a certain hashtag related to the campaign, then an end user can simple add the hashtag in this dictionary to extract these additional Yammer posts for further analytics. Regular expressions are a standard way to write pattern-matching rules. For example, for extracting USA as a country, say the variations in the input text include U.S.A, U.S., and US. Then this can be easily addressed using a regular expression that handles all of these matches. SystemT [26] also provides additional tools for generating a regular expression automatically by providing the labeled matches from the text which makes the rule writing process easier. Sequence patterns provide a richer way of extracting named entities that have a specific underlying pattern. For example, an Email signature in an enterprise typically has a very specific format that includes the username of the person, followed by the “@” and then the name of the domain name of the enterprise. This can be easily expressed as a sequence pattern rule. User defined functions is just one more way of dealing with additional text transformations. All these building blocks help in making the rules richer and concise for gaining a good precision & recall for all the named entities to be extracted.
Rules to classify Yammer posts as "SuccessStories":-

--Extract the candidate success stories based on dictionary keywords

--Ex.: "Keep up the good job claims team. The customers in the Bay area are very happy with your service!! OYS Experience"

create view SuccessStoriesCandidate as
select D.message
from CandidateMessages_Local D
where ContainsDict('SuccessStoryCues.dict','IgnoreCase',D.message) and Not(ContainsDict('JoinedDict_Local',D.message))
consolidate on D.message;

--Filter out the false positives by looking for obvious mentions like "service" etc.

create view SuccessStories_Triples as
select D.message as message,'messageType' as predicate,'SuccessStory' as value
from SuccessStoriesCandidate D
where ContainsRegex(/Service/,'CASE_INSENSITIVE',D.message) and ContainsRegex(/!/,'CASE_INSENSITIVE',D.message)
consolidate on D.message;
export view SuccessStories_Triples;
Contents of the dictionary: SuccessStoryCues.dict :

<table>
<thead>
<tr>
<th>Dictionary Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>on your side</td>
</tr>
<tr>
<td>on their side</td>
</tr>
<tr>
<td>customer experience</td>
</tr>
<tr>
<td>great service</td>
</tr>
<tr>
<td>greatstories</td>
</tr>
<tr>
<td>Oys</td>
</tr>
<tr>
<td>Kudos</td>
</tr>
<tr>
<td>great job</td>
</tr>
<tr>
<td>customer testimonial</td>
</tr>
<tr>
<td>Compliment</td>
</tr>
</tbody>
</table>

Table 6: Dictionary entries for identifying success stories