Use of Porter Stemming Algorithm and SVM for Emotion Extraction from News Headlines

Chaitali G. Patil   Sandip S. Patil

***Abstract* — Abstract - Emotions play an essential role in social interactions, performs important regulatory and utilitarian functions within human body, brain, facilitate rational decision making and perception. Emotion indicating, if the feeling is positive or negative. Emotions may be expressed by a person's speech, face expression and written text known as speech, facial and text based emotion respectively.**

There are six basic emotion classifications:anger, disgust, fear, happiness,sadnss and surpri se. In this paper we are proposing a novel method for extracting the emotions from news headlines. In this approach we are using porter stemming algorithm for preprocessing and SVM classifier for classification.

***Keywords* – Emotion classification, SVM, emotion detection, text categorization, preprocessing.**

I. Introduction

Emotion is one type of affect, other types being mood, temperament, and sensation (e.g. pain).Emotions can be understood as either states or as processes. When understood as a state, for example, being angry or being afraid, an emotion is one type of mental state. As such, the emotion interacts with other mental states and guides behavior.

Emotions may be expressed by a single word or a group of words. Sentence level emotion identification process plays an important role to track emotions or to find out the cues for generating such emotions or to properly identify it. Sentences are the basic information units of any document. For that reason, the overall document level emotion identification process depends on the emotion expressed by the individual sentences of that document which in turn is based on the emotions expressed by the individual words. Emotions may be expressed by a person's speech, face expression and written text known as speech, facial and text based emotion respectively[9]. Sufficient amount of work has been done regarding to speech and facial emotion recognition but text based emotion recognition system still needs attraction of researchers.

In computational linguistics, the detection of human emotions in text is becoming increasingly important from an applicative point of view. Emotion is expressed as joy, sadness, anger, surprise, hate, fear and so on. Since there is not any standard emotion word hierarchy, focus is on the related research about emotion in cognitive psychology domain.

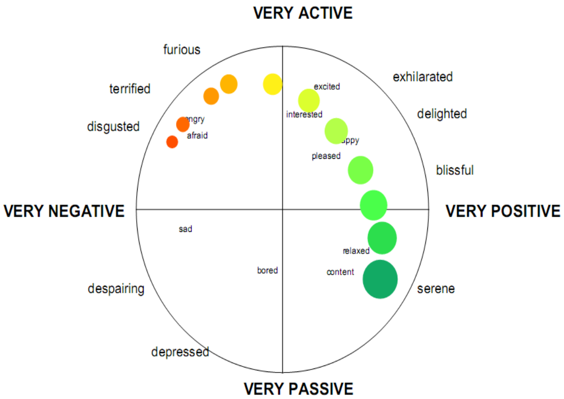

Fig 1: A two-dimensional representation of emotion, derived from [2].

II. Related Work

Research on the sentiment classification is challenging, and more research have been done on this topic. Plaban*et al* [1] introduce a multi-label emotion classification model based on ADT boost MH(a boosting algorithm called ADT boost MH). Word present in the sentences and the polarity of the subject, object and verb are used as features. The classifier performs better than that on singular words only. The average precision was 79.5%.Hu *et al*[2], implemented a Bayes text classification. Their results show that the Naive Bayes classification method achieves a high performance on text classification. However, this method just categorizes texts into only two classes: positive and negative, which excludes reader's emotions such as angry, happy etc .Zornitsa*et al* presented an approach of headline emotion classification based on frequency and co-occurrence information collected from the World Wide Web [3].

Kevin Hsin-Yih Lin *et al* classified news into emotions using various combinations of feature sets and identifying the emotional influences of news articles on readers [4].XU Lin hong*et al* computed semantic similarity of the vocabulary and tagged vocabulary in How Net(http://www.keenage.com/), adopted the derogatory or commendatory terms as features of classification, utilized Support Vector Machine classifier to identify the text orientation, and dealt with the negative sentence via matching negative rules [5]. Chinese character bi-gram, words, metadata, affix similarity, word emotion and emotion categories are used as features. They got the highest prediction accuracy 76.88% for *bored* and 89.66% for useful. Class useful is some kind of vague on emotion expression [6].Yu Zhang et al explored how to incorporate emotional aspects of dialog into existing dialog processing techniques and worked on making a Chinese emotion classification model which is used to recognize the main affective attribute from a sentence or a text [7]. Prem Melville *et al* developed a unified framework in which one can use background lexical information in terms of word-class associations, and refine this information for specific domains using any available training examples [8].

Before going for emotion classification, the first question is "Which emotions should be addressed?" There are many different emotion sets exits including "happy", "sad", "surprise", "fear" and so on. These categories of emotions helps the conversational agent like chatbot or intelligent robot to give more human like responses based on the emotional state of user. Table 1 shows a list of primary emotion.

Table 1 : List of primary emotion

Anger	Disgust	Fear	Joy	Sad	Surprise

III. IMPLEMENTATION

This implementation part explains the system overview of emotion extraction from news headlines using SVM. In this, there are two types of phases i.e. conceptual model for training phase and conceptual model for testing phase. They are as follows:

1. Conceptual model for training phase :

This diagram contains six blocks. From that one block representing the dataset (ISEAR) and another block contain the dictionary (WordNet Affect).

Dataset (ISEAR) :

The ISEAR dataset consists of 7,666 sentences (Scherer and Wallbott, 1994), annotated by 1,096 participants with different cultural background who completed questionnaires about experiences and reactions for seven emotions: anger, disgust, fear, sadness, shame and guilt. The ISEAR dataset contains the emotional statements that in turn contain the emotional sentences. ISEAR contain news sentences. In that 250 is tested data and remaining is for trial.

Extract News :

In this block we extract the news form ISEAR dataset.

Preprocessing :

This preprocessing block performs various operations like filtering, tokenization, stemming and pruning etc.

In the filtering, input data is then filtered i.e. the special characters and special symbols such as @, <, >, $, ^, etc. are removed.

A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. The process of breaking a text up into its constituent tokens is known as tokenization. Tokenization means removing the stop words like a, and, the etc.

In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. For this reason, a number of so-called *stemming Algorithms*, or *stemmers*, have been developed, which attempt to reduce a word to its *stem* or root form.

Pruning also counts the number of times a particular term is occurring in the document which is also called as term frequency.

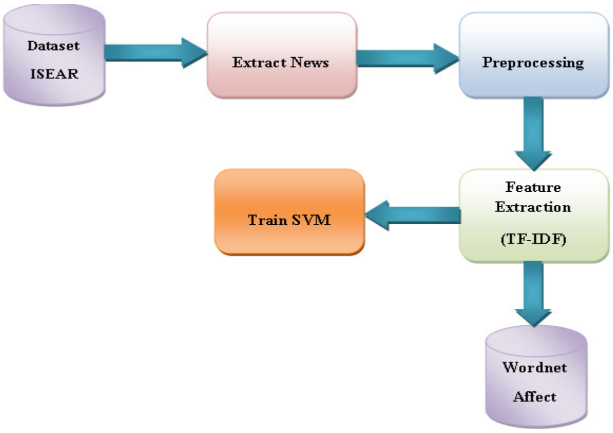

Figure 3 : Conceptual model for training phase

Feature Extraction (TF-IDF) :

This block only contains the emotion's words. And with the help of TF-IDF we can calculate the, how much amount of emotional word it will contain and from which category it will belong.

WordNetAffect :

It contain emotion dictionary. This emotional dictionary dictionary is nothing but it contain all the emotional words in it. For this it will contain the six basic primary emotions i.e anger, disgust, fear, joy, sad surprise[10].

Train SVM :

SVM is the supervised learning method so we have to train our SVM first. So with the help of this emotional dictionary i.e WordNet Affect and feature extraction method we will train our SVM classifier.

2 Conceptual model for testing phase :

This diagram contains eleven blocks. These are as follows. Some blocks are previously discussed in figure 3.

Extract News :

In this block we extract the news form ISEAR dataset.

Classifying using SVM :

With the help of this block we can easily find out the emotion from the news head line.

Preprocessing :

This preprocessing block performs various operations like filtering, tokenization, streaming pruning etc.

In streaming, we stream the words using porter algorithm. The stemmer is divided into a number of linear steps, five or six depending upon the definition of a step, that are used to produce the final stem. The porter algorithm describe as follows :

Porter Stemming Algorithm for Preprocessing

Step 1 : The algorithm is designed to deal with past participles and plurals. The subsequent steps are much more straightforward.
Ex. Plastered → Plaster
Cats → Cat

Step 2 : Deals with pattern matching on same common suffixes.
Ex. Happy → Happi
Relational → Relate

Step 3 : Deals with special word endings.
Ex. Triplicate→ Triplic
Hopeful → Hope

Step 4: Check the stripped word against more suffixes in case the word is compounded.
Ex. Revival → Reviv
Allowance → Allow

Step 5 : Check if the stripped word ends in a vowel and fixes it appropriately.
Ex. Controll → Control
Probate → Probat

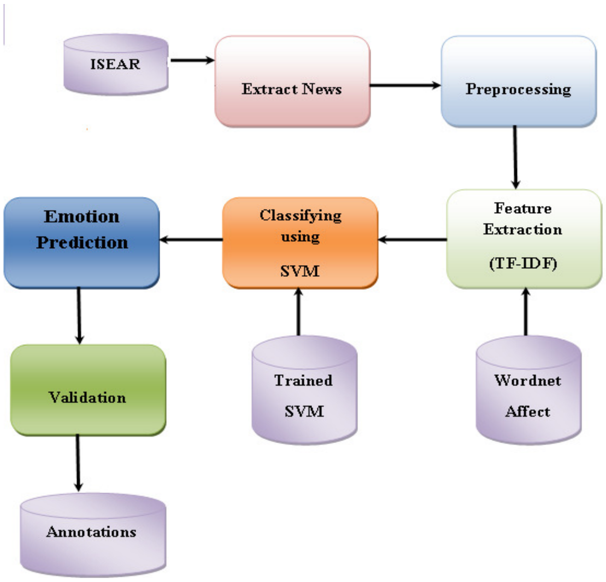

Figure 4 : Conceptual model for testing phase

Emotion prediction :

This block generates the final output of emotion.

Validation :

Whenever our output will get generate that output will get compare with the standard emotion with their valence.

Annotations :

The test data set was independently labeled by six annotators. The annotators were instructed to select the appropriate emotions for each text based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline. Annotation examples were also provided, including examples of headlines bearing two or more emotions to illustrate the case where several emotions were jointly applicable. Finally, the annotators were encouraged to follow their .first intuition, and to use the full-range of the annotation scale bars.

Proposed algorithm for Emotion Extraction from News headlines:

INPUT

Set of documents D = $\{D_1, D_2, . . . ,Dm\}$
Fixed set of categories C = $\{C_1, C_2, . . . ,Cn\}$

STEPS

Step_1 : Wordlist {m, 1}← each word in the document

Step_2 : Wordlist {m, 2} ← Filtering (Wordlist)
Step_3: Wordlist {m, 3} ← Tokenization (Wordlist)

```
Step_4 : Wordlist {m, 4} ← Stemming (Wordlist)
Step _5 : Wordlist ← Pruning (Wordlist)
Step_6 :$tf(t,d) \leftarrow \sum_{x \in d} fr(x,t)$
Step_7 :idf← log(n/N)
Step_8 :$TF - IDF(t,f) = -\log \frac{dt(f)}{N} * \frac{tf(t,d)}{|d|}$
Step_9 : Crate Classifier
Step_10 : Use Classifier C (Y/X)
Step_11 : Emotion Extracted from News headlines.
```

IV. Results And Discussion

We have implemented the system for emotion classification of news headlines. The system annotates the presence of emotions in the text simply based on the presence of word in the WordNet Affect dictionary.

F-measure :

The F-measure is defined as a harmonic mean of precision (P) and recall (R).This is also known as the F_1 measure, because recall and precision are evenly weighted.

$$F = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{(\mathrm{precision} + \mathrm{recall})}.$$

Precision :

Precision is the fraction of the documents retrieved that are relevant to the user's information need. Precision refers to the closeness of two or more measurements to each other.

$$\mathrm{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall :

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\mathrm{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Accuracy :

Accuracy describes the nearness of a measurement to the standard or true value, i.e., a highly accurate measuring device will provide measurements very close to the standard, true or known values.

Table 1: Results for WordNet Affect Testing Dataset

Emotions (ISEAR Testing Dataset)	F1 score	Precision	Recall	Accuracy
Anger	64.07	55.12	72.33	75.8
Disgust	53.35	46.60	62.40	70.83
Fear	72.47	62.90	85.75	68.9
Joy	70.62	64.14	78.55	72.83
Sad	74.51	68.14	82.22	69.86

The table 1 consist of F1 score, Precision, Recall, accuracy result's values for Anger, Disgust, Fear, Joy and sad respectively.

Table 2: Comparative Results for different classifiers

Classifiers	Precision	Recall	F1 Score	Accuracy
Naïve Bayes Classifier	55.3	60.6	61.2	60.8
Vector Space Model	59.1	59.4	28.5	34.8
Support Vector Machine with Porter algorithm (Proposed Method)	**59.38**	**76.25**	**67.004**	**71.64**

This table 2 consist of comparative results for different classifiers. And with the help of this comparative results we can conclude that SVM is the best classifier among them.

Conclusion

. In this paper we have developed the emotion extraction methodology from news headline using support vector machines. In that first of all we extract news from ISEAR dataset. On this obtain news preprocessing operation is perform. This preprocessing contains filtering, tokenization, stemming and pruning. For the stemming we implemented the Porter Stemming Algorithm. After that with the help of WordNet Affect dictionary, TF-IDF is calculated. This gives Term frequency and inverse term frequency. And in the last step the train SVM classifies the emotions from the news headlines which are again validated with annotations. and examine classification performance under different feature settings. Experiments show that certain feature combinations achieve good results. And also from literature survey we concluded that the support vector machine classifier perform well among the all.

References

[1] A. B. Plaban Kumar Bhowmick and P. Mitra, "Classifying emotion in news sentences: When machine classification meets human

classification”. *International Journal on Computer Science and Engineering*, 2(1):98–108, 2010.

[2] AlenaNeviarouskaya, Helmut Prendinger and Mitsuru Ishizuka, “Research ArticleEmoHeart: Conveying Emotions in Second Life Based on Affect Sensing fromText” *Hindawi Publishing Corporation Advances in Human-Computer Interaction Volume 2010*, Article ID 209801, 13 pages doi:10.1155/2010/209801.

[3] AlenaNeviarouskaya, Helmut Prendinger and Mitsuru Ishizuka, “Compositionality Principle in Recognition of Fine-Grained Emotions from Text” *Association for the Advancement of Artificial Intelligence,* 2009

[4] Alexandra Balahur, Jesus M. Hermida and Andres Montoyo, “Detecting Implicit Expressions of Sentiment in Text Based on Commonsense Knowledge” *ACL-HLT 2011*, pages 53–60,24 June, 2011.

[5] Carlo Strapparava and RadaMihalcea, “SemEval-2007 Task 14: Affective Text” *(SemEval-2007),* pages 70–74, June 2007.

[6] Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, Von-WunSoo, ”Towards Text-based Emotion Detection: A Survey and Possible Improvements ”,in *International Conference on Information Management and Engineering*,2009.

[7] C.-H. Wu, Z.-J.Chuang and Y.-C. Lin, “Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models,” *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 5, issue 2, Jun. 2006, pp. 165-183, doi:10.1145/1165255.1165259.

[8] C.Maaoui, A. Pruski, and F. Abdat, “Emotion recognition for human machine communication”, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 08)*, IEEE Computer Society, Sep. 2008, pp. 1210-1215, doi: 10.1109/IROS.2008.4650870.

[9] Radim BURGET and Jan KARÁSEK, ZdeněkSMÉKAL,“Recognition of Emotions in Czech NewspaperHeadlines”*RADIOENGINEERING, VOL. 20, NO. 1, APRIL 201.*

[10] Hugo Liu , Henry Lieberman and Ted Selker “A Model of Textual Affect Sensing using Real- World Knowledge” *IUI’03*, January 12-15, 2003, Miami, Florida, USA. Copyright 2003 ACM 1-58113-586-6/03/0001.

AUTHOR’S PROFILE

Chaitali G. Patil received the B.E. in Information Technology, in 2009 from MIT COE, Pune affiliated to Pune University. Perusing M.E. degree in Computer Science and Engineering from SSBT’s COET, Bambhori, Jalgaon.

Sandip S. Patil received the B.E. degree in Computer Engineering, in 2001 from SSBT’s College of Engineering and Technology, Bambhori, Jalgaon affiliated to North Maharastra University Jalgaon(M.S.), M.Tech. degree in Computer Science and Engineering from Samrat Ashok Technological Institute Vidisha in 2009, Presently working as Associate Professor in department of Computer Engineering at S.S.B.T. College of Engineering and Technology, Babhori Jalgaon. (M.S.) having 12 years of research experience. His area of interests are Soft Computing and machine learning.