CHAT REVIEWS

Project Id: 17-066

Project Proposal Report

Dinesh Lakmal E

Cooray B.S.U.M

Theekshana M.A.H

Harischandra K.P.I.E

B.Sc. Special (Honors) Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

March 2017

CHAT REVIEWS

Project Id: 17-066

Project Proposal Report

(Proposal documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (honors) In Information Technology)

B.Sc. Special (Honors) Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

March 2017

Project ID: 17-066

Authors:

Student ID	Name	Signature
IT14085840	Dinesh Lakmal E	
IT14032066	Cooray B.S.U.M	
IT14029950	Theekshana M.A.H	
IT14093760	Harischandra K.P.I.E	

Supervisor

…….

Ms. Nethmini Weerawarna

DECLARATION

We declare that this is our own work and this project proposal does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

……	……..…….……
Dinesh Lakmal E	Cooray B.S.U.M
……..	……
Theekshana M.A.H	Harischandra K.P.I.E

ABSTRACT

The use of Internet chat applications has benefited many different segments of society. It also creates opportunities for criminal enterprise, terrorism, and espionage. We present a study of a real-world application of chat analysis which will analyze chat messages in four ways such as topic detection, Emotion Extraction, Evaluate healthy and Personal information sharing analysis. Also analyzing chat traffic has important applications for both the military and the civilian world. Here on this document, it compares the results of an unsupervised learning approach with those of a supervised classification approach with regards to chat review application. The paper also discusses some of the specific challenges presented by this chat review application.

Unsupervised learning techniques such as clustering are very popular for analyzing text for topic identification as well as emotion extraction. These techniques have several attractive features, the most significant being that they do not require labeled training examples. This however is also a disadvantage under some circumstances. Therefor meantime we do this research we will discover more and more technologies required for analyzing chat massages based on four different categories such as topic detection, Emotion Extraction, evaluate healthy and Personal information sharing analysis.

With use of this chat analysis application user will be able identify the chatting partner in analytical way. And system will keep an analytical review for each chat session user interacted. Also system will be capable of showing its analytical data in a user friendly manner (in a graphical way).

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
CMC	Computer-mediated Communication
NLP	Natural Language Processing
ML	Machine Learning
ASR	Automatic Speech Recognition
SNS	Social Networking Services
ASU	Automatic Speech Understanding
IM	Instant Messaging

1. INTRODUCTION

1.1. Background and literature survey

Chat is an increasingly important form of CMC (Computer-mediated communication). It is employed by many sectors of society to improve communication, create value, and commit crimes. In this chapter, we explore chat first as it relates to other human language modalities. Then, we explore Natural Language Processing (NLP) and its goals, followed by its applicability to chat. Finally, we discuss the idea of topicality and previous Machine Learning (ML) techniques used to detect topics in chat. The objective of automatically revealing the topic of any form of communication is twofold. The first motive is to increase the knowledge of how humans communicate, to unravel the mystery of information conveyance. The second is to build useful systems. Topic detection and emotion extraction in this context are steps toward automating tasks that would otherwise be untenable, because the sheer volume of data makes it impractical. Section A provides working definitions of CMC, chat, and natural human languages.

1.1.1 Chat's relationship to other natural language constructs

This work considers chat to be a mode of natural language. Natural language is simply a language that humans speak [1]. This definition is a bit restrictive because it does not account for written text, sign language, Braille and manual languages developed by deaf-blind people. However, it is helpful in the sense that it conveys the origin of language people. Speaking, writing, Braille, signing, and CMC are all methods that represent the innovations and phenomena of natural language. We briefly compare and contrast chat to two of these modes speaking and traditional writing.

According to Herring, CMC is "communication that takes place between human beings via the instrumentality of computers" [2]. CMC itself takes many forms to include e-mail, Weblogs (blogs), micro-blogs, video, audio, text chat, text messaging and instant messaging. Social networking services (SNS) and online games also use computers and computer networks to mediate communication by combining many CMC technologies.

Many online gaming venues feature some chat functionality or audio service. SNSs combine nearly all of these CMC forms.

Text chat is a subcategory of chat that is a near-synchronous form of CMC. Chat may be categorized as text chat, voice chat and video chat. This research is dedicated solely to text chat. Any further reference to chat will mean text chat. Further, we narrow our meaning of chat to that which is near-synchronous, multi-member conversation contributed and conversation interleaved. Near-synchronous means that conversation contributors interact in near real-time. They are temporally proximate to each other. Multi-member conversation contributed refers to the fact that there may be more than two people contributing to the chat statements (posts) the same time. Conversationally interleaved indicates that the many conversations (threads) may occurring at the same time.

Chat communication is affected significantly by its technological implementation its computer mediation [3]. All chat implementations share some common characteristics. First, there is a main dialog. It is a relatively large text field, which displays the posts created by all chat room participants. The main dialog is public to all. If a poster wishes to "say" something, the main dialog is where it is displayed for all to see [3]. Second, there is a personal dialog. This is a text field where each user composes his or her particular posts. It is, for the most part, private to each particular user.1 after the user has completed composing her message, she posts the message to the main dialog. The message she composed in private is now public to all participants [3].

1.1.2 NLP and chat

NLP problems are approached in two basic ways rules-based NLP and statistical NLP. Rules based NLP practitioners assume that humans possess a great deal of underlying knowledge of language that allows humans to learn particular languages. Their objective is to model these mental processes in order to create a system that mimics or duplicates the functioning of the human brain. This process usually starts with creating rules that mimic such functionality [7].

Statistical NLP practitioners agree with the rules-based-inclined in that they proceed from the assumption that humans possess something innate that enable them to recognize patterns, which allows humans to learn how to communicate. They, however, differ on the degree to which humans possess this ability [7] the former more and the latter less. Statistical NLP's general approach is to build statistical models of language and then use ML techniques to validate those models. This research approaches the problems using statistical NLP methods.

NLP is used in problems, which involve spoken language or speech after automatic speech recognition (ASR) has been applied. ASR seeks to build a mapping between sounds and strings. Automatic speech understanding (ASU) takes this goal one-step further and tries to understand the words in the broader context of a sentence [1]. Like ASU, conversational agents leverage NLP. Conversational agents such as SGT Star6, a U.S. Army avatar that chats with potential Army enlistees, receives chat input from users and outputs textually, visually and audibly. Just as with telephone menus, SGT Star's capabilities are limited, but indicate an ideal direction for such systems.

1.2. Research gap & Research problem

Identify characteristics of the chatting partner by analyzing chat messages. Nowadays we meet strangers all the time in our day to day life. So we attempt to have partnership with them without any fear at all. Also some time they are more and more smart than we think, then it's very difficult to identify the characteristics by only looking at their messages, because any one can send anything. Nowadays it is highly required to have intelligent way to detect those frauds (may be what that message really mean). Solution is to use analytical application for online chatting. Then we can review our messages in analytical way or it will lead us to think about our chatting partner in analytical way.

It's a web application that analyzes messages in the chat box in short time of period. It simply gives us clear idea about the message by analyzing it in various ways. In that case we can be very much aware about our chatting partner. So this application let you deal with that strange people in understandable way. It shows you auto generated graphs based on emotion, personal information sharing and topics discussed etc. And also we can confirm the outcomes of this system by analyzing public profile, and other related social profiles of particular user or other data sources if possible.

2. OBJECTIVES

2.1 Main objectives

From the study on chat message characteristics, an indicative term-based and single person categorization approach for chat topic detection and study of behavior of the person are proposed. In the proposed approach, different techniques such as sessionalization of chat messages, chat message history and extraction of features from short texts and URLs are incorporated for message pre-processing. And Associative Classification, and Support Vector Machine are employed as classifiers for categorizing topics from chat sessions. This will help to opposite partner to identify their behavior and trustworthy. Nowadays there are so many fake profiles in social Medias. Then this will help to match partners as they wish.

2.2 Specific objectives

- To keep users interested in using the application.
  Provide a social media application to the users mainly focusing on find new friends, sharing the details with friends in real time. So if the user is fake that will be some risk to the other user, so then this guides to identify true people.

- To come up with true peoples
  This will help to identify if the people say true or not by analysing their previous behaviours.

3. RESEARCH METHODOLOGY

3.1 System overview

This application studies the characteristics of chat messages and proposes an indicative term-based categorization approach for chat topic detection, emotion extraction. The proposed approach has been incorporated into an instant message analysis system for both online and offline chat topic detection. The primary objective of chat message characterization is to understand the properties of chat messages for effective message analysis such as message topic detection and emotion extraction. From the study on chat message characteristics, an indicative term-based categorization approach for chat topic detection is proposed. In the proposed approach, different techniques such as sessionalization of chat messages and extraction of features from icon texts and URLs are incorporated for message pre-processing.

Chat language is basically written English. However, due to the real-time and informal conversational environment of IM systems, chat messages are written in a very different way from conventional English. Some of the common usage features in chat language include acronyms, short forms, polysemy, synonyms and mis-spelling of terms.

Using chat box integrated with the system we collect messages for the purpose of analyzing them in for ways and collected messages will be stored in a database. Then messages will be recollected from database according to the chat sessions and Using clustering based algorithms it further group into the four major categories such as topic detection, Emotion extraction, Evaluate healthy and Personal information sharing analysis. Using appropriate algorithms messages will be analyzed and it will generate analytical data which will be further used in graphical review phase as figure 3.1 shown below.

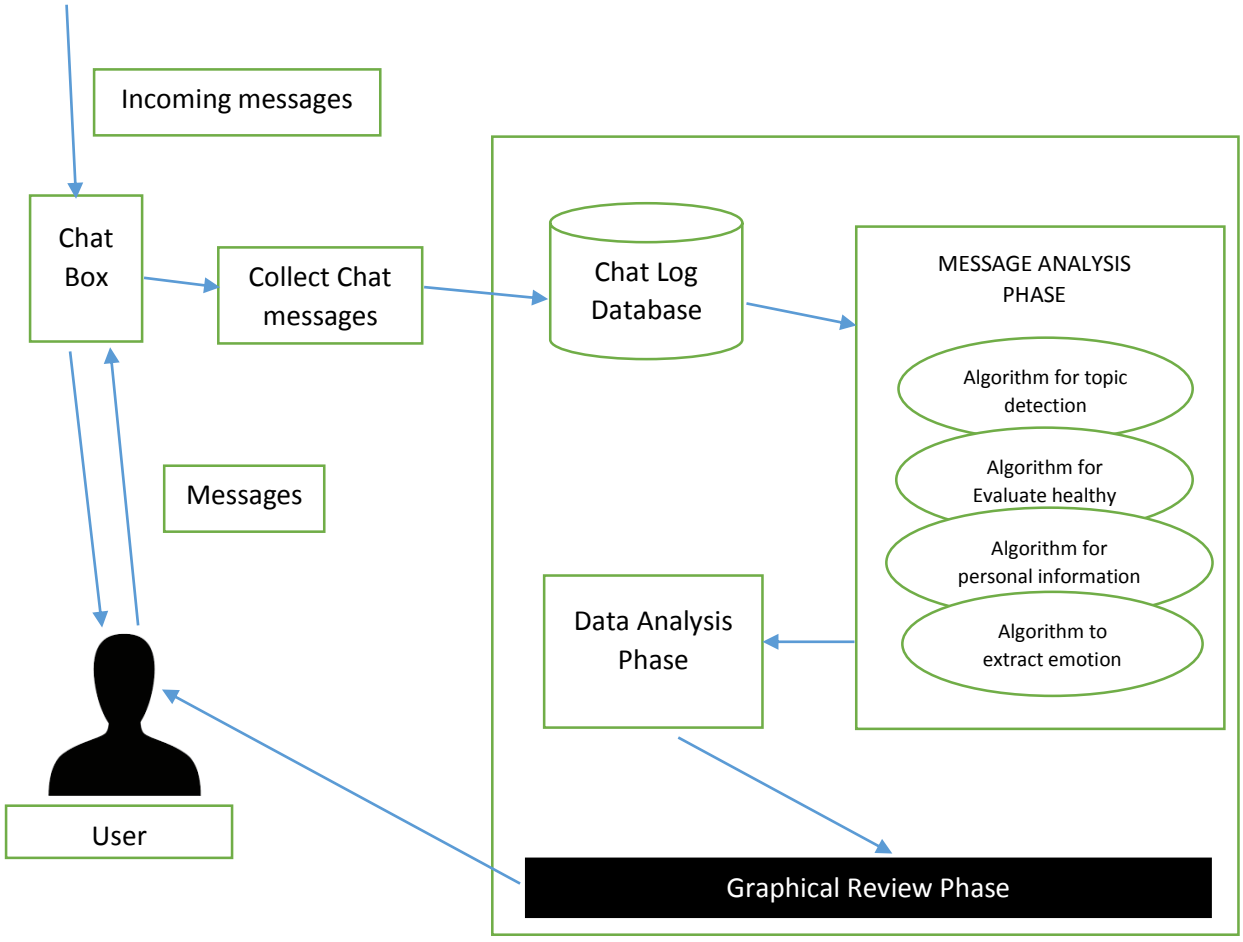


Figure 1: System Overview Diagram 1

3.2 Coming up with advanced chat messages analysis with third party chat box integration.

Extracting chat messages from chat box will be a kind of bit challenging task with the integration of third party chat room which is developed by other party. Fortunately, there are many free chat room services which allow us to create your own room and either provide a simple link to that room, or add that chat room to our web application.

Figure 2: Chat Box integration

Also system will save all the chat messages as text file or any suitable format like .CSV (comma-separated values) for the purpose of analysing them based on topic, emotion, healthy and personal information.

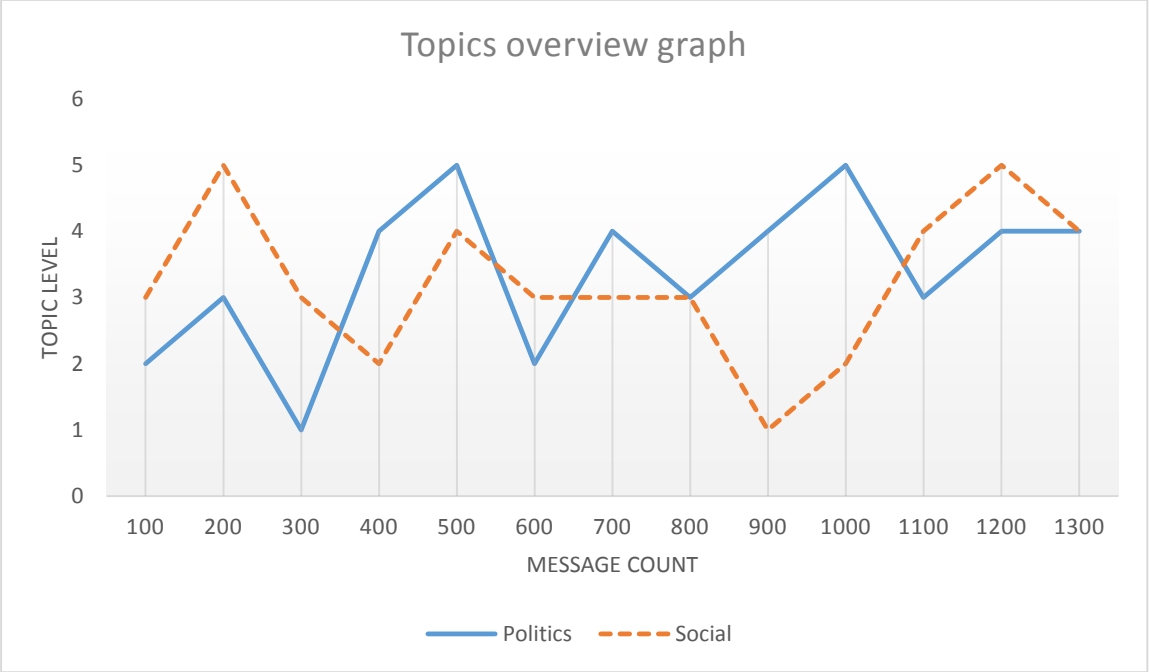
3.3 Algorithms for analysing chat messages.

For analyzing chat messages we will use different similarity-based clustering approaches to group chat messages together according to topic, emotions etc.... One used the term frequency-inverse document frequency (TF-IDF) similarity measure presented in (Adams and Martell, 2008). Here similarity is determined by the number of overlapping words between two messages, weighted by the uniqueness of the words. Also we will use a hierarchical clustering algorithm where each chat message is matched with its nearest neighbor. If one message from this pair is already a part of a cluster, the other is added to it. If each message in the pair belongs to different clusters, these are merged. If neither belongs to a cluster, then a new cluster is created. We modified this basic algorithm to include stemming, filtering of stop words, and a moving chat window. The stop word list consisted of the hundred most common words in English. It also included all the call signs that are used by the team to identify each member. A chat window was introduced in an effort to localize the clusters based on the observation that topics typically consist of subtopics that shift over time. With this modification, each message was paired with a nearest neighbor occurring within a surrounding window (currently set to include 10 messages occurring before and 10 messages after the one under consideration). Finally, the algorithm ignores messages with less than three words (as most of these are related to message acknowledgments).

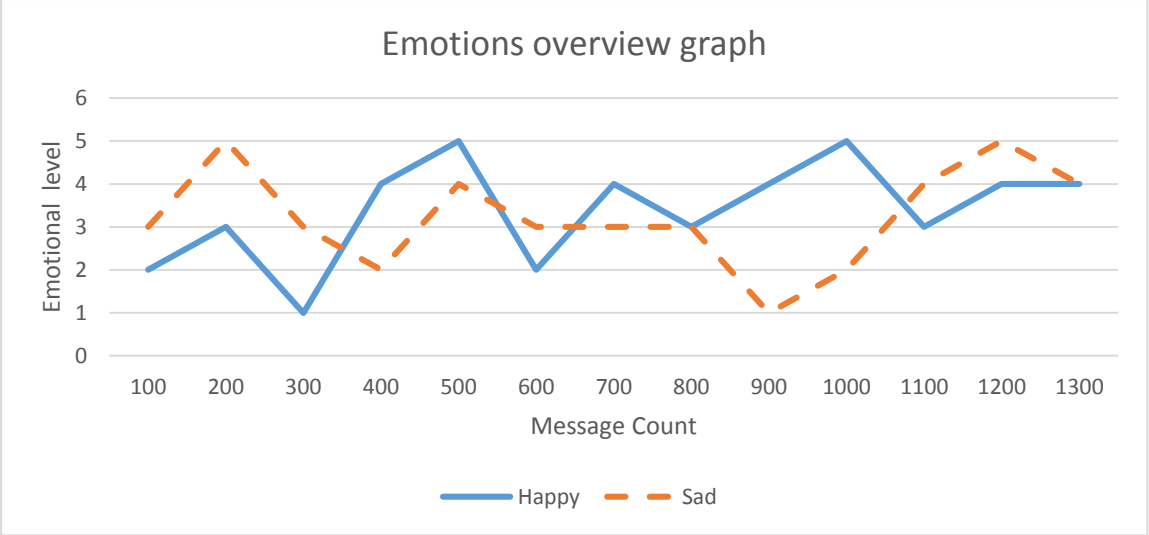
Once clusters are identified, the algorithm then assigns to each cluster a topic label based on Rule 1. One of the following is true about each cluster identified: 1. None of the messages in the cluster were assigned a label by Rule 1, 2. Some messages in the cluster were labeled by Rule 1 and all them are identified with the same topic, 3. Some messages in the cluster were assigned labels by Rule 1 and they are identified with different topics. In the first case, the cluster itself is not assigned any label. In the second case, all the messages assigned to this cluster are assigned to the topic identified. In the third case, the cluster is disregarded because it represents multiple topics and therefore not considered relevant.

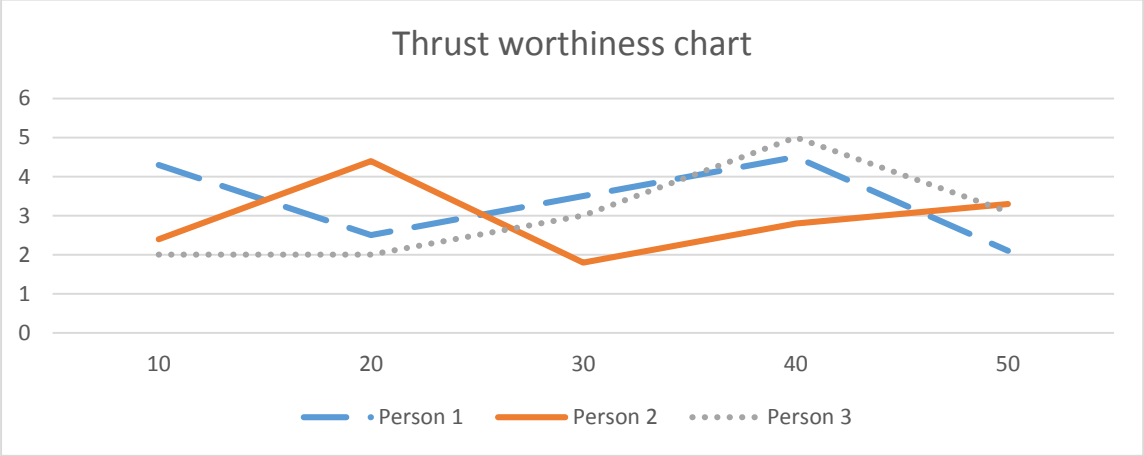
3.4 Visualize the analytical data.

Finally system will generate set of analytical data through algorithms and clustering methods or with any other methodologies which we will use for this chatting analysis application. Based on analytical data system will calculate and will build a review about the trust worthiness of our chatting partner. Also it can review out chat summery in analytical way. Then system also will be capable of showing final outcomes in graphical manner as chart shows below.



From graph shown above it can visualize the user how the topics are being changed with the message count and the point of message count which made it inspired to change the topic. In the above graph system will be capable of showing how topics varies with message count and how it changes its topic level with message count.





Here it takes a number to show up the trust worthiness of our chatting partner as above chart shows it varies with the messages count increment. Then it will be easier for the user to get aware about the chatting partner.

3.5 Gantt chart.

Figure 3: Gantt Chart

3.6 System diagram.

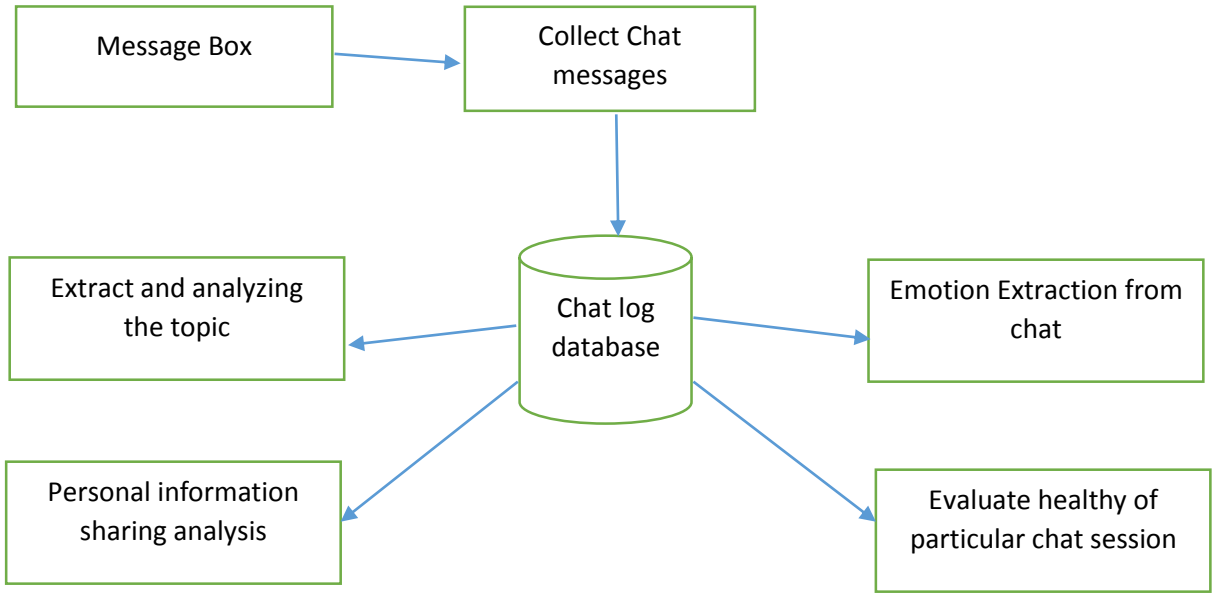


Figure 4: System Overview Diagram 2

4. DESCRIPTION OF PERSONAL AND FACILITIES

Table 1: Personal and facilities.

Member	Component	Task
Dinesh Lakmal .E	Analysis of chat messages for topic detection.	• Collect messages from chat box. • Group chat messages, according to the topic using "similarity based clustering" approaches. • Identifying learning algorithms for chat analysis based on identified topic. • Visualize the outcomes of analyzed data in a graphical manner and generate modules.
Cooray B.S.U.M	Emotion Extraction from chat messages (Negative and positive).	• Group chat messages, according to the Emotion component. • Identifying learning algorithms for chat analysis based on identified emotion set. • Visualize emotional impact and generate modules.
Theekshana M.A.H	Evaluate healthy of particular chat session.	• Extract fewer first-person pronouns, fewer exclusionary words, unusual detail from collected messages. • Identify those tics with message counts using

		appropriate algorithms and visualize. • Compare and analyze the identified module based on analytical data of topics.
Harischandra K.P.I.E	Personal information sharing analysis.	• Collect the information which seems to be personal from particular chat session or task oriented check chat aligned with the topic. • Group the information by analyzing and visualize mutuality status. • Compare and analyze the identified module based on analytical data of emotions.

5. REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. London: Prentice Hall, Pearson Education International, 2009.

[2] S. C. Herring, "Computer-Mediated Communication Linguistic, Social, and Cross-Cultural Perspectives," 1996.

[3] E. N. Forsyth, "Improving automated lexical and discourse analysis of online chat dialog," M.S. thesis, Naval Postgraduate School, Monterey, CA, 2007.

[4] M. Zitzen and D. Stein, "Chat and conversation: a case of transmedial stability?" *Linguistics,* vol. 42, pp. 983–1021, 2004.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, 2003.

[6] M. R. Freiermuth, "Features of electronic synchronous communication: a comparative analysis of online chat, spoken and written texts," 2002.

[7] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, 6th printing ed. Cambridge, Mass.: MIT Press, 2003.

[8] J. G. Shanahan and N. Roma, "Improving SVM text classification performance through threshold adjustment," in *Machine Learning: ECML 2003*, vol. 2837/2003, Berlin / Heidelberg: Springer Berlin / Heidelberg, 2003.

[9] W. Cohen, V. R. Carvalho and T. M. Mitchell, "Learning to Classify Email into Speech Acts," *EMNLP,* 2004.

[10] J. Lin, "Automatic author profiling of online chat logs," M.S. thesis, Naval Postgraduate School, Monterey, CA, 2007.

[11] J. Tam, "Detecting age in online chat," M.S. thesis, Naval Postgraduate School, Monterey, CA, 2009.