Data Analytics Framework and Methodology for WhatsApp Chats

Transliteration of Thanglish and Short WhatsApp Messages

P. Sudhandradevi
Department of Computer Applications
Bharathiar University
Coimbatore, India
psudhandradevi@gmail.com

V. Bhuvaneswari
Department of Computer Applications
Bharathiar University
Coimbatore, India
bhuvanes_v@gmail.com

***Abstract*- Data Analytics has emerged as an important domain in the digital space due to the explosion of tremendous volume of data by various sources such as social media, sensors and business organizations. Social media contribute in generation of huge varied data formats in various representations. WhatsApp has attracted large volume of users because of the easy chat conversations. The chat conversations in WhatsApp support multi languages where user has made their own short conventions in representing communications. In current scenario WhatsApp is used for small scale business, understanding the context in this chat text is important to identify the insights. WhatsApp data is left behind unnoticed as there exist no standard to represent in conventional machine understandable text. The objective of this paper is to design a methodology and framework for transliteration of WhatsApp Chat short messages and thanglish (English and Tamil language combined) text. A MapReduce framework is proposed for the transliteration process. The dataset is acquired from known WhatsApp group. A corpus is created from the framework and the experimental results were found to be interesting. The frequent terms are visualized using word cloud. Around 635 WhatsApp texts are replaced by English words.**

Keywords: MapReduce, Text Preprocessing, WhatsApp, Data Analytics, Word Cloud

1. INTRODUCTION

Social Media has paved way by revolving data which made Big Data to evolve. Data volume is exploding at an exponential rate since past two years particularly through social media. It is reported as follows that in August 2015, over 1 Billion people used Facebook in a single day, within five years there will be 50 Billion smart devices is expert to connect the global.73% of organizations have already invested in Big Data by 2016. (www.statista.com/number-of-monthly-active-whatsapp-users)

In 2020 it also predicted that 1. 7 MB of new information will be created by human by every second in digital through social media. Social media users are found to be highly increasing using Facebook, WhatsApp, Twitter and Instagram. India ranks in the first position in the globe in social media usage.Facebook and messenger chat is maintained by Facebook is excited large number of users of all levels due to the simple chat messaging service. WhatsApp is available in all regional languages. It transfers chat data inform of text, audio, video. Most of the social media data comes under the categories of unstructured data. In WhatsApp most of the chat based on short text message, which is left annotation, widely used in small scale retailer to run their business to communicate through messaging text.

Analysis of the WhatsApp text messaging is evolved as requires attention. The computation Analysis that faces with challenges, such as WhatsApp Messenger is evolved as one of the important communication in social media maintained by Facebook. WhatsApp is an attracting social media chat preferred by large number of users. More than 1 Billion people over 180 countries use WhatsApp to stay in touch with others (www.whatsapp.com). WhatsApp data comes under the category of unstructured data. WhatsApp messenger doesn’t have any standard convention of the text. As it comes in short message the chat become difficult to process using algorithm. Process of WhatsApp data or specific require comes the context of WhatsApp message into common English text. So data processing analysis requires machine understandable text notations for understanding the context.

The objective of the paper is proposed a methodology to convert WhatsApp understandable English text. This paper also provides the methodology for thanglish to English convention. The sections are organized as follows: Section 2 discusses with methodology and framework, Section 3 provides with a detailed view on Framework of MapReduce. In Section 4 the experimental results are discussed followed by Conclusion in Section 5.

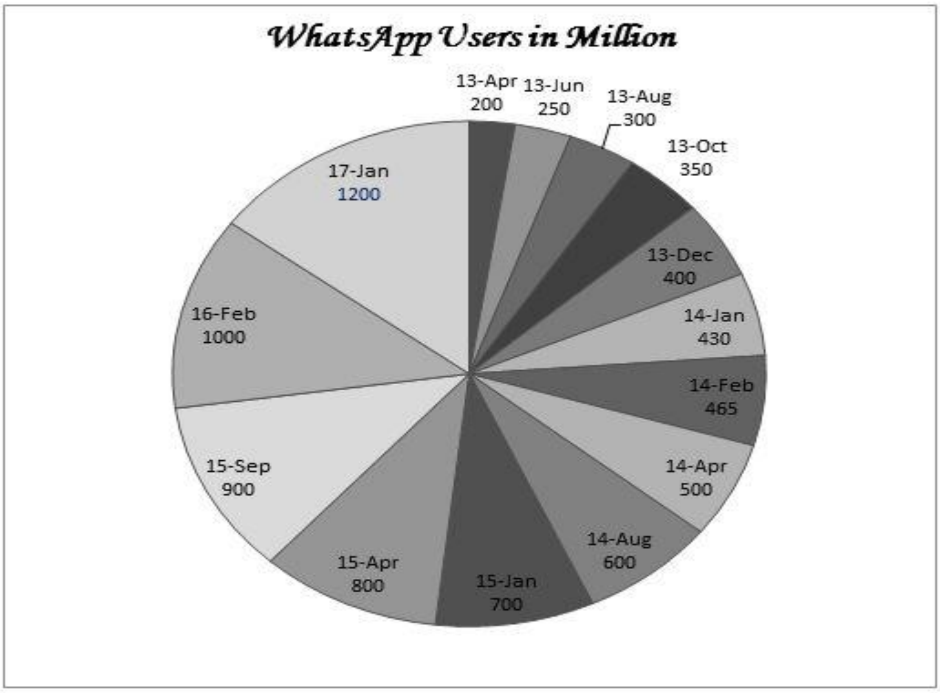


Figure 1. Number of WhatsApp user year wise Statistics (www.statista.com/number-of-monthly-active-whatsapp-users)

2. METHODOLOGY AND FRAMEWORK

In this framework we have designed a messaging chat for WhatsApp data for transliteration of WhatsApp chat text into original English text as given in Figure2. The framework consists of four phases.

- Data Acquisition and Text Pre-Processing
- Creation of WhatsApp Chat Corpus
- Semantic MapReduce Framework for WhatsApp and Visualization

2.1 Data Acquisition

WhatsApp data is collected from WhatsApp Messenger which is in bi-lingual language. 100 WhatsApp data is collected from different WhatsApp groups. The data is collected and saved as text file with the following attributes. This attributes are sending time, data, and sender name followed by chats.

2.2 Text Pre-Processing

WhatsApp data is irregular in formats with ambiguities and inconsistency information. So WhatsApp text composed of formats consists of punctuation marks, upper case, lower cases. The text preprocess is carried to remove the opacities using text preprocessing techniques. WhatsApp users use punctuation marks, upper case, lower cases and short forms of the text mentioned in the Figure3.

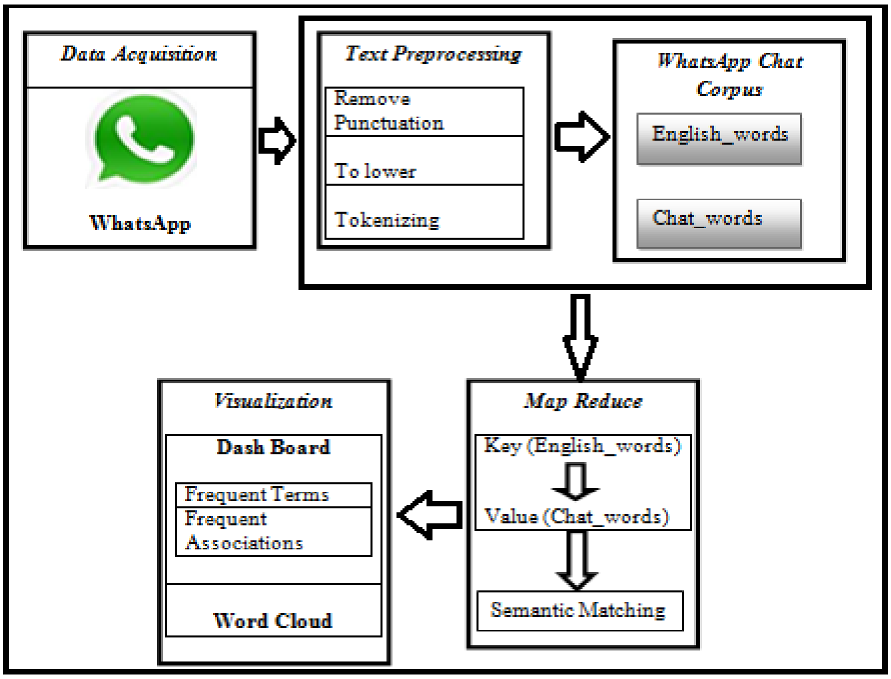
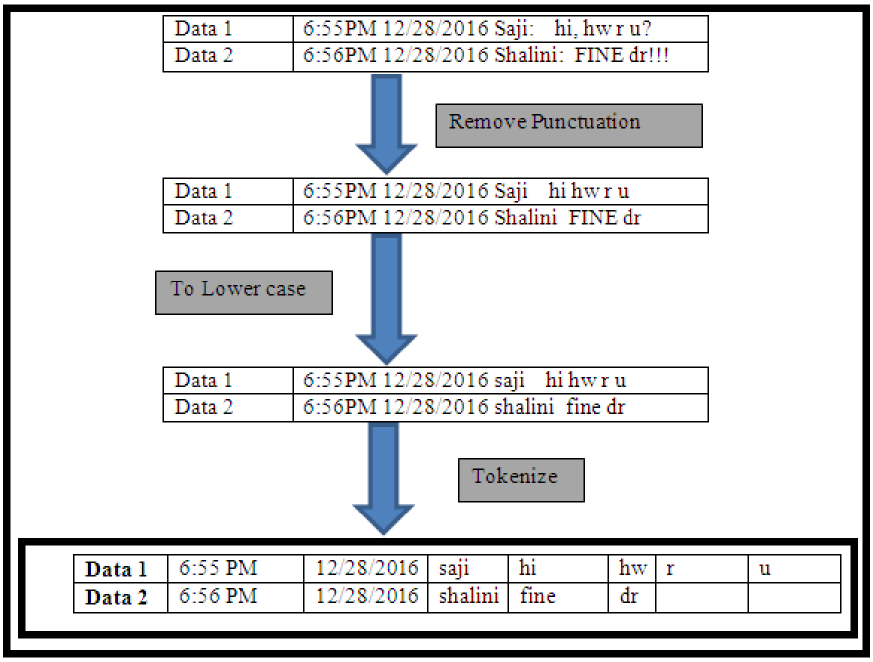


Figure2. Framework and Methodology for WhatsApp Chat Transliteration



Data 1	6:55PM 12/28/2016 Saji: hi, hw r u?
Data 2	6:56PM 12/28/2016 Shalini: FINE dr!!!

Data 1	6:55PM 12/28/2016 Saji hi hw r u
Data 2	6:56PM 12/28/2016 Shalini FINE dr

Data 1	6:55PM 12/28/2016 saji hi hw r u
Data 2	6:56PM 12/28/2016 shalini fine dr

Data 1	6:55 PM	12/28/2016	saji	hi	hw	r	u
Data 2	6:56 PM	12/28/2016	shalini	fine	dr		

Figure3. Text Preprocessing

The following preprocessing is carried out given below. (www.mjdenny.com/Text_Processing_R)

a) Removing Punctuation: WhatsApp message contains punctuation marks. So they are removed to reduce the memory space.

b) Converting to lowercase: WhatsApp users use variety of text cases in chat text typed in upper case, lower case, toggle case and sentence case, which varies for every user. So they are converted into lower case to normalize in a common form.

c) Tokenization: The WhatsApp data is tokenized and split into individual chat words as given in

2.3 Creation of WhatsApp Chat Corpus Model

The WhatsApp chat corpus is created by mapping the original key values with different values. The WhatsApp data consist of both monolingual (Original English Words) and multi lingual languages (Combination of Thanglish and regional Languages. This corpus is modeled to replace English short words and Thanglish short tokens with the original English texts. WhatsApp Chat corpus model is described in Figure 4.

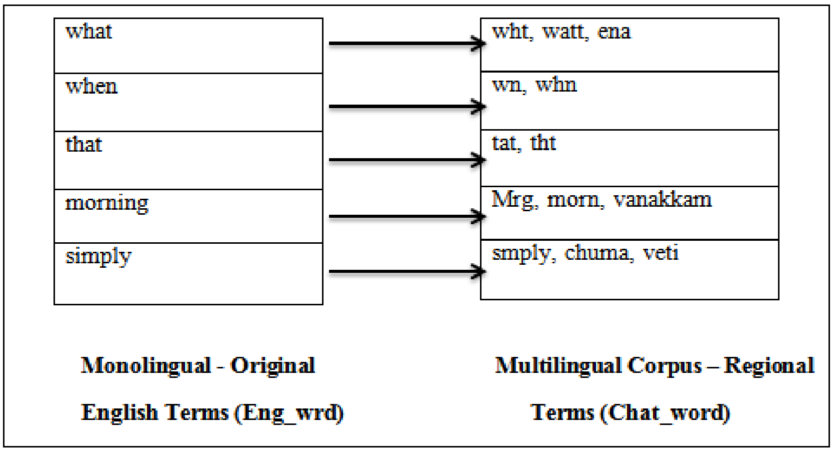

what
wht, watt, ena
when
wn, whn
that
tat, tht
morning
Mrg, morn, vanakkam
simply
smply, chuma, veti
Monolingual - Original English Terms (Eng_wrd)
Multilingual Corpus – Regional Terms (Chat_word)

Figure4. WhatsApp Chat Corpus Model

2.4 Semantic MapReduce Framework for WhatsApp Data

The corpus Eng_wrd and Chat_word consists of 600 objects. In Eng_wrd collection it has the key pair and in Chat_word collection it consists of value pairs. The key and value pairs are mapped using MapReduce task. The mapper function map serves as an input to the reducer function.The reducer function unites all the value pairs with the key pair and then it create sorted list. It typically compresses the list of values to creator the shorter list. Figure5 shows the semantic MapReduce framework for WhatsApp.

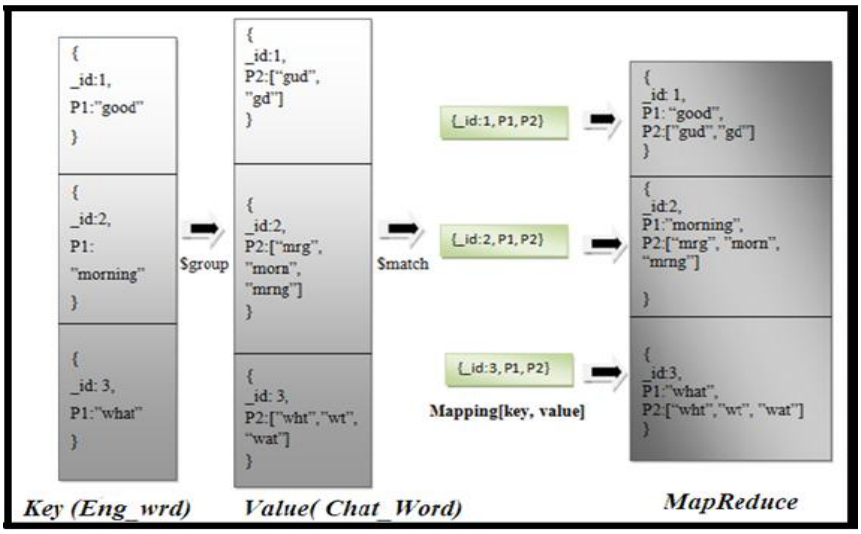

{ _id:1, P1:"good" }
{ _id:2, P1: "morning" }
{ _id: 3, P1:"what" }
$group
{ _id:1, P2:["gud", "gd"] }
{ _id:2, P2:["mrg", "morn", "mrng"] }
{ _id: 3, P2:["wht","wt", "wat"] }
$match
{_id:1, P1, P2}
{_id:2, P1, P2}
{_id:3, P1, P2}
Mapping[key, value]
{ _id: 1, P1: "good", P2:["gud","gd"] }
{ _id:2, P1:"morning", P2:["mrg", "morn", "mrng"] }
{ _id:3, P1:"what", P2:["wht","wt", "wat"] }
Key (Eng_wrd)
Value(Chat_Word)
MapReduce

Figure 5. Semantic MapReduce framework for WhatsApp Data

2.5 Visualization

Visualization of data help people understands the significant of data by placing it in a visual context. Word cloud is a graphical representation of frequently used words in a collection of text files. Tokenized WhatsApp data are visualized using Word Cloud. In word cloud maximum frequent words are plotted and less frequent words are dropped. (www.introduction-to-text-mining-using-R) (www.analyticsvidhya.com)

3. MAPREDUCE FRAMEWORK

This transliteration of WhatsApp chat is implemented using MapReduce framework. The transliteration of chat replacement is modeled as mapper and reducer function for MapReduce programming using R.

3.1 MapReduce Framework View

MapReduce is a process of splitting large files into blocks of equal size, which are distributed across the cluster for storage. Because you always need to consider the failure of the computer in a larger cluster, each block is stored multiple times (typically three times) on different computers. In the implementation of MapReduce, the users apply an alternating succession of map and reduce functions to the data. Parallel execution of these functions, and the difficulties that occur in the process, are handled automatically by the framework. (www.tutorialspoint.com/Mapreduce) Iteration comprises two phases: map and reduce. The MapReduce framework breaks down data processing into

- Map phase
- Reduce phase

3.1.1 Map Phase

The map phase applies the map function to all input. It matches the key with all value pairs. The actual map function is called individually for each of these pairs and emits with a key-value pair.

3.1.2 Reduce Phase

The reducer finally collates all the pairs with the same key and creates a sorted list to the values. The key and the sorted list of values provides the input for the reduce function. The reduce function typically compresses the list of values to create a shorter list. Example: by aggregating the values. Commonly, it returns a single value as its output. Generally speaking, the reduce function creates an arbitrarily large list of key-value pairs, just like the map function.

3.2 Semantic MapReduce Framework for WhatsApp

The corpus Eng_wrd and Chat_word consists of 600 objects. In Eng_wrd collection it has the key pair and in Chat_word collection it consists of value pairs.Map the key and value pairs using MapReduce task. The mapper function map serves as the input to the reducer function.The reducer function unites all the value pairs with the key pair and then it create sorted list. It typically compresses the list of values to creator the shorter list. Figure 6 shows the semantic MapReduce framework for WhatsApp.

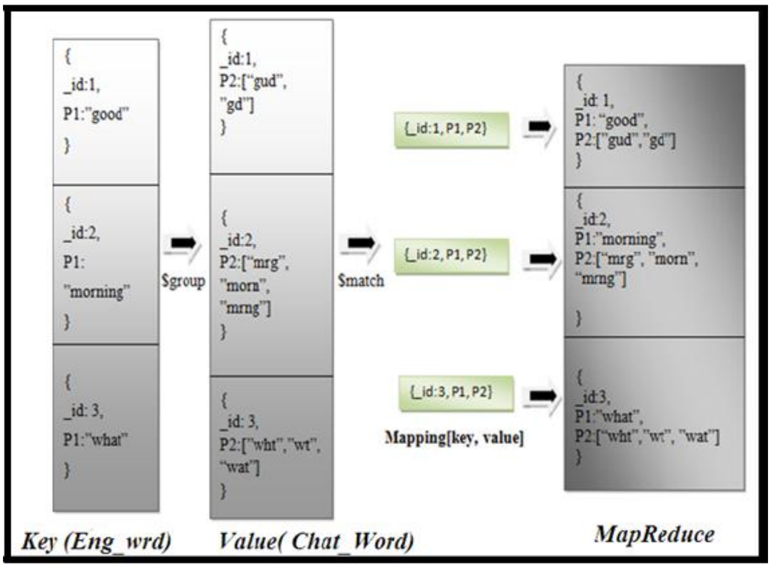

{ _id:1, P1:"good" }
{ _id:2, P1: "morning" }
{ _id: 3, P1:"what" }
$group
{ _id:1, P2:["gud", "gd"] }
{ _id:2, P2:["mrg", "morn", "mrng"] }
{ _id: 3, P2:["wht","wt", "wat"] }
$match
{_id:1, P1, P2}
{_id:2, P1, P2}
{_id:3, P1, P2}
Mapping[key, value]
{ _id: 1, P1: "good", P2:["gud","gd"] }
{ _id:2, P1:"morning", P2:["mrg", "morn", "mrng"] }
{ _id:3, P1:"what", P2:["wht","wt", "wat"] }
Key (Eng_wrd)
Value(Chat_Word)
MapReduce

Figure 6. MapReduce framework for WhatsApp Data

4. RESULTS AND DISCUSSION

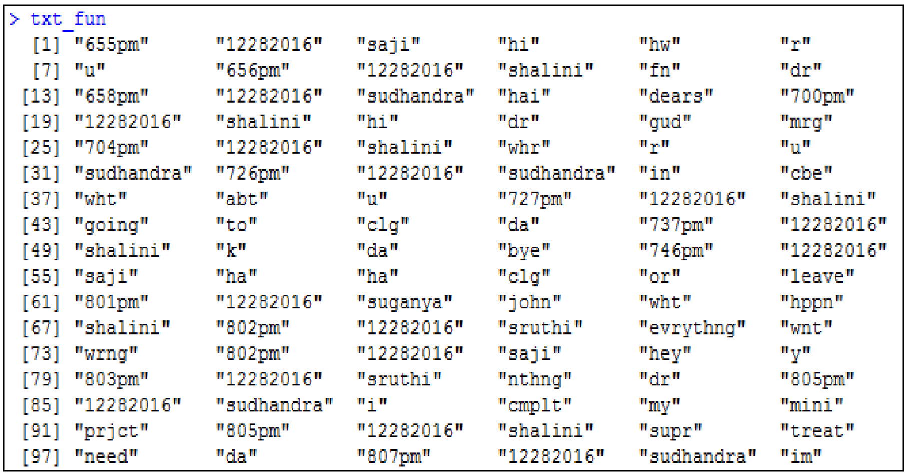
The implementation and results for replacing the WhatsApp Chat Transliteration using the framework for WhatsApp Data is given below. The sample data is shown in Table 1.

Table: 1 Sample WhatsApp Data

Time	Date	Sender Name	Chats
6:55PM	12/28/2016	Saji	Hi, hw r u ?
6:55PM	12/28/2016	Shalini	Fn dr !!
6:58PM	12/28/2016	Sudhandra	Wht dng?

4.1 Experimental Result

The WhatsApp dataset is analyzed for WhatsApp Chat Transliteration of Thanglish and Short WhatsApp Messages by varying regional language. The results are presented and discussed below. The WhatsApp data after text preprocessing is shown in Figure7.

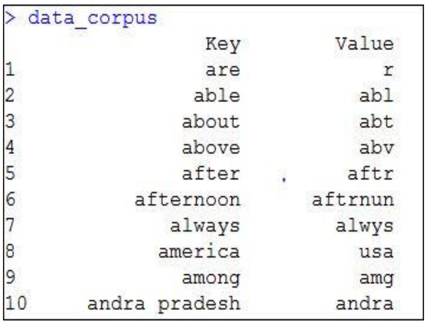


```
> txt_fun
  [1] "655pm"     "12282016"  "saji"      "hi"        "hw"        "r"
  [7] "u"         "656pm"     "12282016"  "shalini"   "fn"        "dr"
 [13] "658pm"     "12282016"  "sudhandra" "hai"       "dears"     "700pm"
 [19] "12282016"  "shalini"   "hi"        "dr"        "gud"       "mrg"
 [25] "704pm"     "12282016"  "shalini"   "whr"       "r"         "u"
 [31] "sudhandra" "726pm"     "12282016"  "sudhandra" "in"        "cbe"
 [37] "wht"       "abt"       "u"         "727pm"     "12282016"  "shalini"
 [43] "going"     "to"        "clg"       "da"        "737pm"     "12282016"
 [49] "shalini"   "k"         "da"        "bye"       "746pm"     "12282016"
 [55] "saji"      "ha"        "ha"        "clg"       "or"        "leave"
 [61] "801pm"     "12282016"  "suganya"   "john"      "wht"       "hppn"
 [67] "shalini"   "802pm"     "12282016"  "sruthi"    "evrythng"  "wnt"
 [73] "wrng"      "802pm"     "12282016"  "saji"      "hey"       "y"
 [79] "803pm"     "12282016"  "sruthi"    "nthng"     "dr"        "805pm"
 [85] "12282016"  "sudhandra" "i"         "cmplt"     "my"        "mini"
 [91] "prjct"     "805pm"     "12282016"  "shalini"   "supr"      "treat"
 [97] "need"      "da"        "807pm"     "12282016"  "sudhandra" "im"
```

Figure7.Text Preprocessing of WhatsApp Data

4.2 WhatsApp Chat Corpus Model

The word corpus is created using Map Reduce framework by associate each chat word as key mapping to corresponding value pairs. Key and Value for corpus is shown in Figure 8. The MapReduce function based on Key, value pairs of the corpus replace the WhatsApp short message chat to the conventional text as English for short English notation and Thanglish words.

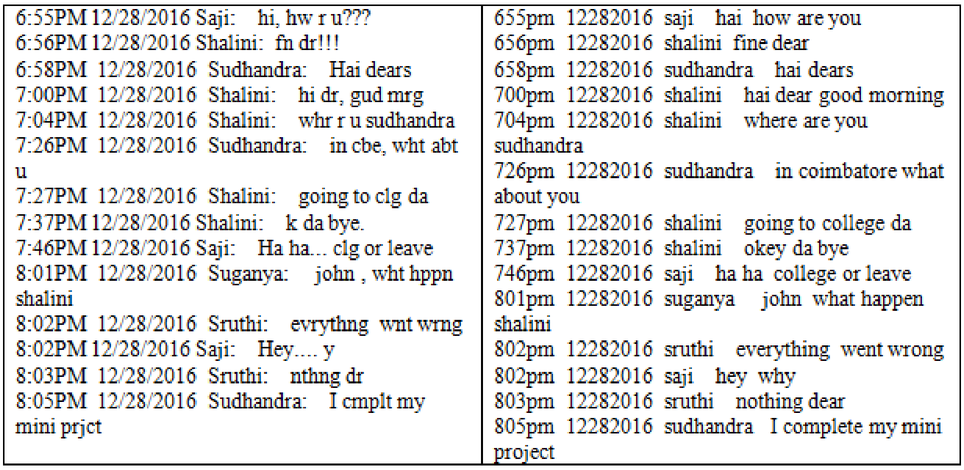


```
> data_corpus
                Key     Value
1               are         r
2              able       abl
3             about       abt
4             above       abv
5             after      aftr
6         afternoon   aftrnun
7            always     alwys
8           america       usa
9             among       amg
10   andra pradesh     andra
```

Figure 8. WhatsApp Chat Corpus Model

4.3 Semantic Match for WhatsApp Data

Using MapReduce function maps the Chat words and English words. The mapper function maps the keys and value pairs. If the key is matched with corresponding values, then data is replaced by keys. Here the English words are known as keys and chat words are known as values. Each key and value pair has the unique id. By using id, key and value pairs are mapped. The Chat words are replaced by original English text. The chat text and replaced text are shown in Figure9.

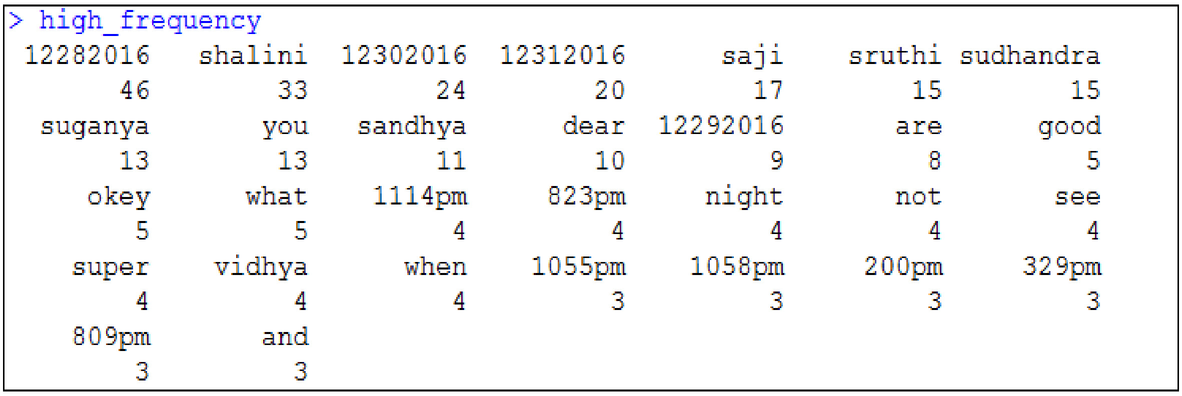


6:55PM 12/28/2016 Saji: hi, hw r u???	655pm 12282016 saji hai how are you
6:56PM 12/28/2016 Shalini: fn dr!!!	656pm 12282016 shalini fine dear
6:58PM 12/28/2016 Sudhandra: Hai dears	658pm 12282016 sudhandra hai dears
7:00PM 12/28/2016 Shalini: hi dr, gud mrg	700pm 12282016 shalini hai dear good morning
7:04PM 12/28/2016 Shalini: whr r u sudhandra	704pm 12282016 shalini where are you sudhandra
7:26PM 12/28/2016 Sudhandra: in cbe, wht abt u	726pm 12282016 sudhandra in coimbatore what about you
7:27PM 12/28/2016 Shalini: going to clg da	727pm 12282016 shalini going to college da
7:37PM 12/28/2016 Shalini: k da bye.	737pm 12282016 shalini okey da bye
7:46PM 12/28/2016 Saji: Ha ha… clg or leave	746pm 12282016 saji ha ha college or leave
8:01PM 12/28/2016 Suganya: john , wht hppn shalini	801pm 12282016 suganya john what happen shalini
8:02PM 12/28/2016 Sruthi: evrythng wnt wrng	802pm 12282016 sruthi everything went wrong
8:02PM 12/28/2016 Saji: Hey…. y	802pm 12282016 saji hey why
8:03PM 12/28/2016 Sruthi: nthng dr	803pm 12282016 sruthi nothing dear
8:05PM 12/28/2016 Sudhandra: I cmplt my mini prjct	805pm 12282016 sudhandra I complete my mini project

Figure9. Semantic match for WhatsApp Chat

4.4 Visualization

The WhatsApp frequent user is viewed in visualization. In WhatsApp chat some terms are occurred frequently. Those words are mostly preferred by WhatsApp senders. Frequent terms are shown in Figure 10. The most occurrence of WhatsApp chats are visualized using chart is represent in the Figure 11.



```
> high_frequency
 12282016   shalini  12302016  12312016      saji    sruthi sudhandra 
       46        33        24        20        17        15        15 
  suganya       you   sandhya      dear  12292016       are      good 
       13        13        11        10         9         8         5 
     okey      what    1114pm     823pm     night       not       see 
        5         5         4         4         4         4         4 
    super    vidhya      when    1055pm    1058pm     200pm     329pm 
        4         4         4         3         3         3         3 
    809pm       and 
        3         3 
```

Figure10.Frequency Terms in WhatsApp Data

WhatsApp senders are associated with the message. First create a term document matrix for the WhatsApp data. In the term document matrix it shows the number of terms (263), documents (100), maximum term length (15) and weighting off the term frequency. Terms hold the WhatsApp senders name and correlation limits for each term in the range from zero to one. The corlimit and terms are shown in Figure12.

The snapshot of the word cloud is shown in Figure 13. The word cloud shows that “12-28-2018”and 12-31-2016” is the dates, where most number of messages is received. The other important words are shalini”, saji”, “sudhandra”, “sruthi” and “suganya”. Then “dear” and “good” are most important adjectives used in the context. Word Cloud for WhatsApp terms are shown below.

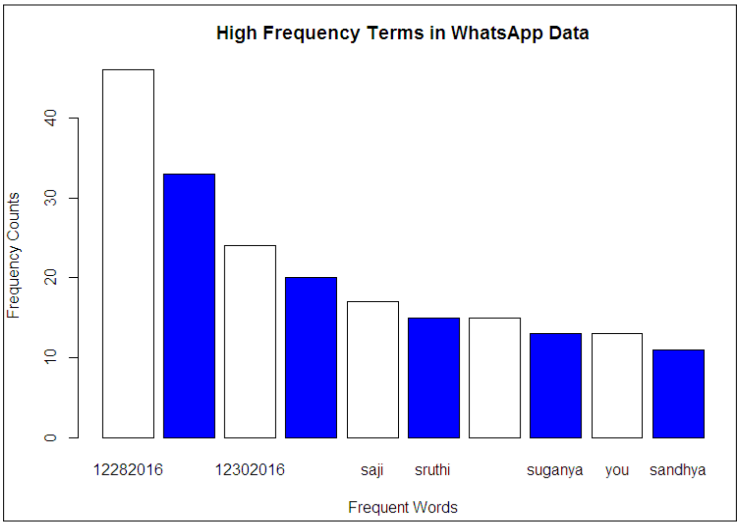
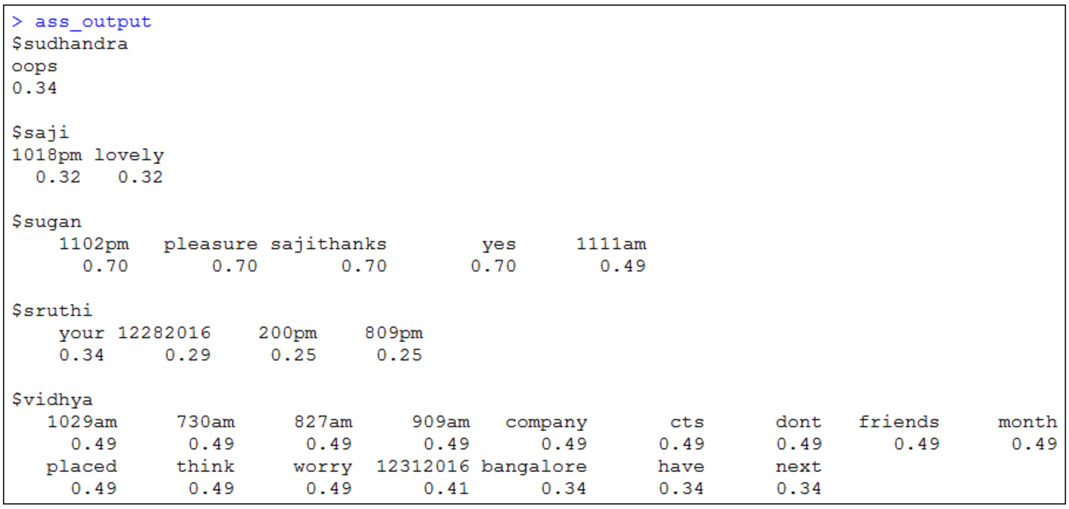


Figure 11. Visualization for Frequent Terms in WhatsApp Data



```
> ass_output
$sudhandra
oops
0.34

$saji
1018pm lovely
  0.32   0.32

$sugan
    1102pm    pleasure  sajithanks        yes      1111am
      0.70        0.70        0.70       0.70        0.49

$sruthi
    your  12282016     200pm     809pm
    0.34      0.29      0.25      0.25

$vidhya
   1029am     730am     827am     909am    company       cts      dont    friends     month
     0.49      0.49      0.49      0.49       0.49      0.49      0.49       0.49      0.49
   placed     think     worry  12312016  bangalore      have      next
     0.49      0.49      0.49      0.41       0.34      0.34      0.34
```

Figure12: Associations of WhatsApp Terms

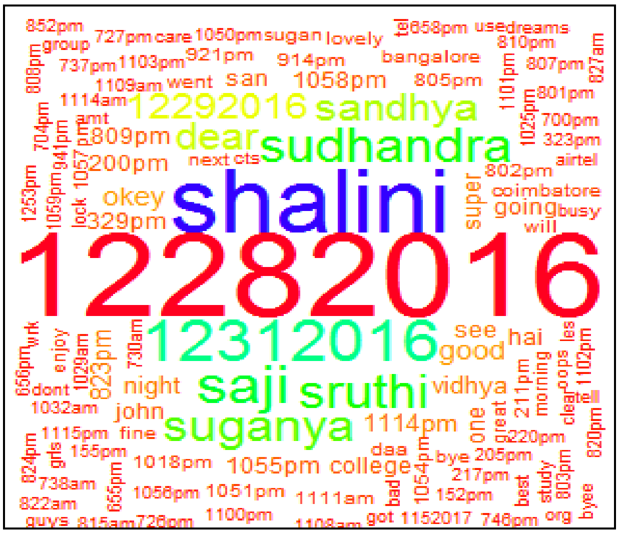


Figure13. Visualization for Frequent Terms in WhatsApp Data

5. DISCUSSION

This work methodology is proposed for converting WhatsApp short message into conventional English notation using MapReduce model. The experimental results found have good accuracy for the dataset used. So we have considered data source only from normal communication group, which is the limitation of the work. In future the work will be extract using the same framework for business chat transliteration. The future work will be extended in the hadoop environment.

6. CONCLUSION

In this work a detailed analysis of MapReduce framework for WhatsApp Chat transliteration is implemented. This MapReduce algorithm is analyzed for WhatsApp data and brief impacts of WhatsApp data on experimental results are discussed. The experimental results were found to be good in mapping the key value pairs. The corpus for WhatsApp chat is created automatically which will serve as the base for other related work. In future this work will be extended and converted as "R-Package" to do chat transliteration automatically. The limitation of the work will of domain corpus of WhatsApp chat will be carried out as future work for large no of WhatsApp text for a specific domain chat text

REFERENCES

Retrieved from www.statista.com/number-of-monthly-active-whatsapp-users.

Retrieved from www.whatsapp.com.

Retrieved from www.mjdenny.com/Text_Processing_R.

Retrieved from www.tutorialspoint.com/Mapreduce.

Retrieved from www.introduction-to-text-mining-using-R.

Retrieved from www.analyticsvidhya.com.

Sawitzki, G. "An Introduction to R", Computational Statistics. CRC Press.

V.Bhuvaneswari. (2016). BIG DATA ANALYTICS - A Practitioner's Approach. Coimbatore: Department of Computer Applications, Bharathiar University.