Volume 6, Issue 2, February 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Intelligent Text Processing for Emotion Detection in Text

Shyamol Banerjee*, Prof. Unmukh Dutta
Department of Computer Science and Engineering
M.P.C.T., Gwalior, India

***Abstract*— This paper present an intelligent text processing technique for identifying the emotion contains in the text data. Users share their opinions and sentiment in the form of comments in different online marketing websites for different products. Emotion detection in these customer reviews for different products in one of the challenging task. In this paper, we proposed efficient technique for detection of emotion in customer reviews for different products.**

***Keywords*— Text Mining, Sentiment analysis, Natural Language Processing, Emotion Recognition, Text Processing, Affective Interface.**

I. INTRODUCTION

Emotion is a particular type of feeling that represents various states of mind such as joy, fear, anger, love and so on. Emotion Detection from text is an important component of Artificial Intelligence; it plays a key role in human-computer interaction [1]. A person can express emotions by speech, facial expressions, hand gestures and written text [2]. Sufficient amount of work has been done in speech and facial emotion recognition but recognition of emotions through text still needs attraction of researchers [3]. From applicative point of view, detecting the human emotions in text is becoming increasingly important in computational linguistics [4].

On the web there is huge amount of textual information which is quite interesting to extract emotions from these textual information for special purpose such as business [5]. For example, in luxurious products available online, the emotional aspects like brand , uniqueness and prestige for purchasing decisions are weighted more by customer as compared to rational aspects such as technical, functional and cost[6]. If a customer is emotionally satisfied he can purchase a product at high price. Emotional Marketing targets the customers emotions to encourage him to opt for a particular brand and so results in increase of product/service sales[7]. Nowadays there is a wide range of products available, but the main target is to create confidence in customer about a product/service he communicates [8].

There are various emotional models used to build emotion recognition system[1,2]. One of the latest model suggested is the hourglass model[1] which has been inspired biologically and psychologically-motivated, and is based on the idea that the emotional states results from the selective activation or deactivation of different resources in brain. Another very common model used in Ekman's model [2] which divides emotions into six universal categories.

Several current approaches associated with emotion detection are based on methods of supervised learning, where a huge set of annotated data(the emotions are labeled as text) is needed to train the model[9]. Since in supervised learning we can obtain good results , but availability of the large data sets with annotations is very low, and when a model gets trained, it is not well translated to another[10].

There are few methods where supervised learning is not used. But most of these methods uses manually designed dictionaries of emotion keywords [11]. There are few methods that do not make use of supervised learning approach. Though, most of these methods use emotion keywords that are taken from manually designed dictionaries. The problem related with such an affect lexicon-based method is that they contain a limited and fixed number of emotion categories in the dictionary.

The other problem is that when a word that expresses emotion is used in a sentence that does not exist in the dictionary, and then it can be considered as unemotional.

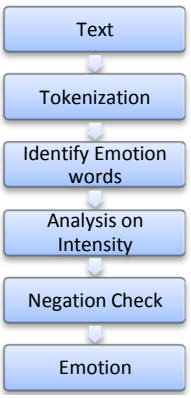

Fig 1: Steps for detection of emotion in text

There are also methods which normally depend on linguistic rules, but designing such rules is not a minor task. Most of these rules are not publicly available. In addition, most current emotion detection methods look at words individually without taking in consideration the context of word.A word may reflect different emotion in different context.

II. RELATED WORK

There have been several work done in emotion detection. Previously Osgood et al.[3,4] has done work to understand expression of emotion in text. Multidimensional has been used to visualize the affective words so that similarity rating between them can be calculated. Osgood used three dimensions, evaluation, potency and activity, where evaluation measures how much a word can refer to a pleasant or an unpleasant event, whether emotional intensity of a word is strong or weak , is computed by potency, and activity refers to the active or passive nature of a word.

Strapparava et al.[5] developed a linguistic resource to represent affective knowledge lexically which has been named as WordNet – Affect [5]. The WordNet-Affect constitutes of a subset of synsets which represent affective concepts matching to affective words. The classification of emotions is then done by mapping emotional keywords that are present in the input sentence with their corresponding in WordNet-Affect.

Ghazi et al. [8] followed the hierarchical classification to classify the six Ekman's emotions. Multiple levels of hierarchy was used while classifying emotions by first classifying whether a sentence contains an emotion or not, then classifying whether the emotion is either positive or negative and finally classifying the emotions on a fine-grained level. They used different features for the classifier for each stage of classification, and a better accuracy of (+7%) was achieved over the flat classification where flat classification classifies the emotions on a fine-grained level directly . The main disadvantage of this approach is that it is not context sensitive.

F. Chaumartin [9] developed a linguistic rule-based system UPAR7, using WordNet [10], SentiWordNet [11] and WordNet-Affect [12] lexical resources. The system uses the dependency graph obtained from the Stanford POS tagger [13], where the graph's root is considered the main subject. For each emotion each word is represented individually. Then the rating of the main subject (main word) is enhanced, as it is more important than the remaining words in the sentence. For Ekmans model of six emotions the best-achieved accuracy of this method was 30% In addition to the low accuracy of this method, it lacks the global understanding of the sentence and is not context sensitive.

Shadi Shaheen et al[16] had proposed a framework that classifies emotions in English sentences where the emotions are considered as generalized concepts that have been extracted from the sentences. They generated an intermediate emotional data representation of a given input sentence which is based on syntactic and semantic structure of the sentence. This representation has been generalized by using various ontologies like ConceptNet and WordNet, which results in an emotion seed that they called as emotion recognition rule (ERR). Finally, a group of classifiers were used to compare the generated ERR with a set of reference. The most famous emotion categories are given by various authors is given in table 1.

Table1. Basic emotion categories identified by various researchers

Tomkins (1962)	**Izard (1977)**	**Plutchik (1980)**	**Ortony (1988)**	**Ekman (1992)**
Joy	Enjoyment	Joy	Joy	Happiness
Anguish	Sadness	Sorrow	Sadness	sadness
Surprise	Surprise	Fear	Fear	Fear
interest	Interest	Surprise	Anger	Anger
shame	Shame	Acceptance	Disgust	disgust
Fear	Fear	anticipation	surprise	surprise
Anger	Anger	Anger		
Disgust	Disgust	Disgust		
	Shyness			
	guilt			

A. Lexicon-based approaches

In this approach only lexical resources are used to detect emotions. Keyword based approaches that are based on predetermining of a set of terms in order to classify the text into emotion categories are also found among these approaches. Strapparava [23], as a baseline, implemented a simple algorithm that can check the presence of affective words in the headlines, and figured out a score that reflects the frequency of the words in this affective lexicon in the text. They had used WordNet-Affect [5].Amongst Lexicon based approaches; we can find the ones that are based on Ontology. (Balahur et al., 2011) used EmotiNet - a resource to detect emotion from text based on common sense knowledge based on concepts, their interaction and their affective consequence to detect emotion.

B. Machine learning based approaches

To prevail over the limitations faced by rule-based methods, researchers created some statistical machine learning techniques which can be divided into supervised and unsupervised techniques. The Supervised machine learning with information from different sciences, like human psychology are also used for identification of emotions[7]. Two machine learning approaches are:

a) SVM classifier for classification of opinion

SVM is a machine learning classifier used for text categorization. The review text to be classified is changed into word vectors. SVM constructs a hyper-plane by use of these vectors which separates instances of data of one class from another. SVM get this hyper-plane using training instances also called support vectors. In the binary categorization of text the hyper-plane which classifies the document d_j as $c_j \epsilon$ (1,-1) can be represented by weight vector $\vec{w}$[29].

$$\vec{w} = \sum a_j c_j \vec{d}j \, , \, \alpha_j \geq \qquad (1)$$

Where α_j is a multiplier and for $\vec{d}j$ that are greater than zero are support vector[29] . Test instance is classified by determining which side of $\vec{w}$'s hyper-plane they fall on.

b) Naïve Bayes classifier

It is based on Bayes theorem of posterior probability [29]. If D is dataset of training data instances X= (X_1, X_2……. X_n) having n attributes and m classes C_1,C_2……..C_n. The classifier predicts text X belongs to class having higher probability values for given conditions. This is shown in equation (2)

$$P(C_i/X) > P(C_j/X) \text{ for } 1 \leq j \leq m, j \neq i \qquad (2)$$

Where $P(C_i/X)$ is calculated using Bayes theorem.

$$P(C_i/X) = \frac{P(X/Ci)P(Ci)}{P(X)} \qquad (3)$$

III. DATA COLLECTION

Data has been collected from top online marketing websites such as Amazon, Snapdeal and flipkart for different Products given in Table 2.

Table 2. Total reviews collected from various sites

Category	No. of Reviews		
	Amazon	FlipKart	Snapdeal
Smartphone	450	105	848
LED	146	168	198
Camera	117	53	221
Laptops	34	84	168
Tablets	73	382	134
Total	820	792	1569

IV. PREPROCESSING

In data pre-processing approach the input text is analysed; at first the sentence segmentation is done followed by tokenization. After tokenization the Stop Words are removed and finally Stemming is done on tokenized input. Sentence Segmentation: Detection of sentence boundary takes place where larger processing units consisting of more than one word are extracted. Figure 2 describe the steps used for pre-processing the data set to remove noise present in the data set.

A. Tokenization

It is the process of breaking a stream of text up into phrases, symbols, words, or other meaningful elements called tokens [7]. The objective of the tokenization is the investigation of the stop words. Textual data is considered as only a textual interpretation or block of characters at the beginning. To retrieve information the words of the data set are required. So a parser is needed which processes the tokenization of the documents. This might be trivial as the text is already stored in machine-readable formats. But there are still some problems that have been left, for e.g., the removal of punctuation marks and other characters like hyphens, brackets etc [8]. The main use of tokenization is to identify meaningful keywords. Another problem is about abbreviations and acronyms which need to be transformed into a standard form.

B. Stop Word Removal

In English language some words are most frequently used which are considered worthless are called stop words [11]. These words are removed in this process e.g. to, the, of, etc.

C. Stemming

It is the process of obtaining root or stem word of each word that emphasizes its semantics. This method removes various suffixes, so that number of words can be reduce, to have exactly matching stems, and to save memory space and time.

In computational linguistics, a stem is the part of the word that will never change even when inflected morphologically. Stemmers are normally easier to implement and run faster, and the reduction in accuracy may not matter for some applications. Stemming refers to a heuristic process that chops off the ends of words in the expectation of achieving this objective correctly most of the time, and often includes the removal of derivational affixes[12]. There is couple of points to be taken into consideration while using a stemmer:

a) The Morphological forms of a word are believed to have the identical base meaning and hence should be mapped to the same stem.

b) Words that do not have the same meaning should be kept apart.

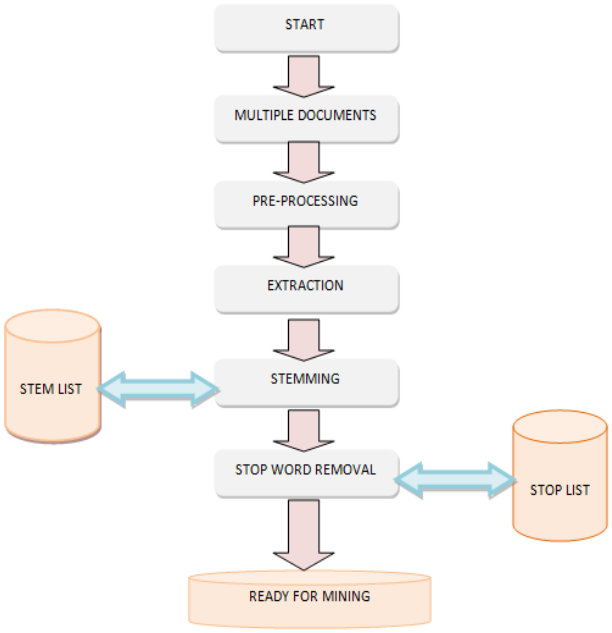

Fig 2: Steps for pre-processing for data

V. EXPERIMENTAL ANALYSIS

In this section we applied classification techniques to divide sentence into two parts: Emotion sentences(E) and neutral sentences(N).Two most popular algorithms[28], Naive bayes(NB) and Support vector machine(SVM) algorithm are used to classifying the data set. We use following words given in WordNet-Affect dictionary [5].

Table 3 Class distribution in the dataset used in emotion/non-emotion classification

Class	Number of Sentences	Percentage
EM	2618	78.7%
NE	708	21.3%
Total	3326	100%

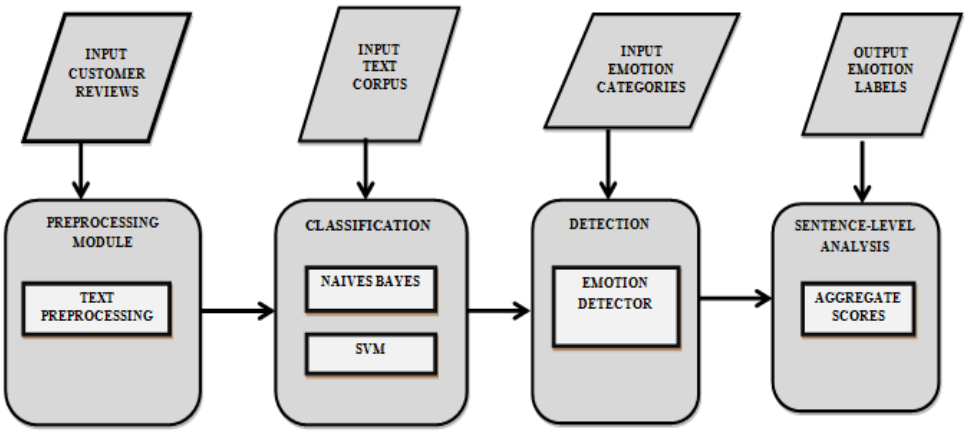

Fig 3: Overview of the emotion detection framework

Data set is divided into emotion category sentences (labeled H, Sd, A, D, Sp, F) to the class emotion class. Following words are collected from the dataset for particular category and matched with the WordNetAffect[5] categories and presented in Table 4.

Table 4 Emotions words corresponding to different categories

Most frequent emotion indicator in data.					
Happiness (H)	Sadness (Sd)	Anger (A)	Disgust (D)	Surprise (Sp)	Fear (F)
Good	Miss	Yelling	Hate	Amazing	Afraid
Lol	Hurt	Angry	dislike	Surprise	Scared

Fun	Sorry	Bitch	shit	wonder	Nervous
Love	Lost	Furious	Stupid	Unexpected	Worry
Happy	Bad	Annoyed	Fucking	Can't believe	Security
Nice	Sad	possed	Disgusting	Weird	Fear
Awesome	Cry	Fucking	Crap	Suddenly	What if
Funny	Stress	Upset	Bitch	Odd	Threat
Great	Wept	Mad	sick	strange	Freak
excited	Longing	Shut up			dangerous

After the fine-grained emotion classification experiments only those sentences from the corpus has been selected for training and evaluation of classifiers which does not contains mixed emotions. The resulting dataset is rich in all types of emotion and their distribution is presented in Table 5.

Table 5 Emotion types and their distribution

Emotion Class	**Number of instances**
Happiness	2184
Sadness	215
Surprise	106
Anger	90
Fear	42
Disgust	6
No-emotion	708

For evaluation purposes we develop a baseline system which a based on word counts. The number of emotion words present in a sentence of each category is being assigned to that category with the largest number of words to the sentence. Again the word lists are extracted from WordNet-Affect[5] for six basic emotion categories. Table 6 shows the precision, recall, and F-Measure values for the proposed baseline system.

Table 6 Performance metrics of the baseline system

Class	**Precision**	**Recall**	**F-Measure**
Happiness	0.793	0.684	0.734
Sadness	0.570	0.462	0.510
Surprise	0.569	0.430	0.489
Anger	0.661	0.439	0.527
Fear	0.591	0.170	0.264
Disgust	0.666	0.176	0.278
No-emotion	0.414	0.593	0..487

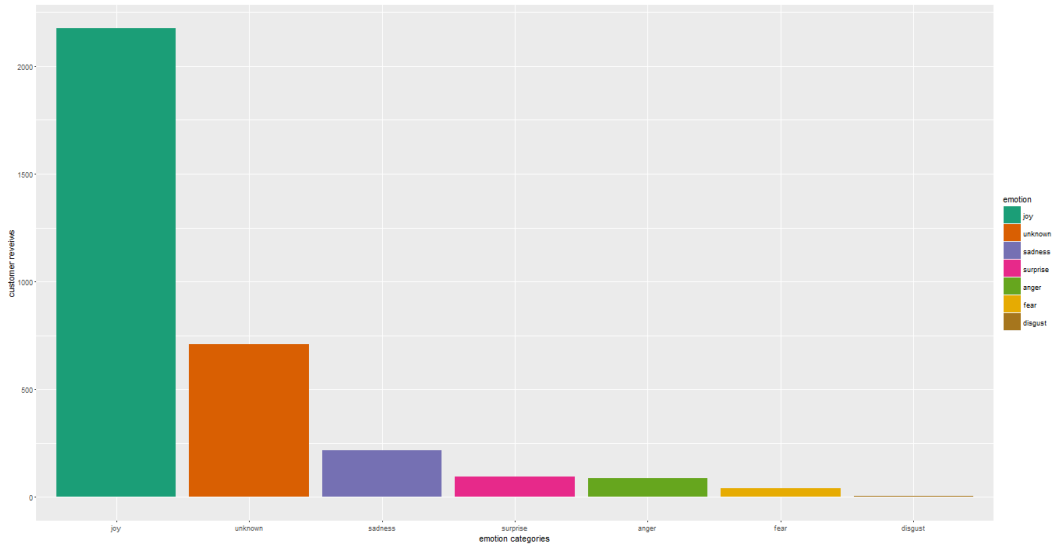

Fig 4 Graphical representation of emotion classes in overall reviews

Fig 6 Resultant Word Cloud

VI. CONCLUSION

The objective of this work is to develop a methodology for automatic detection of emotions from text. Data has been collected from top online marketing websites in the form of customer reviews for different product categories. In this work various data pre-processing step has been applied to remove noise in the data. Classifications of different emotion words present in the sentences are used to find the category of emotion classes. Baseline line system has been developed for classification of multiple emotional words present in the sentences. This system detects 78% emotion sentences from the corpus.

REFERENCES

[1] Erik Cambria, Andrew Livingstone, and Amir Hussain. "The hourglass of emotions." *Cognitive behavioural systems*. Springer Berlin Heidelberg, 2012. 144-157.

[2] P. Ekman, (1999) Basic emotions. In T. Dalgleish and T. Power(Eds.) *The handbook of cognition and emotion*. Pp. 45-60. NewYork.: John Wiley & Sons.

[3] Charles E. Osgood. Cross-cultural universals of affective meaning.University of Illinois Press, 1975.

[4] Charles E. Osgood, and Oliver Tzeng. *Language, meaning, and culture: The selected papers of CE Osgood*. Praeger Publishers, 1990.

[5] Carlo Strapparava, and Alessandro Valitutti. "WordNet Affect: an Affective Extension of WordNet." *LREC*. Vol. 4. 2004.

[6] Cheng-Yu Lu, et al. "Automatic event-level textual emotion sensing using mutual action histogram between entities." *Expert systems with applications* 37.2 (2010): 1643-1653.

[7] Chung-Hsien Wu, Zes-Jing Chuang, and Yu-Chung Lin. "Emotion recognition from text using semantic labels and separable mixture models." ACM transactions on Asian language Information *processing (TALIP)* 5.2 (2006): 165 183.

[8] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. "Hierarchical versus flat classification of emotions in text." *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 2010.

[9] François-Régis Chaumartin. "UPAR7: A knowledge-based system for headline sentiment tagging." *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 2007.

[10] George A. Miller, "WordNet: a lexical database for English."*Communications of the ACM* 38.11 (1995): 39-41.

[11] Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." *Proceedings of LREC*. Vol. 6. 2006.

[12] Carlo Strapparava, and Alessandro Valitutti. "WordNet Affect: an Affective Extension of WordNet." *LREC*. Vol. 4. 2004.

[13] K.Toutanova, KleinD., ManningC., SingerY.,StanfordPOStagger,[Online].Available: http://nlp.stanford.edu /software/tagger.shtml, Stanford, 2003

[14] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. "EmoHeart: conveying emotions in second life based on affect sensing from text."

[15] Second Life Game, http://secondlife.com/

[16] Shadi Shaheen,Wassim El-Hajj, Hazem Hajj, Shady Elbassuoni,"Emotion Recognition from Text Based on Automatically Generated Rules." 2014 IEEE International Conference on Data Mining Workshop. *Advances in Human-Computer Interaction* 2010 (2010)

[17] Lily Dey,Nadia Afroz and Rudra Pratap Deb Nath*,"Emotion extraction from real time chat messenger*",in proceedings of 3rd International Conference On Informatics,Electronics & Vision 2014.

[18] C. Alm, "Affect in text and speech," 2008. [Online]. Available:http://cogcomp.cs.illinois.edu/papers/Alm thesis(1).pdf

[19] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in Proceedings of the Third International Joint Conference on Natural Language Processing, 2008, pp. 296–302.

[20] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical methods in natural language processing, 2002.

[21] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in Proceedings of the AAAI Internatonal Conference on Weblogs and Social Media, 2009.

[22] S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," in Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text,2010, pp. 62–70.

[23] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in Proceedings of the ACM symposium on Applied computing, 2008, pp. 1556–1560.

[24] Alexandra Balahur, Jes´us M. Hermida, and Andr´esMontoyo. 2011. Detecting Implicit Expressions ofSentiment in Text Based on Commonsense Knowledge.In 2nd Workshop on Computational Approachesto Subjectivity and Sentiment Analysis,ACL-HLT 2011, pages 53–60.

[25] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1999. Indexing by latent semantic analysis. Journalof the American Society for Information Science, 41(6):391–407, September.

[26] Alastair J. Gill, Robert M. French, Darren Gergle, and Jon Oberlander. 2008. Identifying Emotional Characteristics from Short Blog Texts. In 30th Annual Meeting of the Cognitive Science Society, pages 2237–2242.

[27] Xuren Wang and Qiuhui Zheng. 2013. Text Emotion Classification Research Based on Improved Latent Semantic Analysis Algorithm. In Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), number Iccsee, pages 210–213, Paris, France. Atlantis Press.

[28] Jain,V.K., Kumar,S.,(2015) An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter, Procedia Computer Science, Volume 70, 2015, Pages 801–807,2015.

