ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

An Overview and Method for Topic Extraction and Sentimental Analysis of Chats

NainySewaney, Hitesh Somani, AdityaThorat, SudhanshuTiwari, CharusheelaNehete

Student, Dept. of Computer Engineering, V.E.S. Institute of Technology, Chembur, Mumbai

Student, Dept. of Computer Engineering, V.E.S. Institute of Technology, Chembur, Mumbai

Student, Dept. of Computer Engineering, V.E.S. Institute of Technology, Chembur, Mumbai

Student, Dept. of Computer Engineering, V.E.S. Institute of Technology, Chembur, Mumbai

Professor, Dept. of I.T, V.E.S. Institute of Technology, Chembur, Mumbai

ABSTRACT: Chat environments pose a challenge to knowledge extraction and data mining due to their very unstructured format and usage of a large variety of slang. In this paper we present a method to extract and derive useful information from a chat session and also perform the sentiment analysis of that chat with respect to the topic being discussed. This provides information on whether the chat session was sentimentally positive towards a topic and has several possible applications in analytics.

KEYWORDS: Sentiment analysis, chats, topic extraction, keyword based analysis

I. RELATED WORK

"An approach towards comprehensive sentimental data analysis and opinion mining" [1] It presented comprehensive sentimental analysis which included parsing and scoring. Parsing is done using Stanford Parser and words are scored depending upon if the word is adjective or noun. If still not found then its synonyms are checked and their average score is taken into account. On the contrary if the word is found then its score is evaluated considering any implications if present. They developed a web-based sentiment analysis system that attempts to automatically extract the relevant features of anything to be analyzed, and summarize the sentiments corresponding to each feature, in a set of positive and negative points.

In the paper "Extracting Product Features from Online Reviews for Sentimental Analysis" [3] authors described the extraction of title and opinion words from feedback of products. Styles of writing are focused upon for extraction of title and opinion words. By doing this high frequency patterns are found. These help in establishing objective sentiment. For establishing subjective sentiment steps such as pre-processing, pattern extraction and pattern matching are followed. Initially, the sentences are divided into sub segments and POS tags are found. The system will calculate the probability for each pattern of POS tags identifying a subjective sentence and choose some effective patterns to create the pattern set. Pattern matching uses the pattern extracted to extract the features and opinion words. The system achieves about 70% precision and 40% recall for title extraction, and about 90% precision and 60% recall for opinion words extraction.

"Fuzzy Logic Based Sentiment Analysis of Product Review Documents" [4] discussed extraction of online reviews which thus aims at the creation of an application using fuzzy logic for sentimental analysis at document level. The proposed system uses: Tree bank model and Fuzzy opinion mining model. The phases involved in fuzzy opinion mining: Pre-processing Phase which involves removal of spelling mistakes or punctuation errors. The feature extraction phase involves POS tagger which tags the given sentence. Frequently occurring multi-word features are extracted based on Tf-idf calculation. . Feature classification phase includes Extract associated descriptors and hedges from the review, identify the polarity and initial value of the feature descriptors based on SentiWordNet score and calculate overall sentiment score using fuzzy functions to incorporate the effect of linguistic hedges. Fuzzy score calculation SentiWord Net score is considered the initial fuzzy score, if there is a preceding hedge then its fuzzy score changes. The accuracy of the system depends on representation of input data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

"Analyzing Sentimental Influence of Posts on Social Networks" [2] describes analyzing sentimental effects of influential posts. A dictionary of emoticons and frequently emotional texts is created. A method is also devised to judge the emotional impact of a post by studying its sentiments. The sentimental influence is also judged by the type of issue i.e. public or personal. Public issue is influenced by external sources like current news. Impact of individual posts can be judged by the number of replies it gets. Emotion detection which basically consists of building emotion dictionary. Emotion dictionary is made up of emotional text and emoticons. Sentimental analysis is then performed and a graph model of posts is built. Sentimental influence can be calculated on the node of the graph and its descendants i.e. the influencer and its descendants. The paper demonstrates the effect of user posts on the audience at large and on other posts.

"Topic detection and extraction from chat" [5] focuses on identifying and extracting the topic being discussed on a chat application by an individual. The paper makes extensive use of vector space model. The conversations are considered as threads. Conversational threads are recovered by using a graph model of posts without relying on metadata. This is done by constructing a connectivity matrix and identifying parent-child relationships between posts. Relevant threads are extracted using an algorithm. This paper compares various techniques like TF-IDF with time distance penalization, Hypernym augmentation and Nickname augmentation. By conducting experiments on a particular dataset we can conclude that TF-IDF with time distance penalization works with great precision.

II. INTRODUCTION

Chats have become a very common mode of communication. Starting from simple online chats to start off, they have now reached a point where we use applications on our mobile devices, such as whatsapp, hike, or online social media such as Facebook, for everyday conversations. They have to an extent replaced old telephonic conversations. Such large usage of chats leads to generation of a large amount of data. This data contains a lot of information but it is difficult to extract it due to the unstructured and dynamic nature of chats.

III. DESIGN FOR TOPIC EXTRACTION

A possible implementation for topic extraction is discussed below with some assumptions with respect to the input messages. The design basically takes chats as input, processes them using the methods we describe below and stores the information in a database. This database can then be accessed via an analytics software or through a programming language.

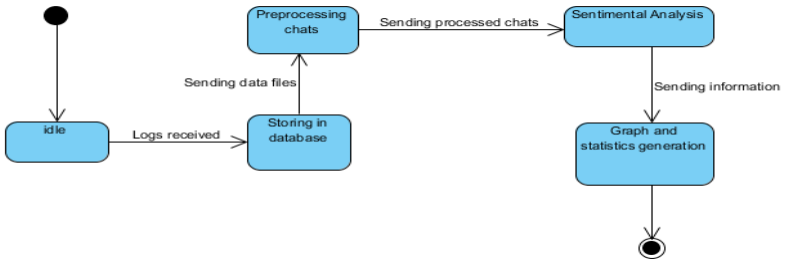

fig.: States Transition Diagram for topic extraction system

Assumptions:

We make the following assumptions for our analysis:

The chat session is between two users. This comes into play in determining weights of the keywords. The weight calculation method would differ for a larger number of users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Chat is to be in English language only. That means no mixed language messaging will be handled. Specific translation management for mixed language messages would be required with knowledge of the languages known by the communicators if mixed language analysis has to be done.

We will also use freely available dictionaries to analyze meanings and other attributes of word as and when needed. One freely available dictionary for this purpose is WordNet.

Approach:

The approach is to first extract keywords from chats to figure out the topic of the conversation and then perform analysis of chats grouped by these keywords. Now it is highly probable that the keyword which is repeated most often between the two users is the topic of conversation. Thus we must find a way to determine which keywords are actually part of the discussion and those which are just there, having nothing to do with the topic. The sentiment value of those topics can be calculated by using a probabilistic algorithm.

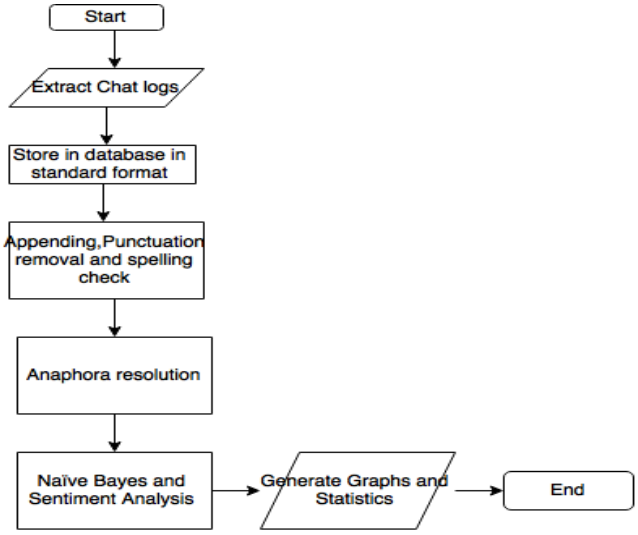

Fig.: System Flow chart

IV. PRE-PROCESSING

As mentioned earlier, the data obtained from chats has in general very little structure. What little structure is observable is in general based on timestamps. There are a lot of differences when we compare chats to standard written english paragraphs. We mention some of them below, and attempt to reduce their effects as much as possible so that we can obtain a modified text form that is simpler to analyzeusing algorithms.

A. *Appending of messages*

It often happens that a user sends multiple messages consecutively without the other user interrupting in a short period of time. It is very likely that these messages are related in some way to each other and belong to the same point being made by the person. So all such messages (that came from one user and with similar timestamp and sent to the same person) are appended together into a single message.

B. *Slang Language Correction:*

Chats can involve a lot of variations in spelling. For instance various abbreviations are used such as LOL (for laugh out loud), l8r/ltr (for later) etc. We can consider commonly used abbreviations such as LOL as easy to read and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

understand for most people and hence it is possible to analyze them by replacing their meaning in the sentence. However, analyzing more obscure terms such as l8r/ltr (someone could translate this as litre, as in the volume unit) is hard to make sense of unless you have past knowledge of their usage and know their meaning. Therefore, to analyze chats you would need to keep a track of possible abbreviations. The algorithm recognizes such slang words and replaces them with their meanings. The words are stored in database with their expanded forms and sentimental value.

C. *Emoticon Detection*

Emoticon contain a lot of sentimental information. They are very important in recognizing the sentiment of the user towards a particular topic. Each emoticon with its meaning, its sentimental value and its text version is stored in the database. The algorithm detects an emoticon and replaces it with its text equivalent. A simple example could be a laughing smiley that show that the person liked the joke that was made in the last message.

D. *Spelling Correction*

As mentioned earlier, we will be using dictionaries for reference. The algorithm will verify each word with the WordNet dictionary and if any word is found with no such spelling then it will be replaced with a word which is most similar to the misspelled word. Even after that if some words remain whose spelling does not exist in WordNet then it is left unchanged assuming it to be a name of some person, place or object (i.e a noun, which generally are common keywords) or a new abbreviation.

E. *Anaphora Resolution*

Anaphora or pronoun resolution involves replacing pronouns with the noun preceding the pronoun or an earlier reference to it. This process can help in simplification of the analysis process.

.

These steps altogether constitute the pre-processing of chat. Now chats are ready for topic extraction.

V. DETERMINING WEIGHTS OF KEYWORDS

Weight is a measurable value assigned to either any word (noun, adjectives, adverbs, verbs) or emoticon. All the messages are checked for their part of speech with the help of WordNet and only nouns are taken as keywords. All the adjectives or adverbs or verbs are considered as helping words because they help a noun to get a special meaning.

Normal keywords carry a value of 1, helping words carry a weight of 2 and all emoticons carry a weight of 4. Now the first chat of that chat session is taken and following steps are followed:

1. If any keyword is found again in that chat then its weight is increased by 1. Simultaneously the same keyword is searched for in the next 10 chats. If the keyword is used by the same user that originally sent the message the its weight is incremented by 1 and if it used by the other user then its weight is incremented by 0.5. If that keyword appears with some verb or adjective then its weight is multiplied by the sum of weights of all the verbs and adverbs and adjectives in that chat. If some keywords appear with an emoticon(s) then the weight is multiplied by the weight of emoticon specified in the database.
2. Step 1 is repeated for all the messages in that chat session(10 - 20 messages). After that the weights of the keywords are summed.
3. After step 2 we will get all the keywords with their weights. Note that there should not be multiple copies of the same keyword for the same user. But the two users can have common keywords with different weights.
4. Now both users have their keywords and their corresponding weights ready. We calculate a threshold value which is given as:

   Threshold (ui) = (Sum(keywords weight of user(ui))/no. of keywords of user(ui))

   where ui = user i.
5. All those keywords whose weights are above threshold are classified as final keywords. Those keywords are likely to be the topic of chat between two users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

6. Now further keywords can be classified using naive Bayes classifier. Naive Bayes classification can classify our keywords according to our needs. If any of those final keyword have remained after classification it means that they are highly probable to be the topic of the chat.
7. The sentiment value of the whole chat session is calculated as

Sentiment value = (Count of (positive words of emoticon) / Count of(Total words of emoticon)) - (Count of (negative words of emoticon)/Count of(Total words of emoticon))

VI . SIMULATION AND RESULTS

Below is a liberally translated chat session that we had earlier. The messages have been preserved in the form they were sent in order to keep the discussion as natural as possible even after translation. The emoticons have been replaced with words. Punctuations have been kept as and where they were. The message have been numbered for understanding purposes.

Table I :Demo Chat Session

User A	User B
1)What have you done the whole day?	
	2.) Filled the exam form. Took DD. Network programming in C very difficult socket programming studied
3) Network programming (astonished)*	
	4) (laugh)*
5) Where did this come from ?	
	6) I was studying for Tejus Networks
7) Tejus	
	8) Tejus networks
9) What is that	
	10) Its a company that will be coming only eligibility criteria should be 6 cgpi

* denotes use of emoticons

The chat discussion can be summarized as follows. A asks about what B did the whole day. B replies what he did. A is surprised by one of B's responses (network programming in C) and then asks why he started doing that suddenly. B explains that he has been studying that as preparation for a company that will be visiting the campus. He also specifies the eligibility criteria for it. So we can conclude broadly that the discussion was about something in network programming and about Tejus.

As mentioned earlier, helping words carry a factor of 2 , emoticons 4.We perform the analysis of the chat session by the weight determination steps mentioned earlier. Numbers in the keyword section denote message numbers as numbered above in the chat.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

Table II :Analysis for user A

Helping words (A)	Keyword (A)	Weight (A)
Whole, done	1.day	1*(2+2)= 4
Astonished	3.network	1*4 = 4
Astonished	3. programming	1*4 = 4
-	7. tejus	1
Total :	4	13

Table III : Analysis for user B

Helping words (B)	Keyword (B)	Weight (B)
Study, difficult	2.form	1*(2+2) = 4
Study, difficult	2.DD	1*(2+2) = 4
Study, difficult	2. Network	(1 + 0.5 + 1 + 1) * (2 + 2) = 14
Study, difficult	2. Programming	(1+1+0.5)*(2+2)=10
Study, difficult	2. C	1*(2+2) = 4
Study, difficult	2. socket	1*(2+2) = 4
Study	6. tejus	(1+ 0.5 +1) *2 = 5
Study	6. networks	(1 + 1) * 2 = 4
-	8. tejus	1
-	8 . network	1
-	10. company	1
-	10. cgpi	1
Total	12	55

Overall network score : 14 + 4 = 18
Overall tejus : 1 + 5 = 6
Remaining keywords only occur once during the entire conversation.
Average value of A's keywords = 13/ 4 = 3 (approximately)
Final keywords for A = network, programming, day (as their weight is more that 3)
Average value of B's keywords = 55/ 12 = 5 (approximately)
Final keywords for B = network , programming , tejus
So the keywords associated with the chat session are: network , programming , day , tejus
Therefore probable topics of this chat include: network, network programming, tejus
Sentiment values of whole chat:
PN=Count(pos.word/emoticons) - Count(neg.word/emoticons)/Count(emoticons)
=(1/2) - (0/2)
=1/2 = 0.5

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

This is almost neutral sentiment (which has value 0).
So the topic of chats can be network, program, day, Tejus.

Assuming we have the following classes for our classification method for chats with keyword programming and network:

Table IV : Naïve Bayes Classes

Networking	Programming	Linux
Network	Program	Linus
IP	language	Torvalds
Address	algorithm	Commands
IPv4	C	Free
IPv6	C++	Open
Linux	Java	Source
Router	Object-oriented	Terminal
Gateway	Procedure-oriented	Ubuntu
Attack	Software	Fedora

Equation for naïve bayes analysis : $P(c|d) = \text{argmax}\{P(d|c)*P(c)\}$ Where, $P(d|c_j)=P(w_1|c_j)*P(w_2|c_j)\ldots P(w_n|c_j)$
$P(w_i|c_j)=(\text{count}(w_{i,},c_j) +1)/\text{Sum}(\text{count}(w,c_j)+|V|)$

where P(c|d) indicates probability of a document belonging to a class

Now,
$P(c_1|d)=P(d|c_1)*P(c_1)$
$=(2/18*1/18*1/18*1/18) *P(c_1) =2/18*P(c_3)*18^{-3} =0.11*P(c_1) *18^{-3}$
$P(c_2|d)=P(d|c_2)*P(c_2)$
$=(1/18*2/18*1/18*1/18) *P(c_2) =2/18*P(c_2)*18^{-3} =0.11*P(c_2) *18^{-3}$
$P(c_3|d)=P(d|c_3)*P(c_3)$
$=(1/18*1/18*1/18*1/18) *P(c_3) =1/18*P(c_3)*18^{-3}=0.055*P(c_3) *18^{-3}$
Assuming P(c) is identical for all classes = 0.33
Hence,
$P(c|d) = \text{argmax}\{P(d|c)*P(c)\}= 0.111*0.33*18^{-3} =0.0363*18^{-3}$
Since there are two topics (Network and program) with maximum probability network and Program were the two topics discussed in the chat session.

On comparison with our manual analysis of the chat we can see that the results deduced are fairly close to what the topic of discussion actually is.

VII. EXTRACTION AND ANALYSIS

We now have information stored regarding the chat messages that we analyzed in a database along with the sentiment values of each chat and the topic associated with each of them. So what can be done with this information? Suppose the chats were analyzed and you figure out that there are many discussions regarding a specific product. We can then use this analyzed collection of discussions for obtaining the general reviews of the people regarding the product. Then we can figure out the likes and dislikes of the people based on the keywords associated alongside the product and the sentiment score of the discussion. The sentiment score by itself will reveal whether the opinion is positive or negative.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 10, October 2015

The information obtained from the earlier step may also be processed further to obtain frequency distributions of various topics and then be used for analysis. The usage of the information that is obtained is only limited by the topics of the discussions and the methods by which the data is transformed and utilized.

VIII. CONCLUSION AND FUTURE WORK

The Sentiment analysis of chats is still an underutilized field.. User privacy is an utmost concern and only public chats can be really analyzed. Application for chat analysis is vast and can be used for various aspects such as product popularity, election results predictions, security, gauging user interests etc. The method discussed to find the topic of discussion is not perfect but provides decent results that can be processed to obtain the required results. It is also fairly general and can be applied to social networking posts/ messages as well with slight variations in calculating method.

1) While the method works fairly well for chat that has been processed properly, preprocessing of chats itself is one of the largest factors affecting chat analysis. Proper preprocessing itself has too many variable factors to be handled easily. While it is possible to create a general analysis method for preprocessing, one really cannot predict the chatting patterns of every possible user in the world. Thus proper preprocessing can affect the results.

2) This analysis method was designed for 2 users. However it can be extended fairly easily for a group of users (by calculating their individual weighted scores for each keyword in the discussion) but the complexity of analysis will rise.

3) An alternate method is instead of calculating sentiment values for the whole session; we can calculate sentiment values for all the filtered final keywords.

4) User privacy: One of the biggest issues with chat sentiment analysis is user privacy. No user feels secure sharing his/her personal data for analysis even if the company assures secrecy. Hence, majority of the possible discussions cannot be really analyzed without user consent. It is fairly understandable that users would not like to reveal their personal data to anyone else. Thus for most parts only publicly declared chat can be analyzed and it is a limitation for chat analysis in general

REFERENCES

1. PoojaKherwa,AjitSachdeva,DhruvMahadeva ,NishitaPande, Prashant Kumar Singh, “An approach towards comprehensive sentimental data analysis and mining”, pp. 602-612 Year 2014, IEEE
2. BeimingSun,Vincent TY Ng, “Analyzing sentimental influence of post on social networks”, pp. 546-551 Year:2014, IEEE
3. HuiSong,Yingxiang Liu, Dao Tao, “Extracting product features from online review for sentimental analysis”, pp. 745-750 Year: 2011, IEEE
4. Indujha K., Raghu Raj, “Fuzzy Logic based Sentiment Analysis of Product Review Documents”, pp. 18-22 Year: 2014, IEEE
5. Paige H. Adams, Craige H. Martel, “Topic Detection and Extraction from Chat”, pp. 581-588 Year:2008, IEEE

BIOGRAPHY

Nainy Sewaney, Hitesh Somani, Aditya Thorat and Sudhanshu Tiwariare Students of Computer Engineering Department, V.E.S Institute of Technology, Chembur, Mumbai.

Mrs.Charusheela Nehete is Professor in the Information Technology Department, V.E.S Institute of Technology, Chembur, Mumbai. They have a keen interest in developing new methods for analysis and application development.