Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 4; July-September, 2017, pp. 263-268

http://www.krishisanskriti.org

Behavior Analysis over Text using Text Mining Ontology Development of Emotion Analysis and Identification

Vaibhav Nagpal[1], Tanya Bhattacharya[2] and Rishi Kumar[3]

[1,2,3]Department of Computer Science, Amity University Uttar Pradesh, Noida-201301, India
E-mail: [1]vaibhav.nagpal@student.amity.edu, [2]tanya.bhattacharya2@student.amity.edu, [3]rishikumar24@amity.edu

Abstract—*Emotion can be articulated in various forms such as textual data, audio, gestures and which can be seen as in facial expression. Further architecture is proposed for detection of emotion from textual data. Emotion detection from textual data have attracted various researchers because of the wide range of application. The survey conducted among various professionals and of different age group to interpret the emotion of the text. This paper also presents the datasets used in analysis of emotion. The emotion detector algorithm is proposed to find the major emotion in given text or data set. Apart from this emotion analysis has been implemented widely in various social media and chat area of application. All these areas are covered in this paper.*

Keyword*: Emotion analysis, ontology, lexical affinity, semantic analysis, mining.*

1. INTRODUCTION

Researchers have proposed that it is very difficult to examine the behaviour of any individual while communicating over text messaging. Text analysis is a research methodology which examine the textual content rather than the structure of the content of the text. This methodology is very vital in the field of research as it avows the researchers to examine the meaning and idea of the text. The text analysis generally includes three factors-why the text was written? at what situation it is written and the audience for the text which includes both senders and receivers. Other than that a specific framework must be made to interpret the text. There are various ways to analysis the text for example to note down the number of times certain phrases or words are used in text, examine the writing style of the author's narration and interpret the meaning of the text. Text analysis has its root of analysis of mass communication from mid 1950s. Text analysis have been used in variety of fields including anthropology, management, political science, LIS (library and information studies),psychology and many more. There are two approaches for text analysis qualitative and quantitative approach as both are used in information studies.

2. RELATED WORKS

Various techniques and methods have been developed for textual analysis among them data mining is the techniques which is widely used to analyse emotion over text.

Leshed et al. (2006) opted classification of emotion on the data cluster obtained from LiveJournal blog posts as the emotion text collection. In this method 50 topmost emotions appearing in the blog posts were taken for analysis. In this support vector machine fused with bag of word model have been used to assign categories of emotion in the blogpost. This method resulted into the accuracy report of 78%.

Mihalcea et al. (2006) considered a collection which was based on the approach for classifying the blogpost from LiveJournal into two categories "happy" and "sad". Bigram feature with naïve Bayes classifier have been used to for classification of emotion. This system reported as the 79.13% accuracy.

F. R. Chaumartin et al. (2007) proposed another such method which imitate rule-based approach against classification of emotions. They word on the data cluster obtained from the news headline for analysis over text. WordNet and target speech have been used for decapitalising he common words in the initial pre-processing step. Each word is classified into emotion classes. Parsing of headline have been used to detect the main theme word, which was rated as higher weight as compared to other textual data. The emotion score given to the particular word occurred on the basis of the coherent relation to the categories in the WordNet. They also considered various other factors for scoring the words such as high tech names, person will, modals, negation and celebs, etc. It was found that 89.43%, 27.56% and 5.69% are the average accuracy, preciseness and recall of this system respectively.

Strapparava et al. (2007) proposed the comprehensive outcome of the three systems that contributed in SemEval (2007). The three systems were rule based, supervised and unsupervised systems. It was seen that outcome of rule based system was reported best in 4 out of 6 emotion classes and supervised between the rest two.

Lin and Chen et al. (2008) proposed the method of sequencing the author's emotions in Chinese news column from a search engine yahoo. They classified the emotions into eight categories or classes for this work they opted the SVM method that is support vector machine for classifying. They used various features for their work as Chinese characters diagram, Chinese news metadata, Chinese characters, and text emotion. This system has been reported as 76.88% accuracy.

Strapparava and Mihalcea et al. (2008) proposed another algorithm that inquiry the presence of significant main words in the headlines and enumerate scores of each word that will reflect the frequency of the word in this affective lexicon in the text using WordNet.

Bellegarda et al. (2010) proposed the comparison between the SVM system and naïve Bayes system. They proposed that support vector machines are much better as compared to naïve Bayes system infor emotion analysis. News headlines and blogs were used as a data cluster for emotion resolution.

Balahur et al. (2011) proposed a method using EmotiNet-resource that has been used to detect the emotion over text dependent on the common sense knowledge on concepts, their communication and their repercussion- to identify emotion. EmotiNet portraits situations as a string of actions and their equivalent emotional result adopting ontological representational data.

Balabantaray et al. (2012) proposed an emotion classifier model that determines the emotion into various classes of the person writing. SVM kernel was used for classifying emotion according to the emotion classification determined by Ekman, 1999.

Sykora et al. (2013) also proposed a method based on the use of ontology approach to elucidate the problem of blended in emotion detection in the text. This approach identifies a range of eight different high level emotions-adness, happiness, fear, excitement, shame, anger, confusion and disgust.

Calvo et al. (2013) proposed separate categories wised approaches on the basis of SVM with reduction methods: PLSA (probabilistic Latent Semantic Analysis), LSA (Latent Semantic Analysis and NMF (Non-negative Factorization). This was done to analysis the efficiency of above three techniques and it was concluded that non-negative factorization (NMF) was the best approach to classification of emotion.

3. RESEARCH METHODOLOGY

1. Emotion over text

In this 21st century speeding, economical, easily accessible and highly technical advanced communication tools like phones, laptops, tablets are accepted as the socially communication devices. As telephonic communication is the most convenient and effective way of connecting but we cannot slip up its limitations and drawbacks too. As text messaging fails to express the nonverbal communication part which leads to misinterpretation of the emotions of the both sender and receiver. It is very difficult to interpret the emotion i.e happy, sad, anger, excited, frustrated, disgust etc., exclusively on the Basis of text. This has been found that people who communicate regularly may lead to risk of interpreting emotion over text. Text messaging has been evolved a lot in these recent years as changes are necessary to understand the emotions of the text message. Now a days emotions are partially interpreted through body gestured i.e. smileys and through voice messaging.

2. Quantitative text and Qualitative text analysis

Quantitative text approach includes following procedures-

i) Create a hypotheses to analyze the content on the basis of related research and significant theory.

ii) Determining the nature of data for testing the hypotheses.

iii) Determining the sample unit and the method. This is known as sampling as it generalizes from the specific to the general.in this a sample is used which symbolizes the whole domain.

iv) Developing the coding schema

v) Analyze the coded information

Quantitative text approach includes following procedures-

i) Categorizing the data

ii) Identify the frame-work

iii) Sorting of data into the frame-work

iv) Developing the coding schema

v) Analyze the coded information

3. Approach for analysing text emotion

3.1 The approach for analysing text emotion is categorized into three types: keyword spotting, lexical affinity technique, machine learning based technique.

3.1.1 Keyword spotting

This is the most popular and naïve approach because of its accessible to use and economy is finding keyword from predefined emotion vocabulary. In that vocabulary words are categorised into sad, happy, anger, surprised, etc. This method includes 5 steps for recognising emotions-Firstly, the text is converted into tokens this process is known is tokenization.

In the second step, identification of emotions words is done from our redefined vocabulary.In the third step, intensity of each emotion is analysed. In the fourth step, negation check is done and the step is our resultant emotion came from the text.

3.1.2 Method of lexical affinity

This method is more sophisticated and extensive version of keyword spotting. This method assigns the probabilistic "affinity" to a selective word rather than finding the emotion word. This method have some limitations such as firstly this method can be tricked by the other sense of the respective word and secondly this method is biased towards a particular word genre. For example a word "accident" in the given the highest probability of negative affect in the phrase used in "car accident" but will contribute incorrectly in the phrases like "I met my ex-girlfriend by accident this morning".

3.1.3 Statistical approach method

This learning based method is used to frame the problem differently. Above two approaches determines the emotions of the given text but now the concern is to collocate the given text among different emotions. Unlike keyword spotting method and lexical affinity method this implies various theories of machine learning like conditional random fields and support vector machine to detect emotions relied on previously trained classifier in order to distinguish in which category the given text belongs.

3.2 Comparison of techniques

From above three techniques key-word spotting technique is the most easily implemented and the result we obtain is quiet accurate. This technique has some limitations: it is not applicable on the large range of domain in which approach is ontology based, obtaining the content of expression lexicon is one-sided. We have to use common sense and improvise recall value but the determination of emotion ontology is very difficult and time draining. Therefore, feature of machine learning algorithm is used as a lexical resources. In case of statistical approaches, the supervised learning is most widely used in detection of emotion because it gives improved results as compared to other un-supervised learning methods. Although these approaches want annotation of each labelling example which is very time draining task. Because of this reason many researcher have determined this task automatically when system process Twitter text, annotation of text can be done through hashtags or emotion it includes. Although results obtained in unsupervised learning were worse than obtained in supervised. In the end it is found that supervised learning approach is more enhanced in detection of emotion. As it is capable of detecting indirect reference of the text.

4. SEMANTIC ANALYSIS OF TEXT

The rapid growth of internet had augmented the pace and measure of information and data circulation. It is quite difficult to search the best suited and relevant data using traditional IR approach in this widely spread ocean of data. Now a days user is overwhelmed by the keyboard-based search which gives thousands of results by hitting in just seconds. In this world people wants want relevant information as quickly as possible so various technologies and approaches are done to fetch the relevant information efficiently.

5. DATA CLUSTER

In this section, we will altercate data clusters that have been discovered for analysis of emotion we divide them into two categories: short text and long text.

5.1 Short Text

Most of the work on analysis of emotion over text have been done on the short text. One of the reason of using short or small text is that huge or longer text acquires large variety of emotions which is hard to express or detect. On the other hand short text as compare to longer be straight and laconic in emotion expression. In short text analysis mainly the datasets used is the phrases of news headlines. The popular technique used in microblogs analysis is Hashtag-based supervision. In this type of technique hashtags used by the author of a tweet is treated as a label. For example, take a tweet as '#sad' it will be considered as sad tweet.

5.2 Long Text

Although long text is difficult to analyse but some prior works have been done on long text for emotion analysis. Among various works on long text the most prominent works are liu(2003), who made the datasets using emails. Alm et.al (2008) who made the datasets using children's stories. But rather than taking the whole story a one big dataset, they divided the whole text into subtext and each text was annotated as single sentence individually. Kang et al. (2016) were the few researchers who used languages other than the English as their dataset. This is relevant because expressions and emotions are not confined to one language it varies across different language.

6. ARCHITECTURE

6.1 Architecture used for emotion analysis:

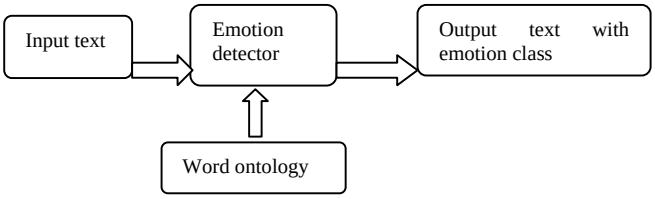
i) Obtain resourceful data cluster

ii) Select your goal

iii) Obtaining I/P from computational linguistics and c linical psychology

iv) Deployment of classifier model

The above are refashioned to extend their performance and capabilities for which a simplified and understandable model is designed as follows:



The above architecture has two domains: Emotion detector and word ontology:

6.2 Emotion Word Ontology

Ontology is basically defined as the idea of expressing domain in machine understandable form which can also be understood by humans. This is achieved by making mathematical formulas, tools like axioms and set of rules. The main concept of ontology is to communicate between institutions, individuals and application system. Hierarchy of emotion word is transformed into ontology. It was proposed by the researcher W.G.parrot. This can be implemented by the ontology tool which is used to develop emotion ontology. This hierarchy has various classes and subclasses relationship.

6.3 Emotion detector

Emotion identification algorithm is used to recognise emotion of textual data. To solve the problem that was identified in above survey, the following algorithm will provide the solution to it.

7. PROPERTIES OF EMOTION

There are four key properties of emotions:

7.1 Precedent

A precedent is an event that has happened earlier which is responsible for emotion cause. It is the triggering effect to the emotion. For example, I case of happiness, the precedent is say, the birth of one's child.

7.1 Signal

It is the psychological method which is used to express the emotion of the humans. Signal is induced when any individual expresses any specific emotion. For example in case of happiness smile on the face is the signal induced.

7.3 Reaction

After the signal has been made by a person, reaction is the expected response by another individual when the individual understands the signal. For example in case of happiness, if a person sees another happy, one will likely to response such a way that person greets or congratulate the happy person.

7.4 Connection

It shows that a particular emotion has similar precedent, signal, and reaction by different individuals. For example put, identical things which make both human being and a cat happy (precedent), both cat and human being will likely to response in the same way (signal).

8. MISCOMMUNICATION BETWEEN USER VIA MAIL AND TEXTUAL MESSAGE

One cannot detect emotion over text as he is not sure of what and how the person wants to implicate a particular data. This problem was identified by conducting a survey. A sample text was sent to see how different people interpret a single common text in different ways consider a text "did you forget to pick up the grocery?" A survey was taken of sample size 100 recipients. The data set of recipients of the text was as follows:

i) Family members- 23

ii) Friends- 27

iii) Fellow Engineering students- 20

iv) Psychology students and professionals- 8

v) Other-22

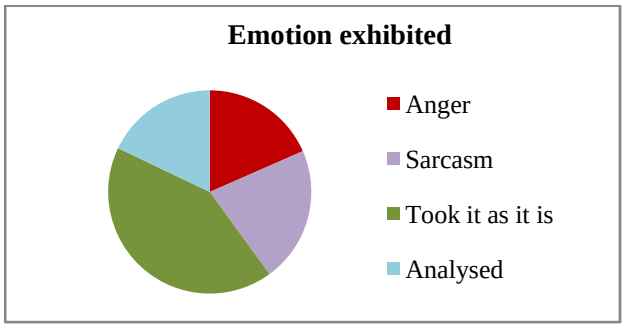


Fig. 1: Emotion perceived by recipients of text

Individual's emotion can't be interpret unless or until he is physical present in front of us. Have you ever hot irritated? Have you ever got annoyed by someone's text? Instead it was a normal text. This happens because of the non-facial interaction between the sender and receiver. Let us take an example how different people interpret a single common text in different ways consider a text "did you forget to pick up the grocery? "People may read this text and start interpreting maybe the sender is saying in this sarcasm or maybe in angry mood or maybe just a simple reminder. This gets more complicated when some kind of relationship is being shared between the respective like son and a mother, husband and a wife, boss and an assistant or many other. We cannot interpret the true emotion accurately but this miscommunication can be partially improved by few methods. Firstly before hitting send button on the hastily written text just count to 10 or even more

and in this gap just ask yourself a question "can I send this text in another way?" Secondly do not interpret the text on initial take try to interpret the text in another possibilities.

9. EMOTION ANALYSIS ALGORITHM

Emotion identification algorithm is used to recognise emotion of textual data. To solve the problem that was identified in above survey, the following algorithm will provide the solution to it.

9.1 Frameworks applied

This algorithm requires calculation of weightage that is designated to various emotion word. This further helps in sorting of those words and emotions. For this an entire series of procedures is followed. Firstly we need to find all the parameters. Jens library helps us to attain this task. Jena API provides functions from its library which enables traversal and parsing of the ontology that we will be framing. Following are the criterion that need to be calculated:

9.1.1 Root and Leaf association

If a particular changes are made in leaf's score then that effect is directly effecting the root, thus root score is also changed. It is so because the data set that is owned by leaf id related to root as well. For modification of score, the traversal of ontology layout or made is done in breadth first manner. This is attained with the help of Jena API. When a particular node is reached all of its leaves are recovered. Then similar procedure is done on every child.

9.1.2 Embeddedness in Ontology

Embeddedness is calculated to get a knowhow of how a particular emotion related term is embedded in the ontology frame. Since its depth is more to according to its embedded degree, higher weightage should be assigned to it. This depth value is procured at the same time the ontology tree is being traversed.

9.1.3 Occurrences in Data set

This parameter signifies that the number of the occurrences of a term will determine its vitality over the text. This is done by parsing the text document and determining the frequency of the terms.

9.2 Algorithm

The following algorithm is given to compute the score for every emotional term by using all the all the parameters of frameworks that we specified above. This will infer that the score is inversely proportional to the how deep the term is embedded in ontology and directly proportional to occurrence of every emotional word. Firstly we will calculate the score of initial stage of emotion class. Lastly, the emotion term class which correspond to maximum score will be signifying the emotion state of the text or the data set adhered to it and that emotion will be labelled on it. The algorithm goes about as:

while a <- 1 to Given Number of node [acquired from ontology]

do root[a] <- root of node a leaf[a] <- leaf of node a

while m<-1 to number of node [acquired from ontology]

do occurrence[b] <- occurrences of b^{th} emotion term

embedded[m] <- embedded b^{th} node present in ontology

Calculation of score:

while b <- 1 to number of root nodes [acquired from ontology]

score(root)= Add(score(root),score(leaf))

return score(root)

while m<-1 to number of root nodes [acquired from ontology]

emotion_term_class <- Max_score[root]

return emotion_term_class

In above algorithm:

1. node[Ontology] gives the otology class,
2. root[a] gives the root class in given ontology framework,
3. leaf[a] gives all the leaf classes in ontology framework,
4. occurrence[b] gives the number of time b^{th} class happens in text or data set
5. Embedded denotes the degree of embeddedness of class
6. score[root] denotes score of root in ontology framework.

The above algorithm will determine the score of algorithm major emotion class. The class having maximum score will determine the main emotion term class of the text or data set.

10. APPLICATIONS

Through this section we will highlight the application of emotion analysis of textual data. Some areas of applications:

1. Emotion analysis and chat applicant

Emotion analysis has been implemented widely in various social media and chat area of application. Chat based user generally have the property of being real-time. Hence being the response time so rapid. In real time communication authors have used expression images (i.e., the image which expresses the emotion) for performing emotion analysis. The labels that author has considered are happy, sad, excited, laugh, disguised, surprise, etc. This is based on input output approach where you input the sentence and output will be the

form of expression image. The output image will depend on the intensity of the text and emotion the of the emotion sentence.

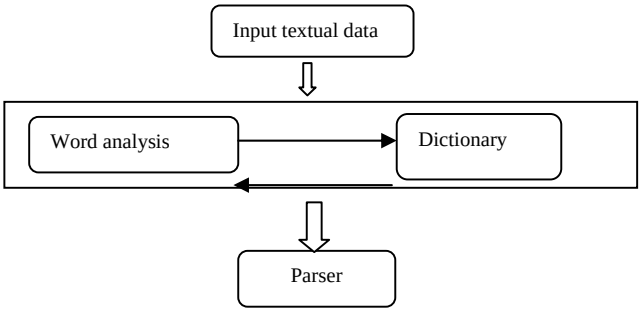
1.1 Approach used in extracting emotion

i) Word is looked up in the dictionary.

ii) When the specific word is not available then it is being stemmed and another possible match will be searched.

iii) If the resultant outcome of above steps is positive then the word will be assigned with appropriate emotion.

iv) Each word symbolizes an emotion therefore a parser is used to demonstrate the syntactic structure of sentence.



2. Emotion analysis and teaching members

For various educational applications different types of automatic tutoring systems have been developed. The teachers is this tutoring system must be possess traditional skills in order to make it effective. One of the kill every tutor should possess is how the students of the class collectively is following you? The goal is to determine the condition of the students as frustrated, bored, happy or in flow.

3. Emotion analysis and mental health monitoring

The biggest risk and danger to the wellness of the human beings is their mental health condition. It is very critical to oversee and monitored well. Today people use the platform of social media to express their emotions and views like Facebook, twitter, etc. it helps the researchers to analysis and understand the people wellbeing very well. Therefore it is very crucial that everyone is affected to mental health risks. A researcher Thomson et al (2004) describes the suicide cases by military officials. He also stated that fear anticipate courage to commit suicide. Since military training leads to nurture them as fearless and gutsiness makes them susceptible to mental risks. The main aim of emotion analysis is to analysis mental health issues.

11. CONCLUSION AND FUTURE WORK

In this paper, we conducted a survey where identification of emotion from textual data was done among people from various professions. There was many interpretation of the same text by different each individual. As we have seen that many research work have been done in identifying emotion from audio and facial information but doing so in the field of textual data is still hot and most noticeable research area. Various techniques of analysis of emotion detection were discussed and a system architecture is discussed which would work efficiently. One of the main concern with the analysis of emotion detection is that humans express their emotion in conceal way independent on the emotive vocab. For example “My delivery job requires me to travel a lot in this non-bearable hot weather on two wheeler” as we can see in this sentence, speaker is conveying extremely negative sentence that he has to travel so hard in his job. Such kind of expression can only be detained with real world knowledge base. Phrases, sarcasm and slangs are form of language which are the main problem in emotion detection. Sarcasm is the form of emotion expression in which surface emotion is exactly opposite of its literal meaning. Therefore a sophisticated method of capturing nuanced forms of sarcasm is needed. In which a system first reproposes the sentence for emotion neutrality and then followed by feeding it to the emotion detector.

REFRENCES

[1] Kato, Y. & Akahori, K., “The accuracy of judgment of emotions experienced by partners during e-mail and face-to-face communication”, ICCE2004 (pp. 1559–1570), 2004.

[2]Kato, Y., & Akahori, K. “Effects of emotional transmissions between senders and receivers on emotions experienced in e-mail communications”, World Conference on Educational Multimedia, Hypermedia and Telecommunications, 723-730, 2005.

[3] Saima Aman and Stan Szpakowicz, “Identifying Expressions of Emotion in Text”, pages 196–205. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[4] Kato, Y., Kato, S., & Akahori, K, “Effects of emotional cues transmitted in e-mail communication on the emotions experienced by senders and receivers”, Computers in Human Behavior, 23, 1894-1905, 2007.

[5] Pierce, T., “Social anxiety and technology: Face-to-face communication versus technological communication among teens”, Computers in Human Behaviour, 25, 1367-1372, 2009.

[6] Riordan, M. A., & Kreuz, R. J., “Emotion encoding and interpretation in computer mediated communication: Reasons for use”, Computers in Human Behaviour, 1667-1673, 2010.

[7] Kato, Y., Kato, S., Scott, D. J., & Sato K, “Patterns of emotional transmission in Japanese young people’s text-based communication in four basic emotional state”, International Journal on E-Learning, 9(2), 203-227, 2010.

[8]Holtgraves T.,”Text messaging, personality, and the social context”, Journal of Research in Personality, 45, 92-99, 2011.

[9]Bart Desmet &V´eronique Hoste , “Emotion detection in suicide notes”, Expert Systems with Applications, 40(16):6351–6358, Nov 2013.

[10] Tanya Bhattacharya, Arunima Jaiswal, Vaibhav Nagpal , “Web usage mining and text mining in the environment of web personalization for ontology development of recommender systems”, Pages: 78 - 85, DOI: 10.1109/ICRITO.2016.7784930, 2016.