CHAT REVIEWS

Dinesh Lakmal E

(IT14085840)

Degree of Bachelor of Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2017

CHAT REVIEWS

Dinesh Lakmal E

(IT14085840)

Dissertation submitted in partial fulfillment of the requirements for the degree of Science

Department of Information Technology

Sri Lanka Institute of Information Technology

October 2017

DECLARATION

I declare that this is my own work and this dissertation1 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 my 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.

Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

…… …../…../…….

Dinesh Lakmal E

IT14085840

The above candidate has carried out research for the B.Sc. Special (Hons) degree in IT Dissertation under my supervision.

…… …../…../…….

Dr. Dharshana Kasthurirathna.

(Signature of the Supervisor)

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 messages are important for both the military and the civilian world. Here on this document, it compares the results of both unsupervised and supervised machine learning approach with regards to chat review application. The paper also discusses some of the specific challenges presented by this chat review application.

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).

Instant Messaging (IM) is a service for users to communicate with each other. There are many Instant Messaging (IM) systems such as MSN Messenger [1] and Yahoo Messenger [2], which are used by millions of users. IM monitoring systems have been developed for monitoring chat messages. Although most of these systems can provide good monitoring functions, they only provide simple message analysis features such as browsing and simple keyword based searching of the recorded messages. In this paper, we explore a system called Chat Reviews that supports intelligent chat message analysis using machine learning techniques and this system provides four IM monitoring modules to analyze IM. Final statistics generated by those four modules in this IM system are used to review chat session of chatting partner.

ACKNOWLEDGEMENT

The work described in this research paper was carried out as our 4th year research project for the subject Comprehensive Design Analysis Project. The completed final project is the result of combining all the hard work of the group members and the encouragement, support and guidance given by many others. Therefore, it is researchers' duty to express their gratitude to all who gave them the support to complete this major task.

The researchers are deeply indebted to supervisor Dr. Dharshana Kasthurirathna and Lecturers of Sri Lanka Institute of Information Technology whose suggestions, constant encouragement and support in the development of this research, particularly for the many stimulating and instructive discussions. We are also extremely grateful to Mr. Jayantha Amararachchi, Senior Lecturer/ Head-SLIIT Centre for Research who gave and confirmed the permission to carry out this research and for all the encouragement and guidance given.

The researchers also wish to thank all colleagues and friends for all their help, support, interest and valuable advices. Finally, the researchers would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged researchers to make this a success.

TABLE OF CONTENTS

DECLARATION i
ABSTRACT ii
ACKNOWLEDGEMENT iii
TABLE OF CONTENTS iv
LIST OF TABLES vi
LIST OF FIGURES vi
LIST OF ABBREVIATIONS vii
1 INTRODUCTION 1
1.1 Background Context 1
1.2 Research Gap 1
1.3 Research problem 2
1.4 Research objectives 3
2 METHODOLOGY 4
2.1 Methodology 4
2.1.1 Classification algorithm for Topic Detection 5
2.1.2 System Design 8
2.1.3 Implementation 8
2.1.4 Topic Detection Overview 9
2.2 Testing and Implementation 10
2.2.1 Implementation Techniques 10
2.2.2 Testing Techniques 10
2.2.3 User Interface of Topic Detection Module 11
2.3 Research Findings 11
3 RESULT AND DISCUSSION 13
3.1 Results of the Topic Detection Module 13
3.1.1 Test Cases 15
3.2 Discussion of the System 16
3.2.1 Flow of the Project 17
4 CONCLUSION 18
REFERENCES 19
APPENDICES 20

LIST OF TABLES

Table 1- Comparison of Exiting Application with the proposed system.....2
Table 2 - Test cases.....15

LIST OF FIGURES

Figure 1- Topic Detection overview diagram.....9
Figure 2 - UI of topic detection module.....11
Figure 3 - Review of result of detection module 13
Figure 4 - Result of algorithm.....14

LIST OF ABBREVIATIONS

Abbreviation	Definition
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 Context

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.

We explored chat monitoring system first and how it is used to monitor chat messages. Then, we explored technologies related Chat monitoring systems and some of them are Natural Language Processing (NLP), Machine Learning (ML), followed by its applicability to chat. And text mining and NLP are commonly used together for different purposes, and one of most common applications is social media monitoring [3], where an analysis is performed on a pool of user-generated content to understand mood, emotions and awareness related to a topic [4].

Nowadays most of chat monitoring system use Machine learning algorithms for text classification. One of the main ML problems is text classification, which is used, for example, to detect spam, define the topic of a news article, or choose the correct mining of a multi-valued word. The Statsbot [5] team has already written how to train your own model for detecting spam emails, spam messages, and spam user comments [6]. And it's impossible to define the best text classifier. In fields such as computer vision, there's a strong consensus about a general way of designing models − deep networks with lots of residual connections. Unlike that, text classification is still far from convergence on some narrow area.

1.2 Research Gap

When comparing Chat Reviews web application with existing applications our one based on mainly analyzing trustworthiness of the chatting partner.

So when we compare existing chat monitoring applications with our one it only allows user for simple chat analysis features. So in that case other than the basic functions like keyword based search, topic identification emotion extraction, we gave a new functions (new message analytical areas) such as Detect Personal Information and Evaluate Healthy. This is not available on current any of the chat monitoring application.

Table 1- Comparison of Exiting Application with the proposed system

Chat Monitoring Application	Keyword based searching	Topic Identification	Emotion Extraction	Message Encryption	Detect Personal Information	Evaluate Healthy
Intelligent Diagnosis System		**Yes**	**-**	**-**	**-**	
Honey Chatting	**-**	**-**	**-**	**Yes**	**-**	
GroupWize	**-**	**Yes**	**Yes**	**-**	**-**	
Chat Reviews	**-**	**Yes**	**Yes**	**-**	**Yes**	**Yes**

Above diagram shows a brief comparison of the existing applications with the proposed system

1.3 Research problem

Nowadays we meet strangers all the time in our day to day life. So we attempt to have partnership with them without any hesitation 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. Nowadays it is highly required to have intelligent way to detect those frauds (may be what that message really mean). We found that the solution is to use chat monitoring 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.

The research problem to be addressed by this research was identified as reviewing chat session in analytical way. A separate background analysis was carried on the usage of chat monitoring system and the possibility of developing such kinds of real world system.

Before developing the Chat Reviews Application based on machine learning concept, the project team went through a large amount of research papers to identify the main problems that need to be addressed when implementing the system. With the researches already done we got to know the existing technologies as well as upcoming technologies and algorithms and how to develop this system by modules.

It became clear that a unique system could be implemented by machine learning algorithms. And it was necessary to search for the most suitable machine learning technologies for implementing our system.

There were different problems that we considered during this research project. Some of them are;

- What are best machine learning algorithms for text classification and text mining?
- How to optimize existing machine learning algorithms?
- How to process large data set (Vocabulary)?
- How to collect training data set for each modules in the system?
- What is best programming language to implement machine learning algorithms?
- How to give real time out put to user?
- How to represent statistics generated by algorithms?
- How to keep the users interest to use our application?

This system tries to address the above problems with regard to user satisfaction and improved service and leave our new chat monitoring system in a user interactive manner.

1.4 Research objectives

- Introduce machine learning algorithms for each modules in the system.
- To review chat session in analytical way.
- To analyze characteristics (trustworthiness) of the chatting partner.
- Give user a graphical review of statistics

2 METHODOLOGY

2.1 Methodology

Several machine learning methodologies were used in the developing process by the development team. In my research module, Machine learning algorithms have been used. At the beginning of the development of the system, a deep study about machine learning concept was carried out and different articles were followed based on those. Weaknesses in the existing systems and new requirements were identified after communicating with different people with different age levels. Nowadays at least one social media application is being used by people belong to each and every age group. So the best way to do a study was to get some ideas and feedbacks from those users. At present social media applications are being used for different purposes. They are being used to interact with friends, for fun and to spend time. Therefor different users are having different perspectives. After the study, security of the system and speed of the algorithms was identified as the most important part missing in the existing applications.

After all the studies that had been carried out, I came up with a new area for developing this topic detection module. Developing an algorithm for topic detection module was was considered as one major segment of this module. The reason why this is important is it is more attracting feature of this chat monitoring system and it will be more useful for analytical purpose in the end. As I mentioned above, if user wants to review a particular chat session by topic, user can view statistics in graphical way.

Technologies and application platforms that we were going to use were decided at the important initial stages. Technical difficulties of the development methodologies that we were going to use and the efficiency of the algorithms we were going to implement were examined after doing a deep literary survey and surfing through internet. In the circumstances like ours, doing discussions and surveys is considered as the best or the only way to solve such technical and practical issues.

This research project was titled as Reviewing Chat session (Chat Reviews) based on different chat reviews module such as Topic detection, Emotion Extraction, Detect personal information and Evaluate healthy. The system must be able to extract the given message and analyze it in real-time (maximum of 10 seconds) also it will update the existing statistics immediately. And the graphical review will be provided as user friendly as possible.

Chat Review application is based on 4 main parts.

1. Topic Detection
2. Emotion Extraction
3. Detect Personal Information
4. Evaluate Healthy

After going through above modules final outcome will be shown to the user in the way of analyzing trustworthiness of particular chat session.

2.1.1 Classification algorithm for Topic Detection

Topic detection module detects the topic associated with messages. There are defined topic classes along with the training data associated with each topic class. For topic analysis in IM Analysis, we aim to identify chat sessions limited only to a number of important topics. Therefore, we have adopted supervised topic detection approaches based on Naïve **Bayes Theorem** [7], and **Bernoulli document model** [8] which have been demonstrated with good performance for classifying text.

Topic detection module use training dataset collected from social networks based on related topics and classification algorithm (supervised learning) [9] to classify message. Text classifiers often don't use any kind of deep representation about language: often a message is represented as **a bag of words** [10]. (A bag is like a set that allows repeating elements.) This is an extremely simple representation: it only knows which words are included in the message (and how many times each word occurs), and throws away the word order! Consider a Message M, whose class is given by C. In the case of topic detection there are set of classes C = S (Sport) and C = P (Politics) etc... We classify M as the class which has the highest posterior probability P(C|M), which can be re-expressed using **Bayes' Theorem** [7]:

$$P(C|M) = \frac{P(M|C)P(C)}{P(M)} \propto P(M|C)P(C) \quad (1)$$

Topic Detection module uses **Bernoulli document model** which represent messages as a bag of words, using the **Naive Bayes** assumption. This models represent messages using feature vectors [11] whose components correspond to word types. If we have a vocabulary V, containing |V| word types, then the feature vector dimension d=|V|.

As mentioned above, in the Bernoulli model a message is represented by a binary vector, which represents a point in the space of words. If we have a vocabulary V containing a set of |V| words, then the t^{th} dimension of a message vector corresponds to word w_t in the vocabulary. Let bi be the feature vector for the i^{th} message M_i ; then the t^{th} element of b_i, written bit, is either 0 or 1 representing the absence or presence of word w_t in the i^{th} document. Let P(w_t |C) be the probability of word w_t occurring in a message of class C; the probability of w_t not occurring in a message of this class is given by (1−P(w_t |C)). If we make the naive Bayes assumption, that the probability of each word occurring in the message is independent of the occurrences of the other words, then we can write the message likelihood P(M_i | C) in terms of the individual word likelihoods P(w_t |C):

$$P(M_i|C) \sim P(b_i|C) = \prod_{t=1}^{|v|}[b_{it}P(w_t|C) + (1 - b_{it})(1 - P(w_t|C))] \tag{2}$$

This product goes over all words in the vocabulary. If word w_t is present, then b_{it}=1 and the required probability is P(w_t|C); if word w_t is not present, then b_{it}=0 and the required probability is 1− P(w_t |C). We can imagine this as a model for generating message feature vectors of class C, in which the message feature vector is modelled as a collection of |V| weighted coin tosses, the t^{th} having a probability of success equal to P(w_t |C). The parameters of the likelihoods are the probabilities of each word given the message class P(w_t |C); the model is also parameterized by the prior probabilities, P(C). We can learn (estimate) these parameters from a training set of messages labelled with class C =k. Let $n_k(w_t)$ be the number of messages of class C=k in which w_t is observed; and let N_k be the total number of messages of that class. Then we can estimate the parameters of the word likelihoods as,

$$\hat{P}(w_t \mid C = k) = \frac{n_k(w_t)}{N_k}, \tag{3}$$

the relative frequency of messages of class C = k that contain word wt. If there are N messages in total in the training set, then the prior probability of class C =k may be estimated as the relative frequency of messages of class C=k:

$$\hat{P}(C=k) = \frac{N_k}{N}. \tag{4}$$

Thus given a training set of messages (each labelled with a class), and a set of K classes, we can estimate a Bernoulli text classification model as follows:

1. Define the vocabulary V; the number of words in the vocabulary defines the dimension of the feature vectors.

2. Count the following in the training set:

- N the total number of messages
- N_k the number of messages labelled with class C=k, for k=1,… K,
- $n_k(w_t)$ the number of documents of class C=k containing word w_t for every class and for each word in the vocabulary.

3. Estimate the likelihoods P(w_t | C=k) using equation (3)

4. Estimate the priors P(C=k) using equation (4)

To classify an unlabeled message M_j, we estimate the posterior probability for each topic class combining equations (1) and (2):

$$\begin{aligned} P(C|M_j) &= P(C|b_j) \\ &\propto P(b_j|C)P(C) \\ &\propto P(C)\ \prod_{t=1}^{|v|}[b_{jt}P(w_t|C) + (1-b_{jt})(1-P(w_t|C))] \end{aligned} \tag{5}$$

Then message will be labeled into specific topic class which is the highest posterior probability computed by above equation.

2.1.2 System Design

The design part is done by dividing it into several parts.

- User interface designing: This will be the first part in designing because the user is interacting with the interface which is provided. So the interface must be user friendly and must fulfill all the needs and requirements of the user. And must be attractive to user's eye.
- Database design: For the database management we decide to use MySQL [12] as it is more recommend for backend operation and Oracle Database 11g Release for topic detection module as it is supported for Data mining and Data analytical purposes. So System will consist with two databases.
- Algorithm design: We have used **Bernoulli document model** for classifying messages into topic.

2.1.3 Implementation

According to the proposed system user have the facility to access our system with any device that has internet facility. And System can receive messages from Facebook pages connected through messenger bot. Messages are going through all four modules integrated in the system within few seconds.

Topic detection module is developed using Java and Oracle Database 11g Release as database. It is integrated using PHP-Java integration methodologies with other three modules in the system.

2.1.4 Topic Detection Overview

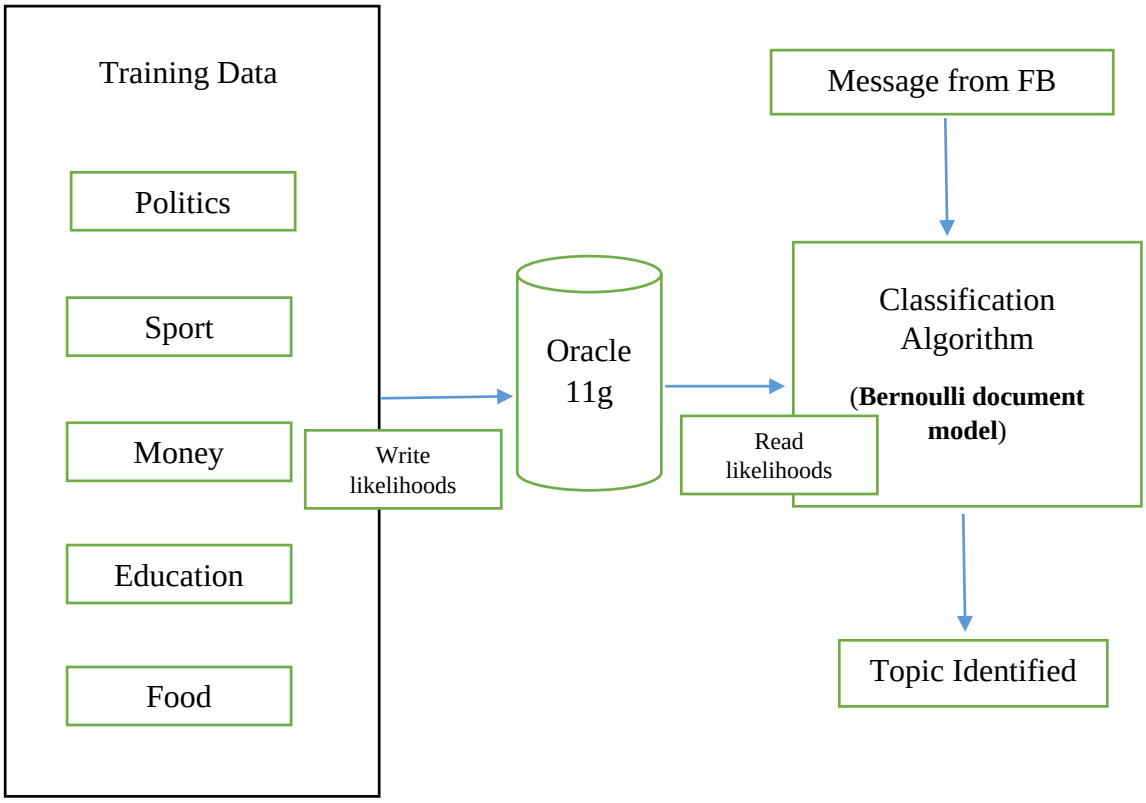

Figure 1- Topic Detection overview diagram

2.2 Testing and Implementation

2.2.1 Implementation Techniques

- HTTPS web host
- Servers required
    - Web Server - Appache Server

- Messenger bot in messenger platform
- To implement web application
    - JavaScript
    - PHP
    - HTML
    - CSS
- Microsoft Office package for Documentation
- Hardware Requirements
    - Internet Connection
    - And device that has the facility to access internet
- Software Requirements
    - Web browser enables with internet access

2.2.2 Testing Techniques

➢ **Testing Methodology**

Any software product should be sent through a testing process in order to identify the weaknesses or faults of the system. This section provides an overview of how the system behaves in the phases recognized in the system. This is accomplished through the testing process where each and every module or phase of the system is tested.

- **Unit Testing**

Unit testing will be carried out during the implementation process. Each unit will be tested by the developer himself.

- **Integration Testing**

Once more than one module is completed, integration testing will begin. Two modules will be integrated and tested. As a third module is joined, all three modules will be tested again.

- **System Testing**

Once all the modules have been successfully integrated, system testing will begin. Three kinds of System tests will be carried out. And for further testing application was given to some of the users and followed them.

2.2.3 User Interface of Topic Detection Module

Following is the user interfaces for the graphical review of topic analysis of particular chat session.

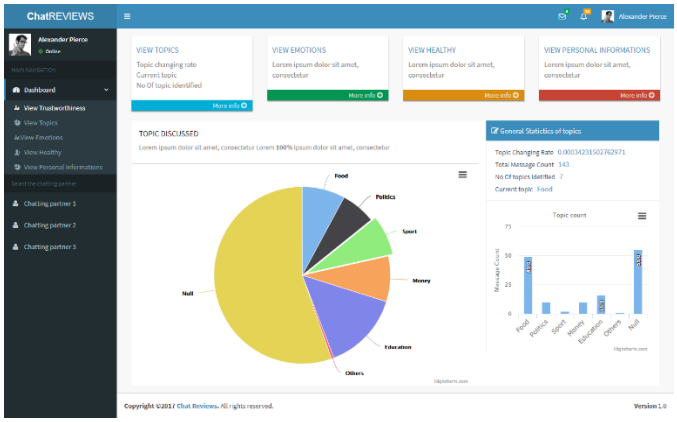

Figure 2 – UI of topic detection module

- User can find topics percentage discussed in the particular chat session in the aim of analyzing trustworthiness of its partner.

2.3 Research Findings

In this research study we came across many new findings. One of them is implementing **Bernoulli document model** (One major document model used for document classification in the world most sophisticated web applications like Gmail, Yahoo for their email classifications) for Instant messages analysis in the aim of detecting topic. And Developing the **Bernoulli document model** with 466,544 no of English Vocabulary was huge challenge since the length of vocabulary affects to execution time of the algorithm.

Also for the Topic detection module we had to find training data (Instant Messages) for each Topic Class separately. Then Likelihoods were generated from collected training data and stored in database (Oracle Database 11g) for purpose of reading it by algorithm developed using Java and reduced the execution time of the whole module greatly.

Furthermore we can find out messages not related to any of the topic in the system from the algorithm improved with confidence is also a new improvement of the **Bernoulli document model.**

3 RESULT AND DISCUSSION

3.1 Results of the Topic Detection Module

The following subsection provides evidence to the implementation results and the solutions provided for the identified research problems. The main user interface of the "Topic Detection" is shown with respective results. So that all the interfaces were designed with a simple, attractive manner to keep the continuous user interaction.

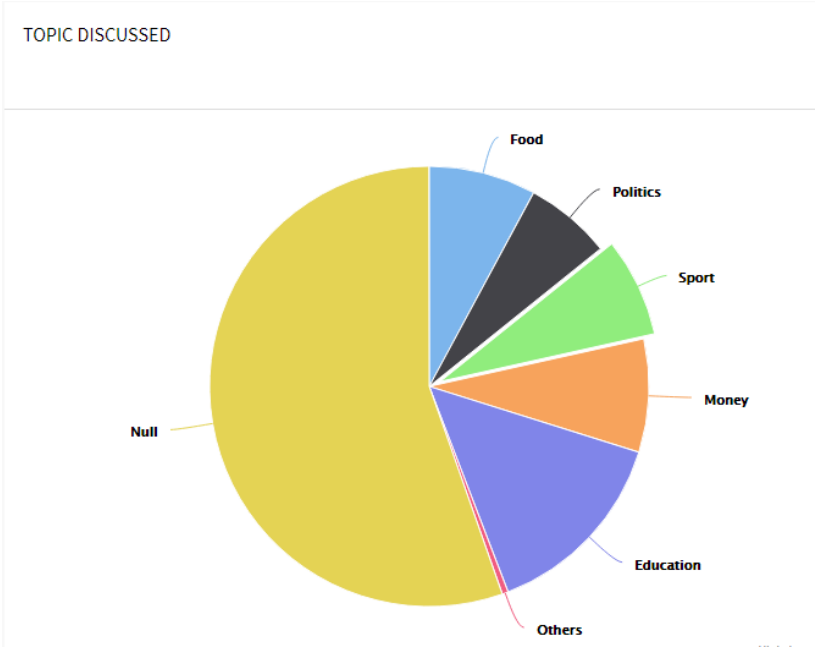

Figure 3 – Review of result of detection module

Results computed by Algorithm:

```
run:
Food: 1.0
Politics: 2.0
Sport: 3.0
Money: 4.0
Education: 5.0
Other: 6.0
Null: 7.0
--------------------------------------------------
--------------------------------------------------
Prints prob array
 1.7546664893333334E-6 1.0 1.0
 2.5054812337481486E-6 2.0 1.0
 3.9360000492E-6 3.0 1.0
 1.9833331270666668E-6 4.0 1.0
 4.840592631997036E-5 5.0 1.0

Prints sorted prob array
 4.840592631997036E-5 5.0 1.0
 3.9360000492E-6 3.0 1.0
 2.5054812337481486E-6 2.0 1.0
 1.9833331270666668E-6 4.0 1.0
 1.7546664893333334E-6 1.0 1.0

topic005
BUILD SUCCESSFUL (total time: 35 seconds)
```

Figure 4 – Result of algorithm

Above output shows the statistics computed for each topics and from last section it shows the sorted probability of each topics. Then top one will be selected as the topic which is having the highest number as shown above. Selected topic is displayed in bottom line as shown above it is "topic005".

For an example;

Message 1: I want to study now

Topic: Education (topic005)

Message 2: I want money

Topic: Money (topic004)

Message 3: He is reading for a degree

Topic: Education (topic005)

3.1.1 Test Cases

Test case #	Description	Pre-conditions	Test Steps	Input	Expected output
01	Messages matching with given topics in the topic detection module	• Messages should be received from Facebook pages	1. Enter Message	Message: I want to eat now	• Message should be classified as "Food", will increment the current percentage of pie chart.
Test case #	Description	Pre-conditions	Test Steps	Input	Expected output
02	Message mismatch with given topic in the topic detection module	• Messages should be received from Facebook pages.	1. Enter Message	Message: Nature is really beautiful	• Message should be classified as "Other", will increment the current percentage of pie chart.
Test case #	Description	Pre-conditions	Test Steps	Input	Expected output

03	Message which is not enough to have a topic.	• Messages should be received from Facebook pages.	1. Enter Message	Message: Hi how are you?	• Message should be classified as "Null", will increment the current percentage of pie chart.

Table 2 - Test cases

3.2 Discussion of the System

As discussed earlier the main target of this application to review a particular chat session and analyze trustworthiness of chatting partner from statistics computed by modules implemented in the system. So by using our system we would provide quality and interesting features for users.

The main system attributes is described as follows,

- Reliability

Reliability is the probability that an application will accurately perform its specified task under stated environment conditions. Simply, that is how much a user can depend on the system. The Purpose of this system is to provide a reliable and efficient service to any person around anywhere who wants to analyze his chatting partner. Other than these, application can be used as a communication media with other users.

- Availability

The specific system outage is not required for the application and system availability is totally dependent on the reliability of the system tools and interface devices. Chat Reviews system will be available at any time with internet connection. If the connection is lost, Application won't be available. User can be access through any device with internet access.

- Security

User will be able to log to the system with username and password which is giving at the registration. Anyone who is having login details can access the system.

- Maintainability

Maintainability is designed as the probability of performing a successful repair action within a given time. In other words, maintainability measures the ease and Speed with which a system can be restored to operational status after a failure occurs. Also the basic thing is in that application falls identification part. Here in every case databases should be maintained properly.

- Privacy

The application can be used by several users after login with username and password. Therefore each user's performance and progress can only be viewed after login to the application using correct username and password.

- Modifiability

The system is able to add new modules with new features. Therefore whole application is built with object oriented and module concepts.

3.2.1 Flow of the Project

- From a research, our team identified problems related to social networks and that's all because of online fake users.

- The project team realized that there is a requirement of a chat monitoring application nowadays that can help people to understand their chatting partner in analytical way.

- When a user deal with the existing systems they got lot of problems. Finally all decided to build a chat monitoring application.

- After the initial discussion we were able to identify the basic problems in current systems and we got a clear idea about some features that were not covered even in places where they were using existing systems.

- All of our members identified the target of the system. We gave much priority to develop a useful and effective chat monitoring system which is better than existing systems.

4 CONCLUSION

And when considering chat monitoring system nowadays most of them just gives basic functionalities. But the proposed system has separate modules called **Topic Detection**, **Emotion extractions**, **Detect Personal Information** and **Evaluate Healthy**. And they provide strong analytical statistics which is used to analyze chat session in the aim of analyzing trustworthiness of chatting partner. Also statistics computed by modules will be displayed in a graphical way for the user in a user friendly manner. Then the user will be able to review result generated by each modules separately easily.

So finally the outcome of our proposed system carries web application with Socializing concepts to analyze our chatting partner. And based on analytical review user will be able to review his chatting partner's characteristics (Trustworthiness). System also support to review multiple chat session by one user then user also be able to compare his chatting partners in analytical way. Analyzing multiple chatting partners is an outstanding feature given by Chat Reviews.

Finally by using this chat monitoring system user can be aware his chatting partner and can be safe from unnecessary things like "giving money for a person never met or seen before" upon a request made by chatting partner. For the safe chatting in social networks or any other chatting application we recommend to use **Chat Reviews**.

REFERENCES

[1] Messenger.msn.com, 'MSN Messenger', 2014. [Online]. Available: http://messenger.msn.com/. [Accessed 13- July- 2017].

[2] Messenger.yahoo.com, 'Yahoo Messenger. Available', 2014. [Online]. Available: http://messenger.yahoo.com/. [Accessed: 13- July- 2017].

[3] Expertsystem.com, 'social-media-monitoring', 2016. [Online]. Available: http://www.expertsystem.com/solutions/corporate-intelligence/social-media-monitoring/. [Accessed 13- July- 2017].

[4] Expertsystem.com, 'Natural Language Processing And Text mining', 2016. [Online]. Available: http://www.expertsystem.com/natural-language-processing-and-text-mining/. [Accessed 13- July- 2017].

[5] Statsbot.co, 'Where analitics happens', 2017. [Online]. Available: https://statsbot.co/?utm_source=blog&utm_medium=post&utm_campaign=text_classifier. [Accessed 13- July- 2017].

[6] Statsbot.co, 'Data Scientist Resume Projects', 2017. [Online]. Available: https://blog.statsbot.co/data-scientist-resume-projects-806a74388ae6. [Accessed 13- July- 2017].

[7] K. Tzeras and S. Hartman, "Automatic indexing based on Bayesian inference networks". In: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), pp. 22-34, 1993.

[8] Mattshomepage.com, 'Bernoulli Naive Bayes Classifier', 2014. [Online]. Available: http://mattshomepage.com/articles/2016/Jun/07/bernoulli_nb/. [Accessed: 20- July- 2017].

[9] Wikipedia.org, 'Supervied learning overview', 2014. [Online]. Available: https://en.wikipedia.org/wiki/Supervised_learning. [Accessed: 19- July- 2017].

[10] Wikipedia.org, 'Bag-of-words model', 2014. [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model. [Accessed: 19- July- 2017].

[11] Wikipedia.org, 'Feature Vector for pattern recognition', 2014. [Online]. Available : https://en.wikipedia.org/wiki/Feature_vector. [Accessed: 14-April- 2017].

[12] Mysql.com, 'Mysql server', 2017. [Online]. Available : https://www.mysql.com/. [Accessed: 14-April- 2017].

APPENDICES

- Database - is a structured collection of data which is managed to meet the needs of a community of users
- PHP - is a server-side scripting language designed for web development but also used as a general-purpose programming language
- Jquery – Java script Library