

FRAMEWORK FOR EMOTIONAL ANALYSIS: AN APPROACH OF SOCIAL MEDIA MINING

¹Gowrav Jain, ²Nitin Kumar, ³Manish Singh, ⁴Mr. Vivek,

^{1,2,3}Student, School Of Engineering and Technology, Jain University (SET JU), Bengaluru.

⁴Assistant Professor, School Of Engineering and Technology, Jain University
(SET JU), Bengaluru.

1. INTRODUCTION

Emotion analysis (EA) refers to computational methods to enable machines to recognize or generate emotions in a human manner. Even though emotion and affect are not the same in psychology, studies in affective computing are mostly focused on emotions. Thus, in this thesis, the terms emotion analysis and affective computing are used interchangeably. So are the terms emotion and affective meanings. A multi-dimensional emotion model is defined in the so called emotion space or affective space. The term sentiment is also used to describe people's feelings

Objectives

1. The main objective of the project is to gather the user's data from user's Facebook wall and generate a chart which depict the emotional nature of the user on the social media.
2. The other objective of our project is to transform unstructured data to structured one using predefined dictionary.
3. The clustering and labelling is done to help the users to quickly navigate through his/her interested information.

Problem statement

The major tasks in the problem of enhancing accessibility of microblogging's messages are:

- a) **Microblogging's Message Clustering:** Let $M = m_1, m_2, \dots, m_n$ be a corpus of n microblogging messages. Among these n messages, there are k different topics and we aim to cluster the n messages into k clusters c_1, c_2, \dots, c_k .
- b) **Cluster Labelling:** For each cluster c_i , we aim to generate human readable cluster labels l_1, l_2, \dots, l_k in which the criteria for creating labels are different for different clusters.
- c) **Emotional Analysis:** Emotional analysis of the data is done and a character certificate of the user is generated which gives the rough idea on how the user is portraying himself in the social media.

2. LITERATURE SURVEY

The important step in software project development process is literature survey where the concentration is on existing systems methodology's to determine the time factor and company strength. By using the drawback of existing system we propose a system that can be useful for the future purpose. Before developing the proposed system it is necessary to take all literature survey for consideration to perform further future work.

Brief Overview of existing system

Brendan O Connery[8] has presented a method on text analysis, where two different methods are considered for information access from microbloggings i.e., (a) Message Retrieval, which determines the messages are collected from microblogging services. (b) Opinion Estimation, which determines whether messages are having positive or negative opinions.

Apoorv Agarwal [12] has proposed a scheme for sentiment analysis in microblogging, where detection and summarization of microblogging messages is done by sentiment analysis. This scheme includes two different methods i.e., two-way task and three-way task.

Xiaohua Hu [11] has presented a mechanism for clustering in microblogging. This mechanism was done by using two different methods i.e., exact-match and related-match to map with wiki data for improving the clustering performance. Once the data is collected from twitter, mapping is done with Wikipedia topics and sub-topics. This process is repeated till complete text is transformed to clusters.

3. SYSTEM ANALYSIS

Emotional analysis in microblogging

Emotional analysis is a task of NLP where it tracks the mood of the people about particular topics. It concentrates on developing a system to examine about product details in blogs, comments of microblogging. Figure 1 explain different stages of emotional analysis.

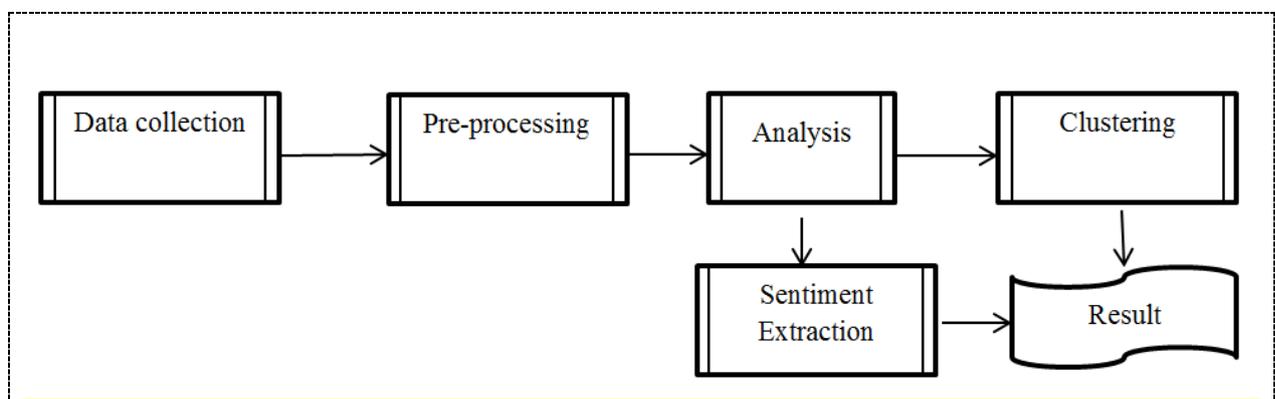


Figure 1

First phase of emotional analysis is data collection, where collection of data is done from different sources like twitter, Facebook and different review web site. After collecting the data, next phase is data pre-processing in which different methods are used to clean the unstructured words. Later, this data is placed into analysis phase. Here, if the data is structured then that is fed to emotion extraction otherwise, it is sent to clustering for further process. Emotion extraction is code part to get opinion of specific product and clustering process can be performed by using Bag-of-word concept.

4. PROPOSED SYSTEM

The various kinds of messages are available in microbloggings, but they are short and unstructured causing difficulty to analyse microblogging messages. By using our proposed system it

iseasy to gauge his/her interested messages from microbloggings. In proposed system following functions can be performed

- Describe the service and messages of microblogging site, such as twitter which contains message acronyms(ex: gr8=great) that allows the users to post real time messages with acronyms because microbloggings are restricted to character length (ex: twitter=140 character length).
- Input messages are unstructured in nature so the proposed system adopts a scheme called pre-processing. It is a scheme which changes the unstructured word to meaningful English word by referring to pre-defined dictionary.
- Cluster is represented using a word i.e., a label. The criteria for creating label depends upon clusters.

5. SYSTEM DESIGN

System design stage is called as stepping stone for deriving the solution for given problem domain. System design is used to identify the modules of the system, and collect the specification for detailed development of these modules to obtain the expected results. System design is helpful in process modelling. A new system can be created by using this approach. System design is a transition from users view to programmers view. It provides a communication bridge between the implementation phase and the requirement specification. High level design and Low level design is the classification of system design phase.

System Architecture

The model which defines behavior, structure and other system views conceptually is termed as System architecture. The representation of system and formal description is organized in a way that structure of the system is supported. The component properties which are visible externally, the system components and the relationship or behavior between these components is comprised by system architecture. It gives an arrangement which could be obtained and forms the frameworks. This gives one of the good frameworks for the system to form the architecture. The system architecture phases are:

1. **Syntactic decomposition:** Collect data in the form of document and it is divided into sentences.
2. **Pre-processing:** Change unstructured word to structured word by referring to pre-defined dictionary. Remove extra space and dots. This phase is helpful for next phase i.e., Clustering
3. **Clustering:** The process of grouping the related sentences is called clustering. Here, collected input sentences are segregated into groups of Named Entity Recogniser based on some condition.
4. **Label:** Word which represents the group. Criteria for creating labels depend upon cluster.

User sequence diagram

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a message sequence chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

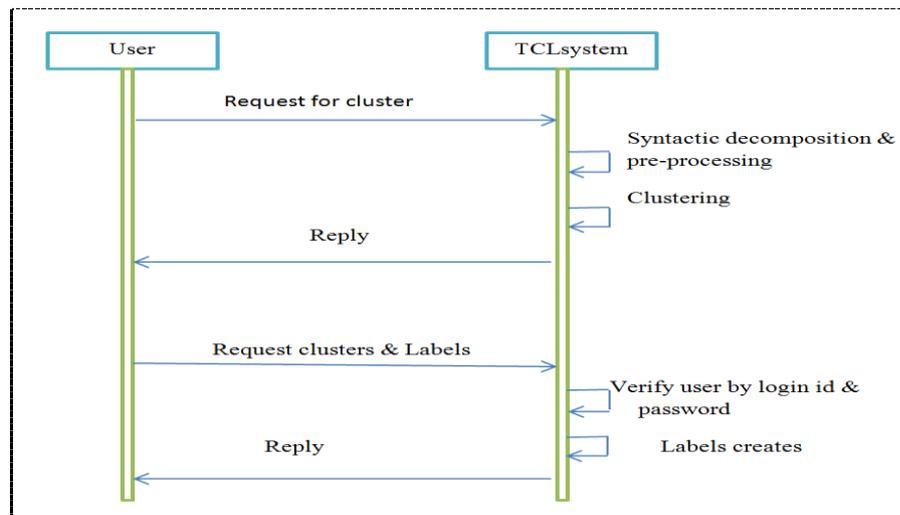


Figure 2: User sequence diagram

The figure 2 explains the sequence diagram, where it explains the clusters and labels of microblogging messages in a sequential manner. In this scenario, we have two classes named as user and Text clustering and labelling (TCL) system, where the communication is through objects. Initial step, users have microblogging messages, then user will request TCL System to convert microblogging message to cluster. TCL system undergoes several processes to get clusters data and it will reply back to the users. Again user requests for label to each cluster. Finally TCL system undergoes some process and reply back with clusters and cluster labels.

6. SYSTEM IMPLEMENTATION

Implementation is a one of the main phase in project development process. All the things which we are concerned till now like requirements, design, coding makes a system to develop this project. When we come to this phase, we will exactly know how our project works. It deals with many of the concepts to get a perfect result.

We have considered different algorithms, programming languages and with software. In this implementation stage, the theoretical design of system is turned into a working project. This is most critical part of achieving success in the project.

Functional Description of Modules

- **Data Extraction:** Collect the data from microblogging.
- **Syntactic decomposition:** Divide the document into sentences.
- **Pre-processing:** Transform the unstructured word into structured one.
- **Clustering:** Grouping the related sentences using NER concept.
- **Labels:** Word which refers to cluster.

Data Extraction

Purpose: Collect the data from microblogging.

Input: Collect the individual account data.

Output: Collected data for input content.

Functionality: In the data collection module, collected data from Facebook or twitter is stored in input.txt file, which will be helpful for the next phase.

Syntactic Decomposition & Pre-processing

Purpose: Generating pre-processed text.

Input: Data stored in input.txt file is input for this module

Output: Collecting the pre-processed data in pre-processed.txt.

Functionality: In the Syntactic decomposition & pre-processing module, clean extra spaces, extra dots and perform some set of operation to transform the unstructured word to structure by referring pre-defined dictionary.

Pseudo code:

Pseudo code 1 : Syntactic decomposition & Pre-processing

Precondition : Special symbol are pre-define dictionary words;

Input : Collected microblogging messages;

Output : Pre-processed microblogging messages;

Description :

1. *Fetch Documents*
 2. *for sentence in Document*
 3. *Parse words in sentence*
 4. *If special symbols exists*
 5. *Replace*
 6. *End if*
-

Labelling

Purpose: Creating cluster labels.

Input: Clusters in data is input for this module

Output: Generating the cluster with appropriate labels.

Functionality: In the labelling module, Depending upon the clusters different pseudo codes are used.

Pseudo code for generating label in Person cluster:

Pseudo code 2 : Label formation in Emotion cluster

Precondition : EmotionWord refers as token;

Input : Cluster in microblogging messages;

Output : Cluster and cluster label in microblogging messages;

Description :

1. *for* token in Emotion -cluster
 2. *label* = token
 3. *for* sentence in document
 4. *if* sentence contain label
 5. Add sentence to label
 6. *else*
 7. *skip*
 8. *end if*
-

CONCLUSION AND FUTURE ENHANCEMENT

The uses of microblogging sites are loaded with large amount of microblogging messages. It is very hard for users to go through his/her messages of their interest.

Since the messages generated from microblogging sites were not in proper format, we had collected the microblogging messages and did the clustering and labelling. We can try to generate large volume of messages from microblogging sites and we can apply our work to those messages to help users for going through their interested information.

REFERENCES

- [1] Xia Hu, Lei Tang, Huan Liu, “Embracing InformationExplosion without hoking:Clustering and Labeling in Microblogging”, IEEE TRANSACTIONS ON BIG DATA, January 2015.
- [2] Logamani.KandPunitha. S. C , “Density Based Clustering using Enhanced KD Tree”, in International Journal of Computer ScienceEngineering and Technology(IJCSET) , November 2014 | Vol 4, Issue 11,314-318.
- [3] Bharat Chaudhari, Manan Parikh, “A Comparative Study of clustering algorithms Using weka tools “International Journal of Application or Innovation in Engineering & Management (IJAEM)Volume 1, Issue 2, October 2012.
- [4] Rami Al –Rfou’StevenSkiena,“SpeedReadA Fast Named Entity RecognitionPipeline”Proceedings of COLING 2012: Technical Papers, pages 51–66,COLING 2012, Mumbai, December 2012.
- [5]David Nadeau, Satoshi Sekine,“A survey of named entity recognition and classification” Higher Education Press and Springer-Verlag Berlin Heidelberg, 2012.
- [6] Y. Hu, A. John, F. Wang, and S. Kambhampati, —Et-lda: Joint topic modeling for aligning events and their twitter feedback.¶ in AAAI, vol. 12, 2012, pp. 59–65.
- [7] YoungjoongKo E., “Automatic Text Categorization by Unsupervised Learning”, in Fourth ACM International Conference on Web Search and Data Mining (WSDM), 2011.
- [8] Brendan O’Connor, RamnathBalasubramanyan, Bryan R. Routledge Noah A. Smith, “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series”, Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

[9] L. Hong and B. D. Davison, “Empirical study of topic modelling in twitter”, in Proceedings of the First Workshop on Social Media Analytics”. ACM, 2010, pp. 80–88.

[10] D. Ramage, S. Dumais, and D. Liebling, “Characterizing microblogs with topic models”, in International AAAI Conference on Weblogs and Social Media, vol. 5, no. 4, 2010, pp. 130–137.

[11] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E.K. Park, Xiaohua Zhou, “Learning to Cluster Web Search Results”, *KDD '09*, June 28-July 1, 2009, Paris, France. Copyright 2009 ACM 978-1-60558-495-9/09/0.

[12] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, “Sentiment Analysis of Twitter Data”, Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March 2009.