Journal of Engineering Sciences
Vol 11, Issue 6,June/2020
ISSN NO:0377-9254

SENTIMENTAL ANALYSIS ON SELECTED DATA SETS

P. Anusha, M.Tech #1, T. Srivalli #2, G. Maneesha #3,
O. Padma #4, K.S.S. Pavan Kalyan Varma #5

#1 Assistant professor, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#2 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#3 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#4 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)
#5 Student, Dept Of CSE, Qis College of Engineering and Technology, Ongole, Prakasam (Dt)

Abstract: Over the past two decades, automatic facial emotion recognition has received enormous attention. This is due to the increase in the need for behavioral biometric systems and human–machine interaction where the facial emotion recognition and the intensity of emotion play vital roles. The existing works usually do not encode the intensity of the observed facial emotion and even less involves modeling the multi-class facial behavior data jointly. Our work involves recognizing the emotion along with the respective intensities of those emotions. Emotion can be expressed in many ways that can be seen such as facial expression and gestures, speech and by written text. Emotion Detection in text documents is essentially a content – based classification problem involving concepts from the domains of Natural Language Processing as well as Machine Learning. In this paper emotion recognition based on textual data and the techniques used in emotion detection are discussed. The results verified that the comparative study could be further used in real-time behavioral facial emotion and intensity of emotion recognition.

Keywords: automatic facial emotion recognition, intensity of emotion recognition, behavioral biometrical systems, machine learning.

I. INTRODUCTION

Emotion Detection will play a promising role in the field of Artificial Intelligence, especially in the case of Human-Machine Interface development. For Emotion Detection from an artificial intelligence different parameter should be taken into consideration. Various types of techniques are used to detect emotions from a human being like facial expressions, body movements, blood pressure, heart beat and textual information. This paper focuses on the emotion detection from textual information. Nowadays within the Internet there's an immense amount of textual data. It's fascinating to extract emotion from various goals like those of business. As an example, in luxury merchandise, the emotion aspect as brand, individuality and prestige for purchasing confirmations, are lot necessary than other aspects like technical, functional or price. In such conditions buyers happy to shop for a product even with high costs. Emotion selling aims to simulate emotions in clients for tying them to brand and then increase the selling of service/product. While nice strides were created as in emotion recognition exploitation multimodal sources[1], such as: face, voice or gestures, there's not yet a strong enough text - based feeling recognition solution, capable of detecting emotions from text, with high accuracy [2], in spite of the text size, and taking into consideration context or one's type of expression. There are four basic methods to detect emotions from text:1) Keyword based detection, 2) learning-based detection, 3) lexical affinity method, hybrid detection. Each and every method contains some strong and weak points while detecting emotions from text. Hybrid Method is the most likely method [3] to get a high accuracy result, as it includes the strength of combined strength of two or more methods. In that also main difficulty is to find the most effective combination. In all the methods these challenges generate problem related to emotion detection:

- **Collection of Data [4]**: what data should be used for feature extraction? And how to cope up with the continuous changes or evolution of textual expressions used in everyday exchanges?
- **Features Choices**: which type of emotion indicators can be present in a speech? How contextual data can be extracted? How to combine those features to get a good result?
- **Labeling of Emotions**: what emotions are going to be assigned in a piece of text? Especially in the case of multiple word combination. And what categories of emotions to be used for the training dataset?
- **Machine learning classifier**: What is the best classifier to use for various textual data? More than one classifier should be used?

II. EMOTION DETECTION METHODS

Emotion detection approaches use or modify concept and general algorithm created for subjectivity and sentimental analysis. There are

many approaches that are being used and explored. However, many of the approaches have few similarities in them. Some of the methods available are presented here.

2.1 Keyword-based Methods

Keywords based approaches use synonyms and antonyms are WordNet to determine word sentiments based on a set of seed opinion words. In a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms [5] in WordNet to predict the semantic orientation of adjective. In WordNet the adjectives are in bipolar cluster form of organization and have synonyms have same orientation. As all the adjectives are linked and it form a pattern and leads to the emotion which the word depict.

2.2 Vector Space Model

Categorical classification is used in the approach of Vector Space Model(VSM). Matrix of co occurrence frequency vectors are used to representing the dataset dimensionally. Words are represented by rows and the columns can represent sentence, paragraph or documents [6]. Therefore, the column and the row depict a relationship. VSM weighs these frequencies using the tf-idf weighting schema. The tf-idf score is the weight of each word in terms of its importance within the dataset of documents. The score is broken down into tf and idf. The tf stands for term frequency and is the frequency of a term within a document. The equation for calculating tf is as follows:

$$tf = nt,d/ kd$$

In this equation, nt, dis the number of times the term, t, appears in the document, d, and kd is the total number of words in the document, d.

2.2.1 PMI Pointwise Mutual Information

Adjectives with same polarity tend to appear together. The affect words (adjectives, nouns, verbs and adverbs) that frequently co-occur together have the same emotional tendency [7]. If two words co-occur more frequently, they tend to be semantically related. There are various models for measuring semantic relatedness and although they use different algorithms [8], they are all fundamentally based on the principle that a word's meaning can be induced by observing its statistical usage across a large sample of language. Point wise Mutual Information (PMI) is a simple information-theoretic measure of semantic relatedness that measures the similarity between two terms by using the probability of co-occurrence. Mathematically, the PMI between two words x and y is calculated as follows:

$$PMI (x, y) = co\text{-}occurrence (x, y) / (occurrence (x) * occurrence (y))$$

where occurrence (x) is the number of times that x appears in a corpus, and co-occurrence (x, y) is the number of times that x and y co-occur within a specified window1 in the corpus. The corpus can be domain-dependent or general depending on the task at hand.

2.2.2 Learning-based Method

Learning-based methods are being used to formulate the problem differently. Originally the problem was to determine emotions from input texts but now the problem is to classify the input texts into different emotions. Unlike keyword-based detection methods, learning-based methods try to detect emotions based on a previously trained classifier, which apply various theories of machine learning such as support vector machines and conditional random fields. To determine which emotion category should the input text belongs.

III. PROBLEM STATEMENT

Sentiment analysis of a text can only say if a particular sentence conveys positive or a negative polarity. If we can classify the text furthermore based on the emotion of the content, it can be used by a product/brand/public figure to make necessary improvements in their respective fields [9]. With the help of this information, the perspective of the users towards that brand can be improved in a positive way. Chatbots are being extensively used for providing services to the customers in various sectors. Identifying the emotion of the user can help them analyze how well they are able to meet the requirements of the customers and make necessary changes to the system [10]. In case the emotion of the customer turns out to be either Angry/Sad they can connect the customer to an executive to take quick actions. Data: We have grouped below four datasets.

1) **Novel Dataset**: This dataset consists of different phrases from a novel. These phrases have been labelled according to their emotion. This dataset has small number of records [11].

2) **Twitter Dataset**: This dataset was collected using Twitter public streaming API. The collected tweets were automatically labelled using the emotion hashtags at the end of each tweet. It consists of 20K records, as this dataset [12] have been already labelled, we have only considered 5 labels namely, Joy, Sad, Surprise, Fear, and Anger.

3) **Kaggle Dataset**: It consists of 4 Million records of labelled data consisting five emotions. We randomly selected 6000 records of each emotion for the model. So that every emotion is normalized. We have reduced the dataset from 2M to 25K due to data processing and modelling constraints by system.

4) **Custom Dataset**: We have also added few of our custom data by referring through blogs, news feeds and custom sentences for initial analysis which we felt as necessary in the dataset. After collecting the data, we have normalized the emotions to Joy, Sad, Surprised, Angry, and Fear so that the data from all the sources will have the same labels.

IV. METHODOLOGY

We have followed below mentioned approaches as a part of word embedding which is one of the language modelling and feature learning technique where words or phrases from the vocabulary are mapped to vectors of real numbers.

1) Frequency based Word Embedding: In this embedding, we have used Count Vector (Bag of Words) and TF-IDF vector (Term Frequency – Inverse Document Frequency) in which we initially measure how frequently a term occurs among all sentences and determine how important a word is to predict the emotion of a given sentence.

2) Prediction based Word Embedding (Word2Vec): In this embedding, all the words are converted into vector from in which similar words share the same spatial position. These word embedded vectors are fed to a Convolution Neural Network (CNN) model which trains the model by processing the vectors thereby, providing the predicted emotions.

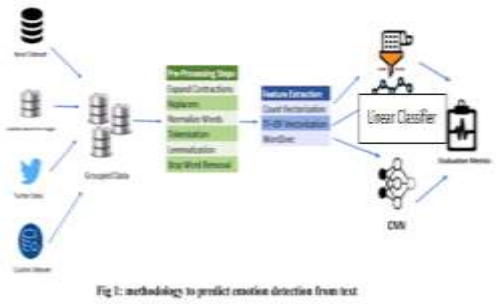


Fig 1: methodology to predict emotion detection from text

Pre-processing: It can be inferred as converting data to a format that a model can understand. Below is the flow of data pre-processing before loading the data into the model:

- **NormalizeWords**: Created a function which will remove non-ASCII values. There may be emoji's, punctuation marks etc. All these characters except letters are removed from the sentences and are converted to lower case letters.
- **RepeatReplacer**: Replaced repeated words with their respective root words. Example: Words like Happyyyyyyyyyy is replaced with happy ExpandContractions : Expanded the contracted words, so that the true meaning of the word will not be changed, and true emotion of the sentence can be captured. Example: Don't is replaced with Do not. Haven't is replaced with Have not etc.
- **Stop word Removal**: A stop word is a commonly used word that a search engine has been programmed to ignore so that we can save processing time, space and avoid giving weights to non-significant terms. For instance, words like is, an, the, and so on are present multiple times in a document which don't contribute significantly in predicting the emotion [13]. Hence, we have removed these stop words. However, negative words like "Not" have significance as they might convey negative emotion (not happy). Hence, we have removed possible negative words from the list of stop words.
- **Tokenization:** We have used Tokenizer to convert the words in a sentence into tokens so that it can be used as an input to a Lemmatizer.
- **Lemmatization**: It is the process of grouping the changed form of words into root word based on the context which helps in providing term frequency for better prediction of the label. As a part of pre-processing, we have tokenized the sentences and tagged with their respective parts of speech and then converted them into their root words [14] accordingly. Example: Words go, going, gone belong to same root word "go". Hence, one sentence might contain word "going" in it and other might contain "go" in it. By converting them to their respective root word "go", we will be able to get rid of unnecessary weights for non-significant words.
- **TF-IDF Vectorisation**: TF-IDF stands for Term frequency and Inverse document frequency. Term Frequency (TF) measures how frequently a word occurs in a document. There are few words which occur many times in a document, however have less information in predicting emotion. Hence, IDF measure is used to decrease the weight for commonly used words and increase the weight for words that are not used much in a set of documents. We have used TFIDF model with bigrams to get the vectorizer.
- **Word2vec**: Word2vec is a group of related models which are used to produce word embeddings. These models are two-layer Neural networks that are trained to reconstruct linguistic context of words. Usually word2vec takes large data as input and produces a vector space which is typically of several hundred dimensions. Each word in the data is assigned a corresponding vector in the space. These word vectors are positioned in the vector space

in such a way that words which share common context are near to one another in the space.

- **Continuous Bag of Words Model**: In continuous bag of words, the current word is predicted by the model using a window of surrounding words. This method takes context of each word as the input and tries to predict the word corresponding to the context.
- **Skip-Gram Model**: The Skip gram model takes every word in the data and takes one-by-one the words that surround it within a defined 'window'. This is then feed to a neural network that after training will predict the probability for each word to appear in the window around the focus word. In our model after finding the word vectors, we have concatenated the CBOW model and SkipGram Model representations to construct a dictionary so that necessary features can be extracted.

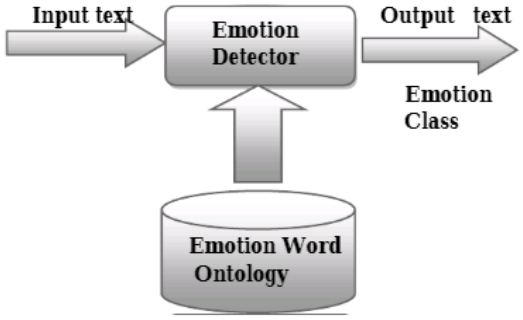


Figure 2. Proposed Architecture

Emotion Ontology

Ontology is an explicit specification of conceptualization. Ontologies have definitional aspects like high level schemas and aspects like entities and attributes [16]; interrelationship is between entities, domain vocabulary. Ontologies provide an understanding of particular domain. Ontologies allow the domain to be communicated between persons, institutions, and application systems. Emotion word hierarchy is converted into ontology. This emotion word hierarchy is developed by W.G. parrot. Protégé [13], an ontology development tool is used to develop emotion ontology. Proposed ontology has class and subclass relationship format. Emotion classes at the primary level in emotion hierarchy are at the top of emotion ontology and emotion classes at the tertiary level are at the bottom of ontology. High weight age is assigned to the upper level emotion classes and low to the lower level emotion classes.

Emotion Detector Algorithm

Emotion of the textual data can be recognized with the help of this emotion detection algorithm.

The algorithm calculates weight for particular emotion by adding weights assigned at each level of hierarchy and also calculates same for its counter emotion, then compares the both scores and greater one is taken as the detected emotion.

Parameters Used

Algorithm is to calculate weight age to be assigned to different emotion words so that they can be sorted according to it. Certain parameters are required for this purpose. The first step is calculation of parameters. This task is achieved with the help of Jena library which allows traversal and parsing of ontology.

Different parameters are calculated as follows:

Parent-Child relationship

If a text document belongs to a child; it also indirectly refers to the parent of it. Hence if a certain value is added to the child‟s score, parent score also need to be modified. This is achieved by traversing the ontology model in a breadth first manner using Jena API. When any node is encountered all of its children are retrieved. Then same method is applied to every child.

Depth in Ontology

This is required as it gives an idea about how specific is the term in relation to its corresponding Ontology structure. The more specific it is the more weight age should be given to it. This value is calculated simultaneously while traversing the ontology tree.

Frequency in Text document

This is also an important parameter as more is the frequency more will be the importance of that term. This value is calculated by parsing the text document and searching for occurrences of the words.

Algorithm

Following algorithm is proposed to calculate the score for each emotion word with the help of parameters from previous steps. This score will be directly proportional to the frequency of the term and inversely proportional to its depth in the ontology. Hence a formula devised for the mth terminology. For every primary level emotion class, a respective score will be calculated. Finally Emotion class having highest score will win the race and declared as Emotion state of the corresponding text document. Algorithm is as follows

for j → 1 to No. of Nodes [Ontology]
do parent [j] → parent of node j
child [j] → child of node j

for m → 1 to No. of Nodes [Ontology]
do freq [m] → frequency of occurrence of mth
depth [m] → depth of mth node in ontology

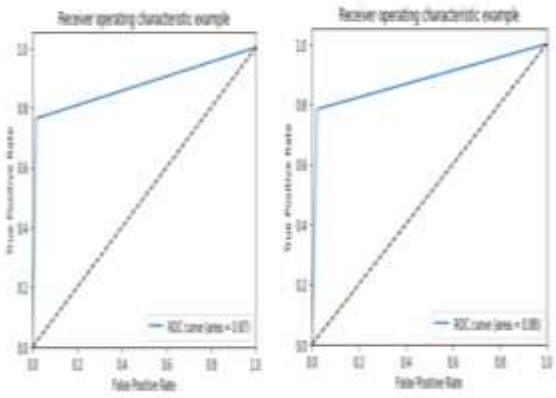
Calculate (x): for m → 1 to No. of Nodes [Ontology]
score (x) → freq [root] / depth [root]
for m → 1 to No. of parent nodes [Ontology]
score (parent) = score (parent) + score (child)
return score (parent)
for m → 1 to No. of parent nodes [ontology]
emotion class → High score [parent]
return emotion class

Where Nodes [Ontology] denotes classes in the ontology, Parent [j] denotes parent classes in the ontology, Child [j] denotes child classes in the ontology, Freq [m] denotes frequency of occurrence of mth class in text, Depth denotes depth of class into ontology, Score [parent] denotes score of parent in ontology. By proposed algorithm we can find out the score of primary emotion classes. Emotion class with highest score will be decided as the final emotion class for the blog.

Evaluation metrics	Twitter dataset		Custom dataset	
	Multinomial NB	CNN	Multinomial NB	CNN
Test Accuracy	0.29	0.198	0.64	0.62
Precision	0.28	0.28	0.603	0.59
F1-Score	0.26	0.188	0.59	0.57
Recall	0.3	0.26	0.59	0.57

V. Test results of Twitter data and Custom Data:

To ensure that model works better on all datasets, we tested the model on Twitter and Custom datasets. We found that Accuracy on Twitter dataset was around 30%. As the twitter data set contains tweets from various users which is not in standard English. Hence, the data was not completely cleaned as a result, the accuracy was low with both the models. When we tried with Custom dataset which contains data from various blogs, Books and Kaggle. We found that the accuracy was increased to 62% which was pretty good as the data contains texts of standard English. We can see that CNN model accuracy was less than Multinomial NB as deep learning model need large amount of data to train.



ROC curves of Multinomial naive Bayes and CNN models respectively(refer appendix for ROC curves of each emotion)

VI. CONCLUSION

Emotion Detection can be seen as an important field of research in human-computer interaction. A sufficient amount of work has been done by researchers to detect emotion from facial and audio information whereas recognizing emotions from textual data is still a fresh and hot research area.

In this paper, methods which are currently being used to detect emotion from text are reviewed along with their limitations and new system architecture is proposed, which would perform efficiently.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, "Emotion recognition in human-computer interaction," in IEEE Signal Processing Magazine, vol. 18(1), Jan. 2001, pp. 32-80, doi: 10.1109/79.911197

[2] Parrott, W.G, "Emotions in Social Psychology," in Psychology Press, Philadelphia 2001

[3] 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

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

[5] N. Fragopanagos, J.G. Taylor, "Emotion recognition in human–computer interaction", Department of Mathematics, King"s College, Strand, London WC2 R2LS, UK Neural Networks 18 (2005) 389–405 march 2005.

[6] C. Elliott, "The affective reasoner: a process model of emotions in a multiagent system," in Doctoral thesis, Northwestern University, Evanston, IL, May 1992.

[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] Z. Teng, F. Ren, and S. Kuroiwa, "Recognition of Emotion with SVMs," in Lecture Notes of Artificial Intelligence 4114, D.-S. Huang, K. Li, and G. W. Irwin, Eds. Springer, Berlin Heidelberg, 2006, pp. 701-710, doi: 10.1007/11816171_87.

[9] C. Yang, K. H.-Y. Lin and H.-H. Chen, "Emotion classification using web blog corpora," Proc. IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, Nov. 2007, pp. 275-278, doi: 10.1109/WI.2007.50.

[10] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining Acoustic and Language Information for Emotion Recognition," Proc. 7th International Conference on Spoken Language Processing (ICSLP02), 2002, pp.873-876.

[11] 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.

[12] C. Elliott, "The affective reasoner: a process model of emotions in a multiagent system," in Doctoral thesis, Northwestern University, Evanston, IL, May 1992

[13] Nicu Sebea, Ira Cohenb, Theo Geversa, and Thomas S. Huangc "Multimodal Approaches for Emotion Recognition: A Survey", USA

[14]http://sail.usc.edu/~kazemzad/emotion_in_text_cgi/DAL_app/index.php?overall=bad&submit_evaluation=Submit+Query".
http://www.wikipedia.org/

Authors Profile

P. Anusha, M. Tech working as an Assistant Professor of CSE Department in QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh, India.

T. Srivalli pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

G. Maneesha pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

O. Padma pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.

K.S.S. Pavan Kalyan Varma pursuing B Tech in computer science engineering from Qis college of Engineering and Technology (Autonomous & NAAC 'A' Grade), Ponduru Road, Vengamukkalapalem, Ongole, Prakasam Dist, Affiliated to Jawaharlal Nehru Technological University, Kakinada in 2016-20 respectively.