GJMS Special Issue for Recent Advances in Mathematical Sciences and Applications-13 GJMS Vol 2. 2 1–9

Mining Usable Customer feedback from Social Networking data for Business Intelligence

Adinarayana Salina[1] and E. Ilavarasan[2]

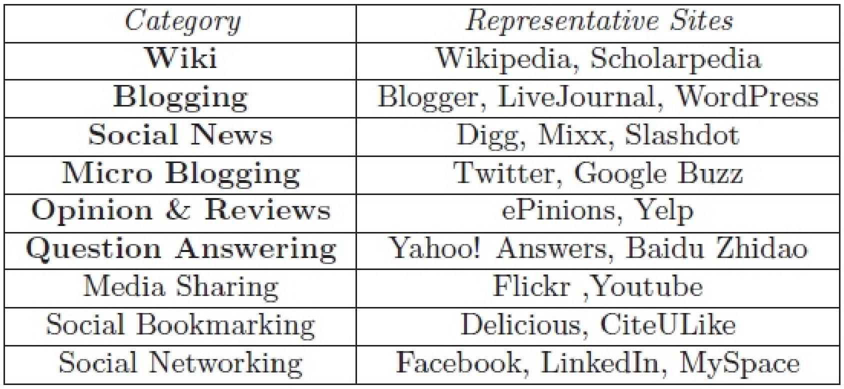
[1]Associate Professor, Dept of IT, Shri Vishnu Engg. College for Women
Bhimavaram, Andhra Pradesh, India. s_adi_2k@yahoo.com
[2] Professor,Dept of CSE, Pondicherry Engineering College
Pondicherry 605 014, India. eilavarasan@pec.edu

ABSTRACT: Business organizations want to use the social media sites (SNS) data to understand the needs and behaviour of their customers or specific targeted groups of individuals with respect to the organizations current or future products or services. SNS frequently discuss consumer products or servicesfrom movies and restaurants to hotels and politics. These shared customers opinions have become highly valuable to businesses and organizations large and small .To increase competitive advantage and effectively assess the competitive environment of businesses, companies need to monitor and analyze not only the data on their own SNS, but also the data of their competitors in different SNS. As a result, a large amount of user-generated unstructured data is freely available on SNS. In this paper we have addressed the application of Text mining methods to harness large amounts of unstructured data and transform it into structured information and then predict variety of customer trends and behaviour which is an important source of input for Business Intelligence to improve the chances of profitability and forecast the future business expansion.

KEYWORDS: SNS,MNC,FI,NLP,IR,KDD,URL,VSM,BOW,DTM,TM,BI.

1 INTRODUCTION

Every internet user in the world especially youngsters, researchers, market analysts and students are more or less give feedback, maintain their social interactions in SNS. With the immense popularity of smart phones recent research indicates that more and more people irrespective of their age, profession are using SNS and online shopping. In this new era of business, essentially anyone can post comments and feedback about companies and their products in SNS,ecommerce websites, which may influence the perceptions and purchase behaviour of a large number of potential buyers. This is of obvious concern to marketing organizations not only is the spread of negative information difficult to control, but it can be very difficult to even detect it in the large space of blogs,forums and SNS. As a result, many companies are adopting SNS to accommodate this growing trend in order to gain customer traffic, increasing customer loyalty and retention, increasing online and offline sales of their products. As the companies have large amounts customer feedback, usually to the point that needles of knowledge cannot be found in the haystack of words. A significant portion of the unstructured feedback collected by companies is in textual format, from SNS, e-mail communication and corporate documents to web pages. Text mining has its academic roots in information retrieval and computational linguistics. Text mining provides the capability of pattern identification, visualization support to aid pattern identification, modelling support to identify or confirm relationships, and drill-down query tools to enable analysts to focus on key problem areas and then focus on related solutions for Business intelligence. The below figure1.1 shows different categories of SNS.



Category	*Representative Sites*
Wiki	Wikipedia, Scholarpedia
Blogging	Blogger, LiveJournal, WordPress
Social News	Digg, Mixx, Slashdot
Micro Blogging	Twitter, Google Buzz
Opinion & Reviews	ePinions, Yelp
Question Answering	Yahoo! Answers, Baidu Zhidao
Media Sharing	Flickr ,Youtube
Social Bookmarking	Delicious, CiteULike
Social Networking	Facebook, LinkedIn, MySpace

Figure 1: Social Media types.

Fig 1.1

Received February 24, 2014 IFNA Publishers, Inc.

From the figure 1.1 we can say that the SNS contain various types of services and thus create different formats of feedback data, in text, image, video formats. For example, the media sharing sites Flickr and Youtube allow to observe people do engaged in the creation and sharing of their personal photography. For example for a recent telugu movie song katama rayuda of pavan kayan starrer attarintiki daredi IIT Kharagpur students had created a video on a dance sequence with a group of friends and uploaded in YouTube and is available at www.youtube.com/watch?v=RsOF1DGL8Y8.This show how SNS influences younger generations. As a result, a large amount of text, image and video data is archived in the sites. Besides, in blogging sites, the users post frequently and create a huge number of textual / text-based data; in social bookmarking sites, ecommerce sites movie reviews, users share with each other tags and URLs. Among the various formats of feedback exchanged in social media, ecommerce sites, text plays an important role. The feedback is mostly stored in text format. For example, micro blogging SNS like FaceBook they allow users to post small amounts of text(SMS language) for communicating breaking news, information sharing, and participating in events. Customers will give feedback on purchased products, seller rating in online shopping. On the other hand, there are also a lot of useful textual data containing in the sites which are concentrating on other domains. For instance, researchers proposed to utilize tag information in multimedia sharing sites to perform video retrieval and community detection.. Under these scenarios, how to mine useful information from textual data presents great opportunities to business intelligence research and applications. Here Text Mining for BI refers to the discovery of usable customer feedback can be found in text archives by applying techniques from Natural Language Processing (NLP),Knowledge Data Discovery(KDD),Information Retrieval(IR) etc.

2 SYSTEM ARCHITECTURE

Data Analytics, Text Analytics and Text Mining are one and the same. The typical phases of the SNS Text Analytics for Business Intelligence is shown in fig 2.1 and they are collecting text corpus from SNS, Pre-process the text corpus, represent[13] the pre processed text with term relationship matrix(DTM),Bag-of-words(BOW)[14], Vector space Model(VSM)and then apply KDD techniques like Classification, clustering and finally extract opinions of the customers expressed in the SNS and transform them towards Business Intelligence.

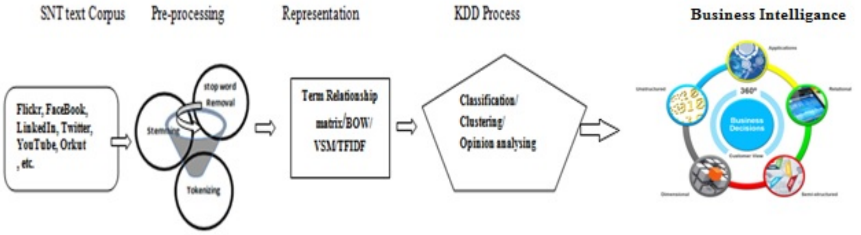


Figure 2: System Architecture

Fig 2.1

In all these aspects pre processing text corpus is difficult process because the text collected from SNS is not structured, in different forms, contains sparseness. In this paper we are addressing this issue and then representing them in any of the representation methods for understanding text. For, pre processing, representation, Visualization of feedback, KDD process and data visualize for Business Intelligence we are using R language tm package. Objective of this work is to filter the vast blogosphere from millions to the thousands of blogs most relevant to the topic of interest. In the simplest case, the topic can be a specific product, and the objective is therefore to identify all blogs discussing this product and perhaps competing products as well.

A typical SNS post or customer feedback consists of multiple entries or posts. It is useful to think of a feedback in a three-dimensional space defined by three dimensions relevance, sentiment and authority. While dimensions relevance and sentiment, are specific to a given post or even smaller section of text, the third dimension authority is topic dependent at the organisation level.

3 METHODOLOGY

Pre-processing text corpus[5,6] aims to make the input documents more consistent to facilitate text representation which is an important task for KDD process. Stop word removal eliminates words using a stop word list, in which the words are considered more general and meaningless. Stemming reduces the diversity of representations of a word to a canonical morphological representation. Pre processing methods for customer Opinion analysis process the message from a syntactical point of view, which requires that the method retains the original sentence structure.. Without this information, it is difficult to distinguish Are you doing right things? and Is your doing things right?, which have overlapping vocabularies. In this case, we need to avoid removing the syntax-containing words. The process of mining text documents involve linguistically and semantically analysis of the plain text, thus structuring the text. Finally relates and induces some hidden traits found in the text, like frequency of use for some words, entity extractions, and documents summarizations. The most common way to model documents is to transform them into term relationship matrix. In this method, the similarity between every term pair is calculated as a basis for determining the clusters. The easiest way to understand this approach is to consider the vector space model. The vector space model is represented by a matrix where the rows are individual items and the columns are the unique words (processing tokens) in the items. The most popular weighting schema is Term Frequency / Inverse Document Frequency (TF-IDF) .This scheme TF-IDF[3] is defined as Given a document collection D, a word w, and an individual document d D, we Calculate wd =$tf \times \lg(|D|/df)$ where tf equals the number of times w appears in d, —D— is the size of the corpus, and df equals the number of documents in which w appears in D and wd is the relative weight of the feature in the vector (Salton & Buckley,1988, Berger, et al, 2000). methods like classification or clustering for KDD. Even we can use NLP based similarity measures for opinion analysis of the customers. The measure increases as similarity grows, As queries are analogous to documents, the same similarity measures can be used to measure doc-doc,doc-query,query-query. A Document or query is treated as n dimensional vector then the most popular cosine similarity is defined as

$$SUM(Q,D) = \sum_{i=1}^{k}(qi \times di)/\sqrt{\sum_{i=1}^{k}(di^2 \times qi^2)} \tag{1}$$

Formula measures the cosine of the angle between the two vectors where q is query, d is document and k is no of document samples .

4 CUSTOMERS OPINION POSTING

Customers preferably internet users will post their opinions[4] like product reviews, feedbacks in online shopping, gives ratings on movie reviews in blogs, upload quotes, self created videos, share ideas in SNS like FaceBook, Twitter. As far as our paper we are concentrating on text data in all these SNS postings. Here our idea is tackle sparseness problem in unstructured text because Certain SNS restrict the length of user-created content such as micro-blogging messages, product reviews, QA passages and image captions, etc .

Twitter SNS allows users to post news quickly and the length of each tweet is limited to 140 characters. Similarly, Picasa comments posting are limited to 512 characters, and personal status messages on Windows Live Messenger are restricted to 128 characters. As we can see, data with a short length is ubiquitous on the web at present. As a result, these short messages have played increasing important roles in applications of social media. Successful processing short texts are essential to data analytics methods.

We have visited, analysed various SNS postings, ecommerce websites product reviews, feedback to review customer opinions. When composing a message, users may follow old adage keep it short and simple (KISS) for instant communication so that they will use or coin new abbreviations or acronyms and symbols that seldom appear in conventional text documents. In this paper we have taken ebay customer feedback as text corpus for implementing text mining. The SNS posts or online shopping feedback have KISS format typical words . For example gud9t for good night, k for ok, tc for take care, yup for yes, nope for no ,5n for fine, clz for college etc. actually these are not English words but they are very intuitive and popular in social media conversations. Short messages, as the most important data format, make people more efficient with their participate in social media applications, however this poses challenges for researchers doing research in text mining ,opinion mining especially in text clustering text classification and sentiment polarity analysis.

5 DETECTING EMERGING TOPICS OF INTREST

Relevance, authority and sentiment[16] provide a useful way for us to focus attention on the SNS posts that we should mined through Text mining. we naturally synthesize this data to identify trends that summarize the feedback. This is perhaps the ultimate objective of Business intelligence. At an even higher level of analysis, Text mining techniques like document clustering is used to identify collections of posts expressing cohesive patterns of discussion. These techniques can be extended with temporal continuity to provide a view of how dominant themes evolved over time, possibly also incorporating feedback from a user on which themes to track or discard from the analysis. We have shown the document clustering process with dendograms created in R.

6 EXPERIMENT RESULTS

To prepare Corpus I have used R language[7]and taken 17 ebay feedback reviews data from http://community.ebay.in/t5/Feedback-to-India-Team/Customer-s-Review/td-p/60605 as input. I have selected first four reviews data contents Mycorpus is created with all these data contents as

$$> mycorpus < -Corpus(VectorSource(c(d1, d2, d3, d4)))$$

The status of the created corpus is by

$$> mycorpus$$

A corpus with 4 text documents. Now to pre-process mycorpus we have used the following code

$$> skipWords < -function(x)removeWords(x, stopwords(''english''))$$

$$> funcs < -list(tolower, removePunctuation, removeNumbers, stripWhitespace, skipWords)$$

$$> mycorpus.proc < -tm_{m}ap(mycorpus, FUN = tm_{r}educe, tmFuns = funcs)$$

Representing the processed corpus to term relationship matrix[7]

$$> mycorpus.dtm < -DocumentTermMatrix(mycorpus.proc, control = list(wordLengths = c(3, 10)))$$

Finally the document term relationship matrix is displayed using

$$> inspect(mycorpus.dtm)$$

the resultant is as shown below in fig 6.1:

```
A document-term matrix (4 documents, 124 terms)
Non-/sparse entries: 142/354
Sparsity           : 71%
Maximal term length: 10
Weighting          : term frequency (tf)
    Terms
Docs additional affect agree also always based big bought buy buyer buyers
   1          0      0     0    0      0     1   0      2   1     0      1
   2          0      1     0    0      1     0   0      0   0     2      0
   3          0      0     0    1      0     0   0      0   1     0      0
   4          1      0     1    0      0     0   1      0   1     0      3
    Terms
Docs buying came can checking click come conclude conclusion confident
   1      0    1   2        1     0    0        1          0         0
   2      0    0   1        0     0    0        0          0         1
   3      0    0   0        0     0    0        0          0         0
   4      1    0   1        0     1    1        0          1         0
    Terms
Docs continuous customer customers decide depend eager ebay ebaybut fact far
   1          0        1         2      0      0     0    0       0    0   1
   2          0        0         0      1      1     1    5       0    0   0
   3          0        0         0      0      0     0    0       1    0   0
   4          1        0         0      0      0     0    3       0    1   0
    Terms
Docs favors fear feature feedback feedbacks finally find flipkart follow forum
   1      0    0       0        2         0       0    0        0      0     0
   2      0    1       1        0         0       0    1        1      0     1
   3      0    0       0        0         0       0    0        0      0     0
   4      1    0       1        3         4       1    0        0      1     0
    Terms
Docs give good google great help idea introduce issue ive kind know launch
   1    0    0      0     1    1    1         0     0   0    0    1      0
   2    1    1      1     0    0    0         1     1   1    0    0      0
   3    0    0      0     0    0    0         0     0   0    0    0      0
   4    0    1      0     0    0    0         0     0   0    1    0      1
    Terms
Docs leave listing luckily major need negative neither never now obvious one
   1     0       0       0     0    0        0       0     0   0       0   0
   2     0       0       1     1    1        1       2     0   1       1   1
   3     0       0       0     0    0        0       0     1   0       0   0
   4     1       1       0     0    0        0       0     0   0       0   0
    Terms
Docs online particular people pethachi philips plays please positive ppl
   1      0          0      0        0       1     0      0        0   0
   2      1          0      1        0       0     1      0        1   0
   3      0          0      0        0       0     0      0        0   0
   4      0          1      0        1       0     0      1        0   1
    Terms
Docs problems process product products provided providing quality quite quo
   1        0       0       4        0        0         0       0     0   0
   2        0       1       5        2        1         1       2     1   1
   3        2       0       0        1        0         0       0     0   0
   4        0       0       2         0        0         0       0     0   0
    Terms
Docs raising reason receive received regards remote request review reviews role
   1       0      0       0        0       0      1       0      1       0    0
   2       1      1       1        0       0      0       0      2       3    1
   3       0      0       0        0       0      0       0      0       0    0
   4       0      0       0        1       1      0       1      0       0    0
    Terms
Docs sales section see seems seller sellers selling share shopping showing site
   1     0       1   0     0      0       0       0     1        0       0    0
   2     1       1   0     2      1       1       0     0        1       0    1
   3     0       0   0     0      1       0       1     0        0       0    0
   4     0       0   1     0      1       2       0     0        0       1    0
    Terms
Docs sites small sold sometime status system the though totally traders treat
   1     0     0    0        0      0      0   0      0       0       0     0
   2     1     0    1        1      1      0   1      1       1       0     0
   3     0     0    0        0      0      0   0      0       0       0     0
   4     0     1    0        0      0      2   0      0       0       1     1
    Terms
Docs uniformly universal ups use value want wanted well whether will works
   1         0         1   0   0     0    1      1    1       0    2     1
   2         0         0   0   1     0    0      0    0       1    1     0
   3         0         0   0   0     0    0      0    0       0    0     0
   4         1         0   1   0     1    0      0    0       0    0     0
```

Fig 6.1

To show the application of Stemming[10,12], we have taken the first feedback of this document and implemented stemming[12] with the following code and results shown in fig 6.2

$$> worder1 < -c(d1)$$

$$> df1 < -data.frame(id = 1 : 5, words = worder1)$$

```
df1
  id
1  1
2  2
3  3
4  4
5  5
```

```
words

 1. It will be great if we can have a customer review section
    on each product so that Customers who bought that product can
    share the feedback about the product which will help other buyers
    to conclude if they want to buy that product based on other
    customer's feedback.\n\nIdea came to me because I was checking
    a universal remote from Philips and wanted to know if it works well
    for who bought it so far.
 2. It will be great if we can have a customer review section
    on each product so that Customers who bought that product can
    share the feedback about the product which will help other buyers
    to conclude if they want to buy that product based on other
    customer's feedback.\n\nIdea came to me because I was checking
    a universal remote from Philips and wanted to know if it works
     well for who bought it so far.
 3. It will be great if we can have a customer review section
    on each product so that Customers who bought that product can
    share the feedback about the product which will help other
    buyers to conclude if they want to buy that product based on
    other customer's feedback.\n\nIdea came to me because I was
    checking a universal remote from Philips and wanted to know
    if it works well for who bought it so far.
 4. It will be great if we can have a customer review section
    on each product so that Customers who bought that product can
    share the feedback about the product which will help other
    buyers to conclude if they want to buy that product based on
    other customer’s feedback.\n\nIdea came to me because I was
    checking a universal remote from Philips and wanted to know
    if it works well for who bought it so far.
 5. It will be great if we can have a customer review
    section on each product so that Customers who bought that
    product can share the feedback about the product which will
    help other buyers to conclude if they want to buy that product
    based on other customer’s feedback.\n\nIdea came to me because
    I was checking a universal remote from Philips and wanted to
    know if it works well for who bought it so far.
```

$$corp1 < -Corpus(VectorSource(df1words))$$

$$> inspect(corp1)$$

```
A corpus with 5 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID
[[1]]It will be great if we can have a customer review section
     on each product so that Customers who  bought that product
     can share the feedback about the product which will help
     other buyers to conclude if they want to buy that product
     based on other customer’s feedback.
     Idea came to me because I was checking a universal remote
     from Philips and wanted to know  if it works well for who
     bought it so far.
[[2]]It will be great if we can have a customer review section
     on each product so that Customers who  bought that product
     can share the feedback about the product which will help
     other buyers to  conclude if they want to buy that product
     based on other customer’s feedback.
     Idea came to me because I was checking a universal remote
     from Philips and wanted to know  if it works well for who
     bought it so far.
[[3]]It will be great if we can have a customer review section
     on each product so that Customers  who bought that product
     can share the feedback about the product which will help
     other buyers to conclude if they want to buy that product
     based on other customer’s feedback.
     Idea came to me because I was checking a universal remote
     from Philips and wanted to  know if it works well for who
     bought it so far.
[[4]]It will be great if we can have a customer review section
     on each product so that Customers who bought that product
     can share the feedback about the product which will help
     other buyers to conclude if they want to buy that product
     based on other  customer’s feedback.
     Idea came to me because I was checking a universal remote
     from Philips and wanted  to know if it works well for who
     bought it so far.
```

```
[[5]]It will be great if we can have a customer review section
     on each product so that Customers who bought that product
     can share the feedback about the product  which will help
     other buyers to conclude if they want to buy that product
     based on other customer's feedback.
     Idea came to me because I was checking a universal remote
     from Philips and  wanted to know if it works well for who
     bought it so far.
```

Removal of stop-words by:

$$corpus1 < -tm_{m}ap(corpus1, removeWords, stopwords(''english''))$$

Stemming[8] is done by:

$$tm_{m}ap(corpus1, stemDocument)$$

create the term-document matrix for corpus1

$$dtm < -DocumentTermMatrix(corpus1)$$

$$inspect(dtm[1:5, 100:105])$$

```
A document-term matrix (5 documents, 6 terms)

Non-/sparse entries: 6/24
Sparsity           : 80%
Maximal term length: 13
Weighting          : term frequency (tf)

    Terms
Docs competitive competitors concentrating consider considered consistent
   1                3                     1                     0            0          0
   2                0                     0                     0            0          0
   3                0                     0                     1            0          0
   4                0                     0                     0            0          0
   5                0                     0                     0            1          1
```

Fig 6.2

we want to find those terms that occur at least five times,then we can use the findFreqTerms() function[13]

$$dtm < -DocumentTermMatrix(mycorpus)$$

$$> findFreqTerms(dtm, 5)$$

```
[1] "about"   "and"     "ebay"    "for"     "product" "that"    "the"
[8] "they"    "this"
```

After applying cluster analysis[9,15] and obtained dendograms and graphs we can analyse the different set of data posted in SNS for effective feedback analysis. Result Analysis with Dendograms applying cluster analysis for mycorpus by considering mycorpus.dtm as input.

$$> mydata < -na.omit(mycorpus.dtm)$$

$$> mydata < -scale(mydata)$$

Ward Hierarchical Clustering[9]and we can observe the dendograms for analysis

$$. > d < -dist(mydata, method = ''euclidean'')$$

distance matrix

$$fit < -hclust(d, method = ''ward'')$$

to display the dendograms

$$> plot(fit)$$

7 VISUALIZING REVIEW ANALYSIS

Analyzed the mined feedback (reviews) after preprocessing with the text mining techniques in R and Rcommander[6]. The analysis of top 5 and worst 3 features of different ipods are shown in Fig 7.1

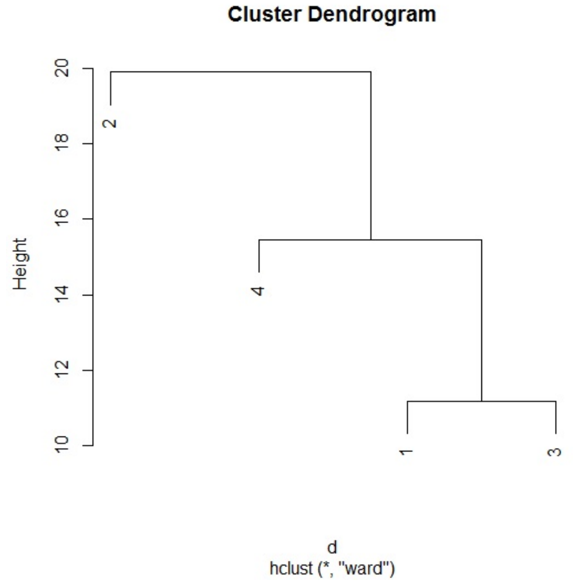


Fig 6.3 cluster dendogram.

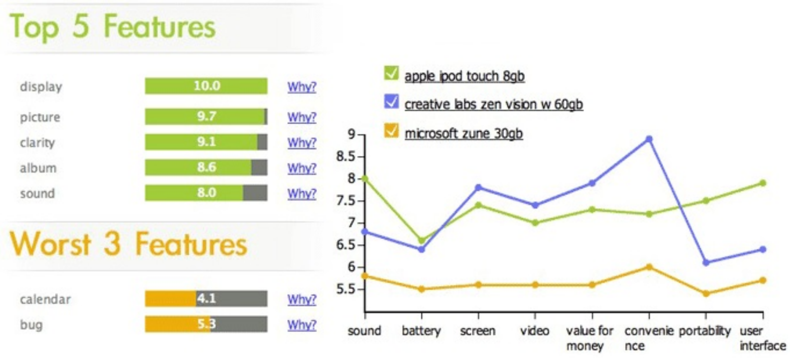


Fig 7.1 analysis of top 5 and worst 3 features of different ipods .

8 CONCLUSION AND FUTURE ENHANCEMENT

As of now I have reviewed existing clustering, classification techniques [8] and applied them on mycorpus how ever because of page limitation we are not including them. So far my investigation on pre-processing the text corpus for removing the sparseness and structuring the unstructured text is major challenge and we have handled effectively. As a future scope of this paper we want to extend the existing clustering and classification techniques, apply Business intelligence software Jasper-soft on Text mined data for a better results. We also identified a new research area topic modelling to identify collections of posts expressing cohesive patterns of discussion to observe business trends .This will be discussed in our subsequent papers.

ACKNOWLEDGMENT

We would like to thank all our colleagues who encouraged and helped in completing this work and also for r-project for its help on R.

REFERENCES

[1] htdp://gn.wikipetia.ore/wiki/Data_analysis.

[2] Mirco Speretta and Susan Gauch,-Using Text Mining to Enrich the Vocabulary of Domain Ontologies.

[3] Marko Grobelnik, Dunja MladenicJ. Stefan Institute, Slovenia,Text Mining tutorial.

[4] Bo Pang and Lillian Lee- Opinion Mining and sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 2, Nos. 12 (2008) 1135.

[5] Processing text, projects.iq.harvard.edu/undergradscholars/book/export/html/6585.

[6] Introduction to the tm Package Text Mining in R ,cran.r-project.org/web/packages/tm/vignettes/tm.pdf.

[7] Text Mining with R,

$$www.zinkov.com/posts/2010-10-21-slides_from.../tm_slides.pdf.$$

[8] Michael W. Berry and Malu Castellanos,Survey of text mining: Clustering, Classification, and Retrieval, Springer Second Edition

[9] Distances between Clustering, HierarchicalClustering, http://www.stat.cmu.edu/ cshalizi/350/lectures/08/lecture-08.pdf

[10] Stemming,en.wikipedia.org/wiki/Stemming.

[11] Durga Shankar Baggam, Salina Adinarayana, Raghuraman Koteeswaran and Yagnesh Balasubramanian, SNS Registration Coggling Social Capitals, European Journal of Scientific Research,Volume83,Issue1, August, 2012.

[12] Word Stemming in R,Duncan Temple Lang Department of Statistics,UC Davis,August 4, 2004.

[13] Basic-text-process,

$$http://pages.cs.wisc.edu/\ jerryzhu/cs769/text_preprocessing.pdf$$

[14] Guy Lebanon, Yi Mao, Joshua Dillon- The Locally Weighted Bag of Words Framework for Document Representation, Journal of Machine Learning Research 8 (2007) 2405-2441

[15] Brandon M. Stewart , Practical Skills for Document Clustering in R, faculty.washington.edu/jwilker/tft/Stewart.LabHandout.pdf

[16] Word Stemming in R,Duncan Temple Lang Department of Statistics,UC Davis,August 4, 2004.

[17] Salina Adinaryana ,Detecting Identification Anomalies in Social Networking with Cluster based re-ranking and Slink Algorithms, International Journal of Modern Engineering Research (IJMER),Vol.2, Issue.4, July-Aug. 2012 pp-2839-2842; ISSN: 2249-6645