2012 SUMMER PROJECT REPORT
Sentiment Analytics

Nationwide Enterprise

By Fan Fu

Team members: Fan Fu, Ramiya Venkatachalam
Supervisor: Bruce Craig, Tara J Paider
# 2012 Summer Project Report-Sentiment Analytics

## Contents

Yammer Sentiment Analytics .................................................................................................................. 1

Sentiment Analysis ................................................................................................................................. 1

1. What is Sentiment Analysis? ........................................................................................................... 1

2. What is Employee sentiment analysis? ......................................................................................... 1

3. Why Sentiment is hard to decipher? .............................................................................................. 2

4. Sentiment Dictionary ....................................................................................................................... 2

   4.1 AFINN Sentiment Dictionary .................................................................................................. 2

   4.2 Nationwide (NW) Sentiment Dictionary ................................................................................. 3

Analysis process based on CRIPS_DM steps ....................................................................................... 4

1. Business Understanding .................................................................................................................. 4

   1.1 Determining Business Objectives ......................................................................................... 4

   1.2 Assessing the situation ............................................................................................................ 4

   1.3 Determining the data mining goals ......................................................................................... 5

   1.4 Producing a project plan ......................................................................................................... 5

2. Data Understanding ........................................................................................................................... 5

   2.1 Collecting initial data ............................................................................................................... 5

   2.2 Describing data ...................................................................................................................... 5

   2.3 Exploring data ....................................................................................................................... 6

   2.4 Verifying data quality ............................................................................................................ 6

3. Data Preparation ............................................................................................................................... 6

   3.1 Selecting data ......................................................................................................................... 6

   3.2 Cleaning data .......................................................................................................................... 7

   3.3 Constructing new data ............................................................................................................. 7

   3.4 Integrating data ...................................................................................................................... 7

   3.5 Formatting data ....................................................................................................................... 7

4. Modeling ........................................................................................................................................... 8

   4.1 Selecting modeling techniques ............................................................................................... 8

   4.2 Generating a Test design .......................................................................................................... 8

   4.3 Building the Models .................................................................................................................. 13
4.4 Assessing the Model .............................................................................................................. 14

5. Evaluation .................................................................................................................................. 14

5.1 Evaluating the results ............................................................................................................... 14

5.2 Refine the sentiment analytics .............................................................................................. 14

5.3 Determining the application of sentiment message ............................................................ 15

6. Deployment ................................................................................................................................ 17

6.1 Planning for next steps research ............................................................................................ 17

Reference: ...................................................................................................................................... 19
Yammer Sentiment Analytics

1 What is Sentiment Analysis?

Sentiment analysis is an application of text analytics techniques for the identification of subjective opinions in text data. It normally involves the classification of text into categories such as "positive", "negative" and in some cases "neutral". Over the last five years, there is an increasing demand for sentiment analysis tools by companies willing to monitor customer’s or employee’s opinions of the company and on its products and services. To fulfill the increasing demands for such kinds of tools, more and more researchers and companies are releasing products to perform sentiment analysis, many of them claiming to be able to perform sentiment analysis of any type of document in every domain. Unfortunately, experience has shown us that, an "out-of-the-box" sentiment analysis tools working across domains does not yet exist. The main reason sentiment analysis is so difficult is that words often take different meanings and are associated with distinct emotions depending on the domain in which they are being used. There are even situations where different forms of a single word will be associated with different sentiments. For example, in customer feedback that the word "improved" was associated with positive comments, but "improve" was more often used in negative ones. All sentiment analysis tools rely, at varying degrees, on lists of words and phrases with positive and negative connotations or are empirically related to positive or negative comments.

2 What is Employee sentiment analysis?

In the past few years, some companies have been using text analytics software to analyze positive and negative words or phrases appearing in social media and other electronic posts to figure out what customers think about their products, service and policies. For example, firms in the travel industry can monitor comments posted on websites for users’ opinions about hotel properties, or perhaps about specific features such as housekeeping or food service. Less well-known but a growing reality is the use of this "sentiment analysis" technology by companies to gather information about their own employees.

Why are companies venturing into this area? One of the most basic is that this kind of analysis may provide a more effective tool for assessing key factors such as job satisfaction than internal surveys or other traditional methods. In mining unstructured text data, HR leaders and management teams can obtain insights on the degree to which employees are engaged, and IT help desk can gain insights on the degree to which employees meet problem. By gaining a clearer picture of employee sentiment, companies may identify areas where employees are dissatisfied and design strategies for enhancing engagement and in turn, improving productivity and employee retention.
3 Why Sentiment is hard to decipher?

Online comments don't fall neatly into "positive" and "negative" buckets. There's a range of consumer sentiment that challenges even the most sophisticated natural language processing technologies. In a Sentiment Analysis Symposium, Catherine van Zuylen, VP of products at Attensity, a social analytics software vendor, provided this list of difficult comment-analysis problems:

- **False negatives:** The words "crying" and "crap" suggest negativity, but then there is "I was crying with joy" or "Holy crap! This is great." Here's where simplistic tools might be fooled.
- **Relative sentiment:** "I bought a Honda Accord.", which is great for Honda but bad for Toyota.
- **Compound sentiment:** Doing work for movie studies. Such as "I loved the trailer but hated the movie." Big mobile phone companies encounter mixed messages such as "I love the phone but hate the network."
- **Conditional sentiment:** "If someone doesn't call me back, I'm never doing business with them again." Or "I was really pissed, but then they gave me a refund."
- **Scoring sentiment:** Vendors are expected to measure relative sentiment, but how positive is "I like it" versus "I really like it" versus "I love it"?
- **Sentiment modifiers:** "I bought an iPhone today :-)" or "Gotta love the cable company ;-<". Emoticons are straightforward, but what words are they connected to?
- **International sentiment:** Japanese have unique emoticons, like (;_;) for crying. Italians tend to be far more effusive and grandiose, whereas Brits are generally drier and less effusive, making those relative scoring challenges mentioned earlier all the more complicated.

Sophisticated systems can be optimized to handle these kinds of problems. But no amount of tuning will lead to perfection, so it's best to focus the extra effort on developing insight about and acting on the majority of clear-cut sentiments.

4 Sentiment Dictionary

4.1 AFINN Sentiment Dictionary

AFINN is a list of English words rated for score with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. The word list initiated from a set of obscene words as well as a few positive words. It was gradually extended by examining twitter postings. This sentiment dictionary also includes words from the public domain Original Balance affective word list by Greg Siegle. And added internet slang by browsing Urban Dictionary including acronyms such as WTF, LOL, and ROFL. The most recent additions come from the large word list by Steven J. DeRose, the compass DeRose Guide to Emotion Words.

There are two versions:

AFINN-111: Newest version with 2477 words and phrases. There are 878 positive words (873) and phrases (5), and 1599 positive words (1581) and phrases (18).
AFINN-96: 1468 unique words and phrases on 1480 lines. Note that there are 1480 lines, as some words are listed twice. The word list is not entirely in alphabetic ordering.

Since this word list is based on social media data, it is better for yammer sentiment analysis research. I used the AFINN-111 in my research.

4.2 Nationwide (NW) Sentiment Dictionary

We use a public AFINN sentiment dictionary for sentiment analysis testing, but we also built a Nationwide sentiment dictionary based on unique word occurrence in yammer message. A lot of effort is needed to develop a domain-specific sentiment dictionary and to identify the proper vocabulary associated with the expression of positive and negative feelings. (Detail is in below review process part)

In the Nationwide sentiment dictionary, there are 655 positive words and 329 negative words.

*Based on the yammer message as bellow:

"193922753","193922753","1488476328","Kathleen Hawthorne","(Processor II) has [Tag:9205:joined] the Nationwide network. Take a moment to welcome Kathleen.","2012-07-18T12:57:31.679-07:00"

Although these two words —“joined” and “welcome”—are classified into positive word list, they have no useful meaning that can present the positive sentiment. Since this is system automatic message. Therefore, these two words would be excluded from both sentiment dictionaries. In AFINN sentiment dictionary, there are 877 positive words and 1599 negative words. In Nationwide sentiment dictionary, there are 653 positive words and 329 negative words.
Analysis process based on CRIPS_DM steps

1. Business Understanding

1.1 Determining Business Objectives

Social media is an important data resource of Big Data. In social media, employers can get a lot of the latest and huge data, which contain various kind of information. Analyzing this kind of big data can help the employers understand more about employees, such as satisfy employee’s need and improve employee’s productivity. Thus, this research is commissioned the following objectives:

- Build a Nationwide Yammer Sentiment Dictionary
- Understand employee’s real-time sentiment by monitoring sentiment dictionary
- Analyze the employee’s sentiment change trend in specific period

Tentatively, the research will be judged a success if:

- The precision and accuracy of Yammer Sentiment Dictionary can higher than AFINN
- Employee’s real sentiment can be matched with sentiment analysis result
- The research finishes on time and under budget

1.2 Assessing the situation

It is the first attempt for big data analytics on employee sentiment in Nationwide Enterprise, and it is also a new research field in insurance industry. We are on the experiment and testing road to find business opportunity in the future big data analysis as much as possible.

Personnel: For our summer project research team, it is our first time to do research on big data analysis. We learn by learn, and we are confident for this research. Additionally, Nationwide hopes the research results would become a part of a continuing big data analytics example. Thus, managers should consider whether any position created for the future work focus on this area.

Data: After several weeks’ effort and discussion with several other departments, we decided to use the Yammer stream data in streams sentiment analytics. Yammer is an internal social media platform used by Nationwide employees, and data is collected by the SPOT department, which is convenient for us to get data privilege and analyze it. Moreover, yammer is a social media data, which can satisfy our initial data needs.

Risks: Since there are three persons form a research group to do research on this area, there is not a great deal of immediate risk in the venture. However, time is important and limited. We only have 9 weeks to accomplish the whole project, include learning new software, construct testing environment, write SPL code, generate testing design, and complete the research report. Additionally, there is no need much budget on equipment for this research, the experiment environment and data can be constructed in internal server.
1.3 Determining the data mining goals
The sentiment analytics goals for the initial study to be completed as below:

- Scan yammer data file to find interesting topic. Such as harsh tag, topic, sentiment, etc
- Use IBM InfoSphere Streams to analyze the sentiment in each yammer message and classify yammer message as positive (score greater than 1), negative (score less than 1), and natural (score equal to 0) three categories. Exclude message without sentiment words or phrases.
- Refine the sentiment analytics model to get higher precision and accuracy.

1.4 Producing a project plan
The overview plan for the research is as follow:

<table>
<thead>
<tr>
<th>Steps</th>
<th>Time</th>
<th>Resources</th>
<th>risks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business understanding</td>
<td>1 week</td>
<td>IT architecture, IBM</td>
<td>Data mining goals</td>
</tr>
<tr>
<td>Data understanding</td>
<td>2 weeks</td>
<td>IT architecture, IBM</td>
<td>Data resource problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Technical problem</td>
</tr>
<tr>
<td>Data preparation</td>
<td>2 weeks</td>
<td>IT architecture, SPOT department</td>
<td>Data resource problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Technical problem</td>
</tr>
<tr>
<td>Modeling</td>
<td>2 weeks</td>
<td>IT architecture Yammer</td>
<td>Technical problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model problem</td>
</tr>
<tr>
<td>Evaluation</td>
<td>1 week</td>
<td>IT architecture Yammer</td>
<td>Inability to evaluate result</td>
</tr>
<tr>
<td>Deployment</td>
<td>1 week</td>
<td>IT architecture Yammer</td>
<td>Inability to implement result</td>
</tr>
</tbody>
</table>

2. Data Understanding

2.1 Collecting initial data

**Yammer Messages**: the yammer messages contain much information file: Files, Groups, Messages, Users, and Topics. It needs to identify and extract useful data for further analysis

**Additional data**: Public sentiment dictionary—AFINN, which would be used in the streams application testing process of sentiment analytics.

2.2 Describing data

There are huge records in the yammer database. The initial data package we got from SPOT department contains five different CSV files: Files.csv, Groups.csv, Messages.csv, Topics.csv, and User.csv. The message file contains 222,821 message records from 2008-09-11 to 2012-07-18. Most of
the value types in these data source are string, like user name and body. Some of these are symbolic, like dates and time. And some of these are Boolean.

Data quantity:

- The data is CSV format
- The database is huge, we use output data for streams sentiment analytics

Data quality:

- The data contain much useful information about employee. It is useful for sentiment analysis
- The data types are various, number, string, and time, etc
- The effort would spend on finding insights in data, and form use case for future study

2.3 Exploring data

First, look up the attribute and records in each file to find the useful one. We identified the messages.csv file as research data, and we find the content of body attribute in message file is unstructured test data. Second, scan and read the body content to find is there sentiment words or phrases. Third, this exploration help clarify research idea about yammer data. At last, we decide use the body content of yammer message for employee sentiment analytics.

2.4 Verifying data quality

Missing data: the known missing data includes the blank posted by employee, or blank employee information

Data error: most of the data are automatically generated, so it is not a great worry. Typographical errors in the database can be found during the exploration process

Measurement error: the great potential source for measurement error is the questionnaire. Since the yammer data is automatically generated in to database, there is none measurement error.

Coding inconsistencies: none. All the data have their own value for each attribute.

Bad metadata: none. Data are generated automatically.

3. Data Preparation

3.1 Selecting data

Selecting items: There is research assumption that employee’s reply message to others cannot stand for the real internal sentiment. Thus, we would focus on original post message by employee in sentiment analytic research.
Selecting attributes: It is important to filter attributes such as id, thread_id, sender_id, sender_name, body, created_at from initial data resource. Since these attributes are enough for understand employee’s sentiment.

3.2 Cleaning data

Missing data: none

Data error: none

Measurement error: none

Coding inconsistencies: none

Bad metadata: none

3.3 Constructing new data

After Streams application process, it would generate a new data attribute named sentiment score. This score means the sentiment value of specified yammer message. If the score is greater than 0, it means positive words more than negative words. If the score is lower than 0, it means negative words less than positive words. If the score is equal to 0, it means negative words are equal to positive words. If there is no sentiment score for this yammer message, it means there are no sentiment words or phrases in this yammer message. Additionally, there is a new date and time format would be generated to replace the initial created_at attribute. Thus the final new data has seven useful attribute: id, thread_id, sender_id, sender_name, body, created_time, and sentiment score.

3.4 Integrating data

There are two integrations through this research process. The first is integrate SentimentScore (body, sentiment score) with OriginalMessage (id, thread_id, sender_id, sender_name, body, created_at) to form SentimentMessage1 (id, thread_id, sender_id, sender_name, body, created_at, sentiment_score). The second is integrate Time(created_at1, created_time) with SentimentMessage1 (id, thread_id, sender_id, sender_name, body, created_at, sentiment_score) to form the final output SentimentMessage(id, thread_id, sender_id, sender_name, body, created_time, sentiment_score)

3.5 Formatting data

The output data of Streams application is CSV format. For one thing, it keeps the consistency of input data format and output data format. For another thing, CSV format output data is more convenient be read by the tools that for data visualization and graphing.
4. Modeling

4.1 Selecting modeling techniques

In this research, it uses IBM InfoSphere Streams and Text Analytic toolkit (AQL) to generate sentiment analysis process, and uses Microsoft Visual Studio 2010 (C# development) to generate the unique word occurrence from yammer message content, which is used to build the Nationwide Yammer Sentiment Dictionary.

4.2 Generating a Test design

We use the Message.csv file as initial input data resource in sentiment analysis research process.

Total records: 222821

Number of attributes: 20

Here are the detail testing steps as bellow:

1. Replace the double quote “” with space in Note. Because the double quote would cause error when streams studio deal with csv format file. (In streams studio, the first quote is regarded as a start signal, and the next following quote is regarded as an end signal)

2. Use SPL text extract operator to generate some core annotator statistics, including address, city, country, date time, email address, location, organization, person, phone number, state or province, URL, and zip code.

3. Structure the input 1messages.csv files into SPL structure format, and use permissive parameter to skip incorrectly format record.

```csharp
composite Main {
    //define the Message1 type, which is used to specify and store the input data
    Message1 = rstring id, rstring replied_to_id, rstring thread_id, rstring conversation_id,
    rstring group_id, rstring group_name, rstring participants, rstring in_private_group,
    rstring in_private_conversation, rstring sender_id, rstring sender_type, rstring sender_name,
    rstring sender_email, rstring body, rstring api_url, rstring attachments, rstring deleted_by_id,
    rstring deleted_by_type, rstring created_at, rstring deleted_at;
    //define the Message2 type, which is used to specify and store six selected attributes data
    Message2 = rstring id, rstring thread_id, rstring sender_id, rstring sender_name, rstring body, rstring created_at;
    //define the Message3 type, which is used to specify and store sentimentscore data
    Message3 = rstring body, int32 sentiment_score;
    //define the Message4 type, which is used to specify and store SentimentMessage data
    Message4 = rstring id, rstring thread_id, rstring sender_id, rstring sender_name,
    rstring body, rstring created_time, int32 sentiment_score;
```
There are four incorrect format records are skipped. It is lower percentage than total records.

Use the "Punctur" operator to select particular attributes - leave out attributes like "sender_type" etc. For sentiment analysis, we only focus on the messages posted by user, not by system or others. Records that without user value are skipped.

Use the "Filter" operator to filter the message into two streams: one is original messages that posted by user, the other is reply messages. For sentiment, the original post can demonstrate the employee thought better than reply. Sometime, the reply message is very short and not the really though t of the employee.

There are 77704 records are original message.

There are 89662 records are reply message.

Ex:
4 Using AFINN positive word list extract positive record and calculate positive score from the original message. The positive record file has three attributes (body2, sentimentword2, score2). For example, if one yammer message has two positive words, it would generate 2 separate records, each one with corresponding positive word and score 1. Then, calculate the total positive score for every yammer message. The positive score file has two attributes (body3, score3). For example, if one yammer message has two positive records, it would sum the positive score from each record as 2.

```
{stream<rstring body1, rstring sentimentword1>PositiveMentioned;
stream<rstring body2, rstring sentimentword2, int32 score2> PositiveRecord;
stream<rstring body3, int32 score3> PositiveScore;
= com.ibm.streams.text.analytics::TextExtract(OriginalMessage) {
    //define a TextExtract operator to calculate the positive score for OriginalMessage
    param
        //define a specified text extract rules in AQL
        AQLfile : "Sentiment.aql";
        outputViews : "PositiveMentioned","PositiveRecord","PositiveScore";
        text : "body";
}
```

Ex:

```
PositiveScore.txt
"testing this out.... happy",1
"wow, great job, test@test.com",2
"good better bad ",2
"bad worse good",1
"good better",2
```

5 Using AFINN negative word list extract negative record and calculate negative score from the original message. The negative record file has three attributes (body2, sentimentword2, score2). For example, if one yammer message has two negative words, it would generate 2 separate records, each one with corresponding negative word and score -1. Then, calculate the total negative score for every yammer message. The negative score file has two attributes (body3, score3). For example, if one yammer message has two negative records, it would sum the positive score from each record as -2.
6. Generate the sentiment score for each yammer message. The sentiment score file has two attributes (body4, sentiment_score). For example, if one yammer message has three positive words and two negative words, it would generate the sentiment score 1 by sum the positive score and negative score.

   (stream<rstring body1, rstring sentimentword1> NegativeMentioned;
   stream<rstring body2, rstring sentimentword2, int32 score2> NegativeRecord;
   stream<rstring body3, int32 score3> NegativeScore)
   = com.ibm.streams.text.analytics::TextExtract(OriginalMessage) {
     //define a TextExtract operator to calculate the negative score for OriginalMessage
     param
     //define a specified text extract rules in AOL
     AQLFile     : "Sentiment.aql";
     outputViews : "NegativeMentioned","NegativeRecord", "NegativeScore";
     text         : "body";
   }

   Ex:

   ![NegativeScore.txt](image)

   "a bad thing happen",-1
   "good better bad ",-1
   "bad worse good",-2
   "bad worse",-2

7. Generate the yammer sentiment record by union PositiveRecord file and NegativeRecord file. The sentiment record file has three attributes (body2, sentimentword2, score2). For example, if one yammer message has two positive words and two negative words, it would generate four sentiment records with each sentiment words and corresponding score. This file store the sentiment words extract from yammer body content, which would be used for sentiment words analysis.
stream<rstring body2, rstring sentimentword2, int32 score2> SentimentRecord = Union(PositiveRecord; NegativeRecord) {
  //Union Positive Record and Negative Record
}

Ex:

<table>
<thead>
<tr>
<th>SentimentRecord.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;testing this out.... happy&quot;, &quot;happy&quot;, 1</td>
</tr>
<tr>
<td>&quot;Wow, great job, <a href="mailto:test@test.com">test@test.com</a>&quot;, &quot;wow&quot;, 1</td>
</tr>
<tr>
<td>&quot;Wow, great job, <a href="mailto:test@test.com">test@test.com</a>&quot;, &quot;great&quot;, 1</td>
</tr>
<tr>
<td>&quot;A bad thing happen&quot;, &quot;bad&quot;, -1</td>
</tr>
<tr>
<td>&quot;good better bad&quot;, &quot;good&quot;, 1</td>
</tr>
<tr>
<td>&quot;good better bad&quot;, &quot;better&quot;, 1</td>
</tr>
<tr>
<td>&quot;good better bad&quot;, &quot;bad&quot;, -1</td>
</tr>
<tr>
<td>&quot;bad worse good&quot;, &quot;good&quot;, 1</td>
</tr>
<tr>
<td>&quot;bad worse good&quot;, &quot;bad&quot;, -1</td>
</tr>
<tr>
<td>&quot;bad worse good&quot;, &quot;worse&quot;, -1</td>
</tr>
<tr>
<td>&quot;good better&quot;, &quot;good&quot;, 1</td>
</tr>
<tr>
<td>&quot;bad worse&quot;, &quot;bad&quot;, -1</td>
</tr>
<tr>
<td>&quot;bad worse&quot;, &quot;worse&quot;, -1</td>
</tr>
</tbody>
</table>

8. Generate the Yammer sentiment message by two steps: first, integrate SentimentScore (body, sentiment score) with OriginalMessage (id, thread_id, sender_id, sender_name, body, created_at) to form SentimentMessage1 (id, thread_id, sender_id, sender_name, body, created_at, sentiment score); second, integrate Time(created_at1, created_time) with SentimentMessage1 (id, thread_id, sender_id, sender_name, body, created_at, sentiment_score) to form the final output SentimentMessage(id, thread_id, sender_id, sender_name, body, created_time, sentiment_score)
4.3 Building the Models

Based on the Streams SPL application process, the model flow as bellow:
4.4 Assessing the Model

Based on this Stream SPL model, it can deal with yammer sentiment analysis without any errors and can complete the process with huge big data in limited time, which can accomplish our goals successfully as business understanding part mentioned.

5. Evaluation

5.1 Evaluating the results

Capture 100 latest sentiment message records and read the body content one by one to judge whether it is positive, negative, or natural. Based on AFINN sentiment dictionary, there are 25 records’ actual sentiment doesn’t match with the calculated sentiment score. Thus, this SPL sentiment analysis model can perform well in judge sentiment from yammer message, and the precision of AFINN sentiment dictionary is about 75%.

5.2 Refine the sentiment analytics

Sentiment analysis rely, at varying degrees, on lists of words and phrases with positive and negative connotations or are empirically related to positive or negative comments. In order to better understanding the employee’s sentiment, it is necessary to self-build a sentiment dictionary based on Yammer message posted by employees own.

1. Build a C# word count application to token the text data of body attribute from original message:
   - Split all the sentence and phrases in to single words
   - Calculate the occurrence for each unique word.

2. Build a Nationwide Yammer Sentiment Dictionary based on the word count result:
   - The C# application generates 68502 unique words, numbers, and other characters
   - Delete number and incorrect characters. There are 56928 left.
   - Rank left 56928 words based on occurrence. Then, choose words more than 10 times. There are 9282 left
   - In these 9282 words, remove stop words, typo words, noise word, and natural words. Then, classify the left words into positive and negative categories. There are 984 words left, including 329 negative words and 655 positive words. Exclude the “joined” and “welcome” in positive list, the positive dictionary contain 653 words, and the negative dictionary contain 329 words

3. Take the same process as Step 4 Modeling to get new output result.

Capture 100 latest sentiment message records and read the body content one by one to judge whether it is positive, negative, or natural. Based on Nationwide yammer sentiment dictionary, there are 20 records’ actual sentiment doesn’t match with the calculated sentiment score. Thus,
the Nationwide yammer sentiment dictionary performance better than AFINN sentiment dictionary in judge employee’s sentiment through yammer message data. The precision of Nationwide yammer sentiment dictionary is about 80%.

4. Comparison with AFINN sentiment dictionary and Nationwide yammer sentiment dictionary

<table>
<thead>
<tr>
<th></th>
<th>AFINN_DIC</th>
<th>without joined &amp; welcome</th>
<th>NW_DIC</th>
<th>without joined &amp; welcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sentiment word message</td>
<td>49194</td>
<td>38244</td>
<td>65099</td>
<td>37832</td>
</tr>
<tr>
<td>No. of positive word message</td>
<td>46932</td>
<td>35653</td>
<td>63219</td>
<td>35779</td>
</tr>
<tr>
<td>No. of negative word message</td>
<td>12182</td>
<td>12182</td>
<td>11025</td>
<td>11025</td>
</tr>
<tr>
<td>No. of without sentiment word message</td>
<td>28510</td>
<td>39460</td>
<td>12605</td>
<td>39872</td>
</tr>
<tr>
<td>Precision</td>
<td>75%</td>
<td></td>
<td>80%</td>
<td></td>
</tr>
</tbody>
</table>

5.3 Determining the application of sentiment message

After get the sentiment messages file based on Nationwide yammer sentiment dictionary, it can be used to process other visualization application results as below:

1. An employee sentiment change graph within specific period

We can select any employee we want to analysis, using excel to draw a line graph as below:

- Employee A’s sentiment change in 2012
2. An employee sentiment score distribution graph within specific period

- Employee A’s sentiment score distribution in 2012

- Comparison of Employee A’s sentiment score distribution in 2010 and 2011
3. An employee sentiment words cloud within specific period

- Employee A’s sentiment words cloud in 2012

![Employee sentiment words cloud](image)

4. All employees’ sentiment change in specific period

- Employee’s sentiment change in July 18th 2012

![Employee's sentiment change](chart)

6. Deployment

6.1 Planning for next steps research

- Refine the dictionary based on new yammer data

The nationwide yammer sentiment dictionary can be updated every specific period, like every month. The updated sentiment dictionary can collect the new and updated sentiment words or phrases that are
not in the current version. And the SPL sentiment analytics model can perform better within updated sentiment dictionary.

- **Build personal own sentiment dictionary**

As we all know, everyone has different propensity to talk about. In order to know more accuracy about employee’s personal sentiment, there is necessary to build the own personal sentiment dictionary for analysis. For example, “KISS” is the highest frequency word in Jeff Schumann word cloud, but “KISS” in Jeff’s yammer is mean a Nationwide's KISS Collaboration Station managed by SPOT team. This word has no sentimental meaning for Jeff.

- **Use HashTag to extract topic from yammer body, find relationship between sentiment and topic**

The yammer body content, there is one kind of words or phrases with tag mark. Using IBM streams text extractor to extract the hashtag information from yammer body, then, find interesting relationship between the tagged content and the sentiment of corresponding yammer message.
Reference:


