Emotion Extraction using Rule-based and SVM-KNN Algorithm

Mohini Chaudhari * and Sharvari Govilkar **
* Department of Computer Engg. PIIT, New Panvel, India
3mohinichaudhari89@gmail.com
** Department of Computer Engg. PIIT, New Panvel, India
sgovilkar@mes.ac.in

Abstract: Language and emotions generally goes hand in hand. Language forms a firm base for communication and expression of emotion. Computers are not capable of understanding and taking decisions on its own. In order to make it analyze the language and detect emotions, Natural Language Processing (NLP) techniques have been applied to automatically identify the information content in text. As emotions influences human thinking, perception and behavior, they play an important role in decision making, and learning, and can even overcome reason under stress conditions. Extraction of emotions from a given input text is one of the most popular research areas today and involves lots of challenges. We need computational approaches that would successfully analyze the online emotion rich content, recognize and aggregate relevant information, and draw useful conclusions. Recognizing emotions conveyed by a text can provide an insight into the author's intent and sentiment, and can lead to better understanding of the text's content. Thus, the idea is to propose a system that will accept single document as input in English NLP and processes the input and extract emotions by using a hybrid system.

Keywords: List five or more keywords related to your work.Emotion Detection, Natural Language Processing,Machine Learning, Support Vector Machine (SVM), Rule-based, K-Nearest Neighbor (KNN).

Introduction

Emotions are our subjective feelings and thoughts. People express emotions as part of everyday communication. Emotions can be judged by a combination of clues viz. facial expressions, prosodies, gestures, and actions. Emotions are also articulated by written texts. Considering the inspiration by work done in sentiment analysis, the thesis explores hybrid approach for detection of emotions in text. Emotion recognition in text is just one of the several dimensions of the task of making the computers make sense of emotions. It is essentially a content-based
Classification problem involving concepts from the domains of Natural Language Processing as well as Machine Learning.
Understanding and expressing emotions between people is a complex interactional phenomenon that forms an intricate web. People are able to perfectly distinguish the expressed emotions because they are capable of understanding the meaning of the words and phrases. They also are able to generate expressions and sentences for various emotions. But, developing a computer system that would analyze and interpret different emotions in a given text is a different task.
The paper presents, a combination of three machine learning techniques to detect the emotion labels present in the given text in pure English language. In section 2, related work has been discussed. Section 3 deals with the detailed proposed system architecture for the emotion detection system. Working of system is mentioned in detail in section 4. Section 5 explores accuracy obtained by system. Finally, paper is concluded in section 6.

Related work

In this section we cite the relevant past literature that use the various emotion extraction techniques. Most of the system extracts the emotions from sentences, blogs, and text document.
Ruchika Sharma and Amit Arora [1] have extended and compared the work proposed by Mullen and Collier (2013). The system comprises of two phases viz., a feature Extraction phase and a learning phase, which forms the basis of the analyzing the overall sentiment of the document. Machine learning approaches such as Naives Bayes, Maximum Entropy, SVM and Kernels were explored and out of all of these approaches Multiple Kernel outperforms. It produces an accuracy of 90% and for cross validation in 5 fold and 10 fold respectively is 92%.
Arti Buche, Dr. M. B. Chandak and Akshay Zadgaonkar [2] proposed the technique to detect and extract subjective information in text document that is opinion mining. Sentiment classification or Polarity classification is the binary classification process that labels an opinionated document as expressing either an overall positive or an overall negative opinion.

V.S. Jagtap and Karishma Pawar [3] focus on different approaches used in sentiment classification for sentence level sentiment classification. It aims at analyzing a solution for sentiment classification at a fine-grained level in which the polarity of the sentence can be given as positive, negative or neutral. According to them, Sentiment Analysis is the process of extracting knowledge from the peoples' opinions, appraisals and emotions towards the entities, events and their attributes. With the evolution of web technology, there is large amount of data present in web for the internet users. These users use the available resources in the web, and also give their feedback, thus generating additional useful information.

Pravesh Kumar Singh and Mohd. Shahid Husain [4] concludes that though opinion mining is in an inchoate stage of development, the ground is still set for dense growth for researchers. They made an attempt to appraise different techniques involved in feature extraction. According to their convergent point of view, textual classification can be easily done by Naïve Bayes, clustering is suitable for consumer services and for biological reading and interpretation SVM gives better results.

Azadeh Nikfarjam, Ehsan Emadzadeh, [5] has proposed a system which is a combination of machine learning and rule based techniques. They had described this approach as a hybrid approach for sentiment analysis. The system consists of a rule-based engine and a trained Support Vector Machine (SVM) classifier for each possible emotion. A set of syntactic and semantic features are extracted from sentences to build the rules and train the classifier. In order to generate the sentence features we propose a new approach to identify a sentence‘s clauses and its constitutive grammatical elements and to use them to measure the polarity (a quantitative measure of the positive or negative feelings reflected in it) of a given sentence. In order to handle the problem of multiclass classification, they implemented a component emotion detector for all of the 15 emotion categories.

G. Vinodhini [6] has proposed the techniques used for sentiment classification which includes Naïve Bayes (the basic idea is to estimate the probability of categories given a test document by using the joint probability of words and categories), SVM (Method of statistical classification based on the structural risk minimization principle), Centriod Classification, K-nearest neighbor Method, Winnow (Well-known as online mistaken-driven method), and Ensemble technique (Combines several base classification output to generate an integrated output).

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka [7] describe a lexical rule-based approach to recognition of emotions from text and an application of the developed Affect Analysis Model in Second Life. Based on the result of the Affect Analysis Model, the developed EmoHeart (―object‖ in Second Life) triggers animations of avatar facial expressions and visualizes emotion by heart-shaped textures. They propose a twofold focus in their research: (1) recognition of affective content conveyed through text, and (2) automatic visualization of emotional expression of avatars, which allows avoiding manual control by the user and enriching remote communications effortlessly.

Tony Mullen and Nigel Collier [8] introduces an approach to sentiment analysis which uses support vector machines (SVMs) to bring together diverse sources of potentially pertinent information, including several favorability measures for phrases and adjectives and, where available, knowledge of the topic of the text.

Joy, sadness, anger, surprise, hate, fear, so on are some of the emotion expressions. W. Gerrod Parrot [9] wrote a book in the year 2001, named ― Emotions in Social Psychology, in which he explained the emotion system and formally classified the human emotions through an emotion hierarchy in six classes at primary level i.e., Love, Sadness, Joy, Anger, Fear and surprise as shown in following table. There are many other words that also fall in secondary and tertiary levels.

Proposed system

We propose a system to extract the emotions using a Grimm's dataset which combines different fairy tales to identify nine artistic emotions as described as: pleasure, mirth, anger, energy, fear, disgust, astonish, serenity, sorrow.

The input to the system is a single text document in English Natural Language and the output is the emotion label for each sentence. The proposed approach undergoes two phases viz.,

1 Pre-processing Phase
2 Emotion Extraction Phase

Figure 1 shows the proposed architecture of the system. The system consists of two phases, i.e., Pre-processing phase and Emotion Extraction Phase. The input given to the system is in the form of text document in pure English Language. This text document in pure English Language when given to the system first enters the pre-processing phase and once the preprocessing is done, it is further sent to the emotion extraction phase.

Pre-processing Phase

The pre-processing phase consists of following modules.

Sentence Segmentation: The given input text is chunked into sentences and these sentences are stored in different text file for further use.

Tokenization: It is the process of splitting the sentences into words by identifying the spaces, comma and special symbols between the words. So list of sentences and words are maintained for further processing.

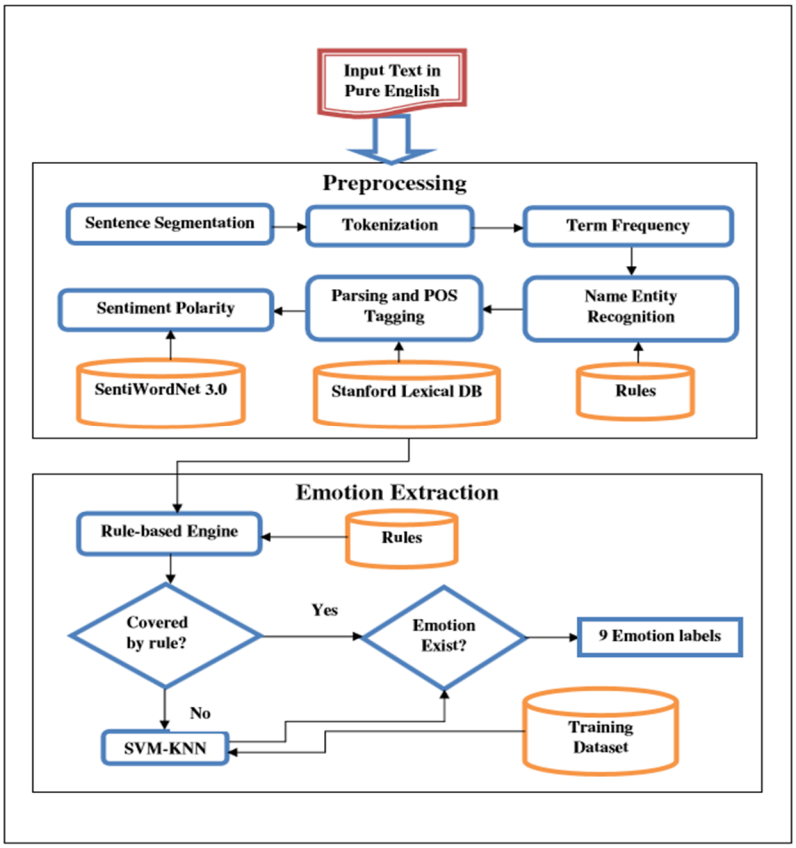


Figure1. Proposed System

Term Frequency: Term frequency is defined as the number of times a word occurs in the document. Term frequency generally depends on the length of the document, i.e., a term may occur more frequently in a large document as compared to a small document.

Name-Entity Recognition: Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors.

POS Tagging: Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, adjective, verb, etc.

Sentiment Polarity Calculation: Sentiments are generally classified into two groups such as positive or negative. But determining a sentiment as positive or negative is not just enough. It is also necessary to analyze the intensity of that sentiment i.e. how positive or how negative the sentiment is. This is done with the help of SentiWord Net 3.0 [15]. As per SentiWordNet, Sentiment Polarity is given as follows.

$$\text{Sentiment Polarity} = 1 - (\text{Pos Score} + \text{Neg Score})$$

Emotion Extraction Phase

In this phase, the preprocessed document is given as a input. This phase works in three modules, viz., Rule-based engine, SVM classifier and KNN. This is the main phase of the system wherein the emotion labels are extracted using rule-based engine, SVM and KNN. In this phase set of rules are applied to the input text document. If the rule based engine does not extract the emotional label then such sentences are forwarded to the SVM. SVM build a hyper-plane based on the polarity calculated. Here the positive values are at one side of the hyper-plane while the negative values are at the other side. While implementing this system on the Grimm's dataset it was observed that there are some polarity values that lie on the hyper-plane and the emotion labels for such valued sentences are not extracted. In order to overcome this problem, KNN algorithm is used. Those sentences whose polarity value belong to the hyper-plane where thereafter given as an input to the KNN algorithm which considers an optimal value of k and finds all the nearest neighbors and finally provides the emotion label.

Working of System

This is the main phase of the system wherein the emotion labels are extracted using rule-based engine, SVM and KNN. In this phase set of rules are applied to the input text document. If the rule based engine does not extract the emotional label then such sentences are forwarded to the SVM. SVM build a hyper-plane based on the polarity calculated. Here the positive values are at one side of the hyper-plane while the negative values are at the other side. While implementing this system on the Grimm's dataset it was observed that there are some polarity values that lie on the hyper-plane and the emotion labels for such valued sentences are not extracted. In order to overcome this problem, KNN algorithm is used. Those sentences whose polarity value belong to the hyper-plane where thereafter given as an input to the KNN algorithm which considers an optimal value of k and finds all the nearest neighbors and finally provides the emotion label.

Any text document in pure English language can be given as an input to the system. At the initial stage, when the system starts, we need to open the text document for which we wish to extract emotions. Once the text document is open, the system first checks for script validation and opens the document by eliminating those script which are not in English language. The following figure shows how the system accepts the input document.

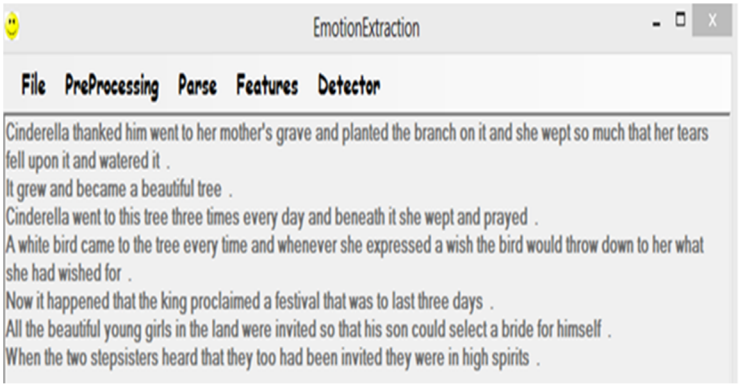


Figure 2. GUI of the system

Considering the following sample text document, let us understand the working of the system.

Sample text document

Cinderella thanked him, went to her mother's grave, and planted the branch on it, and she wept so much that her tears fell upon it and watered it. It grew and became a beautiful tree.

Step1: In order to accept the input text document, the system provides an open option in the File menu from where the user can open any text document for which he requires the emotion detection. The figure below shows the GUI for accepting a text document in pure English language as input.

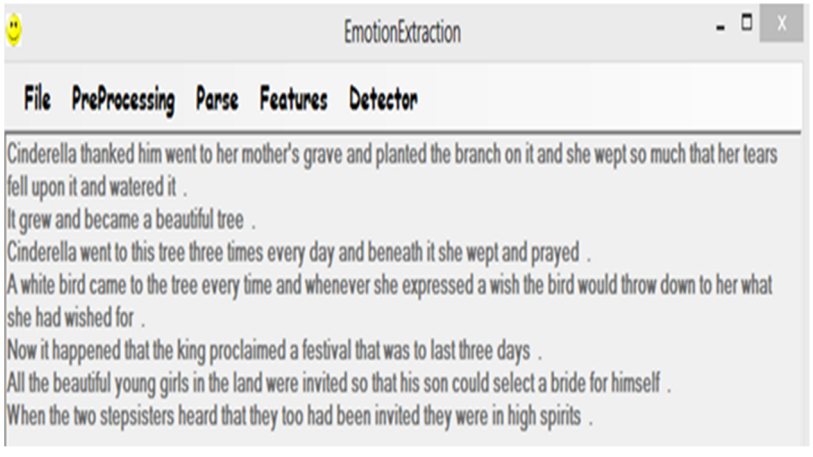


Figure 3. GUI for opening the input text document

Step2: Once the text document is given as input, the input needs to be preprocessed. The preprocessing step consists of Sentence Segmentation, Tokenization, Parsing and POS Tagging. The process of breaking a string of text in sentences is known as sentence segmentation. In English language this can be done particularly by analyzing the full stop. The figure below shows the GUI for Sentence Segmentation.

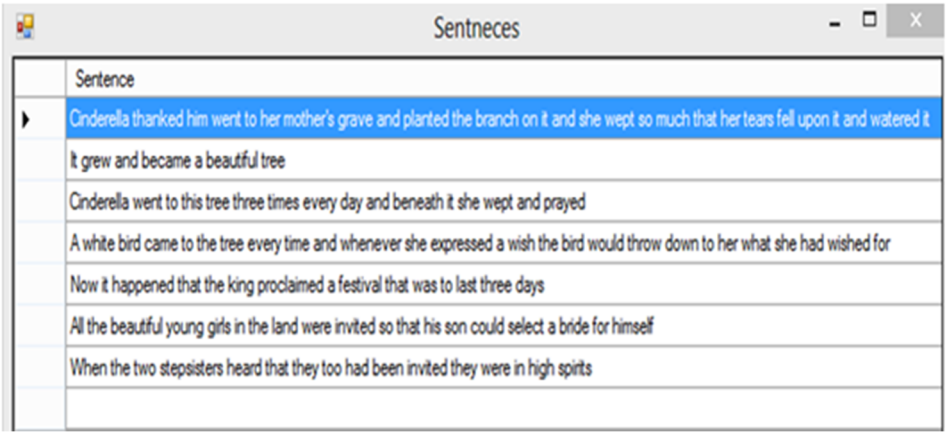


Figure 4. GUI for Sentence Segmentation

Step 3: After segmenting the complete text into sentences, tokenization is performed. It is the process of breaking the string of text into words (Tokens). This is done by analyzing the space between to the two tokens. These tokens are then saved to a different file for further use. The figure below shows the GUI for tokenization.

Figure 5. GUI for Tokenization

Step 4: After performing tokenization, Parsing is done. Parsing is the process of recognizing the sentence and how it is constructed. Constructing was a bit difficult for the system as the sentences provided in the input text document are too long, so the system has assigned different colors to each token is the sentence. The figure below shows the GUI for Parsing.
Step 5: In this step, the system assigns part-of-speech tag to each token in the sentence. This process of assigning Pos to each token is known as POS Tagging. The figure below shows the GUI for POS Tagging.
Step 6: After completing POS Tagging, the Feature Extraction phase begins. This phase consists of Term frequency calculation, Name-Entity Recognition and Sentiment Polarity Calculation. This step determines how frequently a term occurs in the document. The figure below shows the GUI for Term Frequency Calculation.
Step 7: In this step the system assigns the atomic text elements to the predefined categories such as name of persons, organization, location, etc. This process is known as Name-Entity Recognition. The figure below shows the GUI for NER.
Step 8: After recognizing the atomic elements of the text, the system calculates the sentiment polarity of each sentence in the document with the help of SentiWordNet. The figure below shows the GUI for Sentiment Polarity Calculation.

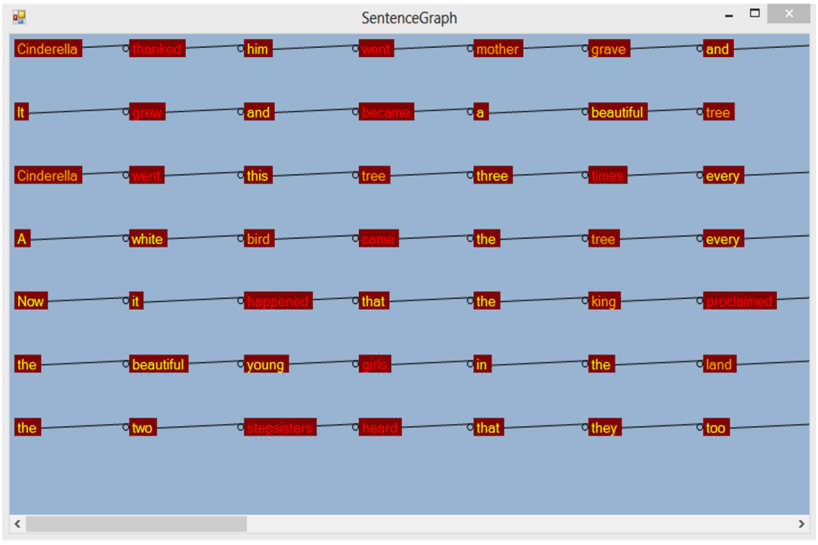


Figure 6. GUI for Parsing

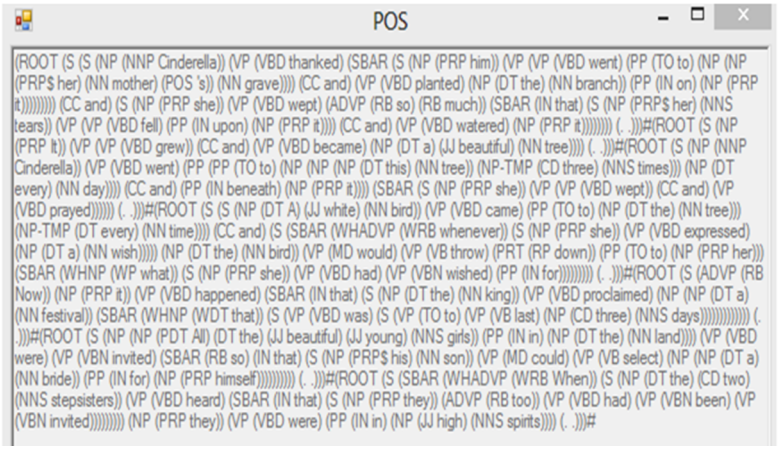


Figure 7. GUI for POS Tagging

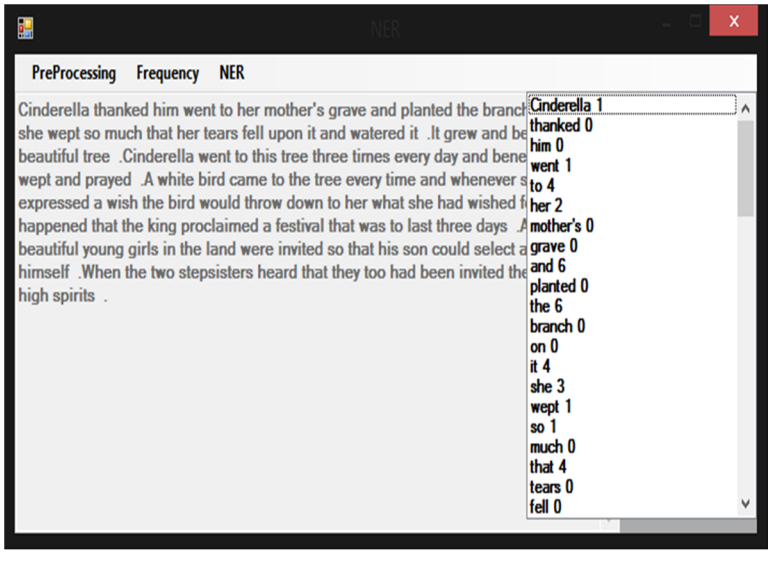


Figure 8. GUI for Term Frequency Calculation

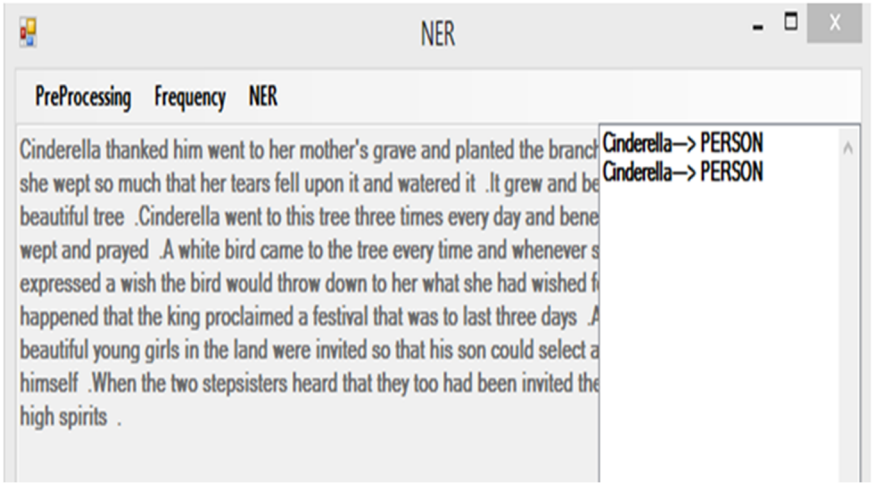
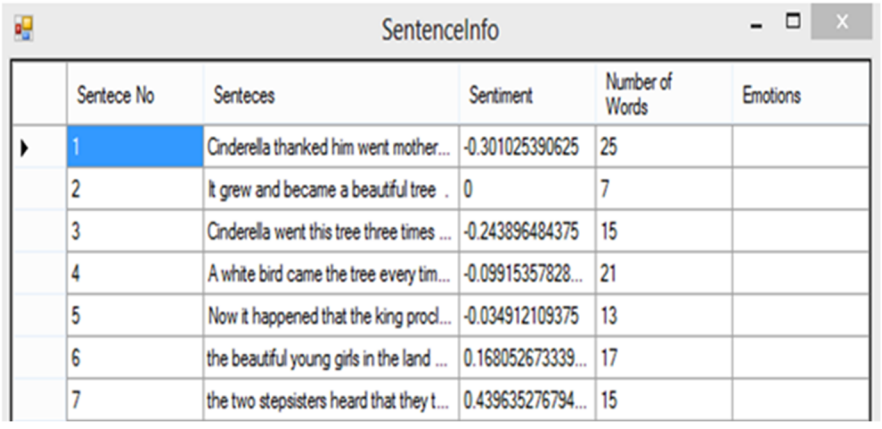


Figure 9. GUI for NER



	Sentece No	Senteces	Sentiment	Number of Words	Emotions
▶	1	Cinderella thanked him went mother...	-0.301025390625	25	
	2	It grew and became a beautiful tree .	0	7	
	3	Cinderella went this tree three times ...	-0.243896484375	15	
	4	A white bird came the tree every tim...	-0.09915357828...	21	
	5	Now it happened that the king procl...	-0.034912109375	13	
	6	the beautiful young girls in the land ...	0.168052673339...	17	
	7	the two stepsisters heard that they t...	0.439635276794...	15	

Figure 10. GUI for Sentiment Polarity Calculation

Step 9: In order to train the classifier, the system need to load the training set before analyzing the emotion labels which is done in this step. The figure below shows the GUI for loading the training set.

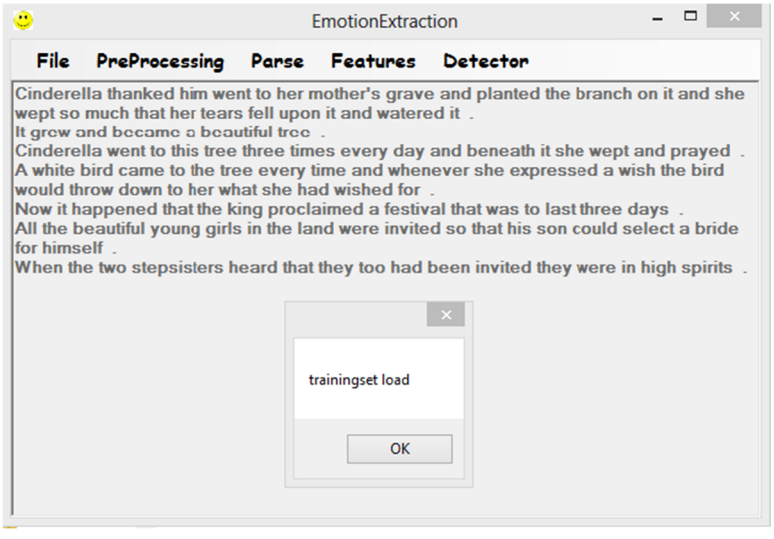
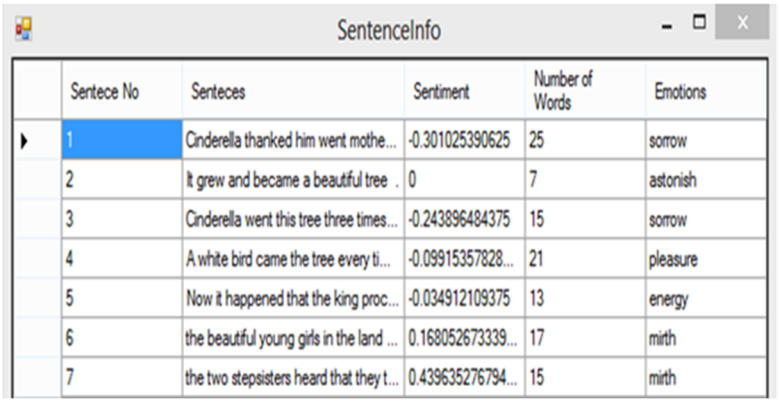


Figure 11. GUI for Loading the Training set

Step 10: This is the last step where the system analyzes each and every sentence of the text document in order to assign them with the correct emotion label. The figure below shows the GUI for Emotion Detection.



Sentece No	Senteces	Sentiment	Number of Words	Emotions
1	Cinderella thanked him went mothe...	-0.301025390625	25	sorrow
2	It grew and became a beautiful tree .	0	7	astonish
3	Cinderella went this tree three times...	-0.243896484375	15	sorrow
4	A white bird came the tree every ti...	-0.09915357828...	21	pleasure
5	Now it happened that the king proc...	-0.034912109375	13	energy
6	the beautiful young girls in the land ...	0.168052673339...	17	mirth
7	the two stepsisters heard that they t...	0.439635276794...	15	mirth

Figure 12. GUI for Emotion Detection

Performance Analysis

The system is evaluated to check whether the output generated by the current system is efficient than that of the existing system or not. Various performance measures such as accuracy, precision, recall and f-measure score has been used to check the same. To check the accuracy and efficiency of the system number of sentence in the document are used as unit for evaluation. In order to perform these evaluations, some notations are used. They are as follows: D1, D2, D3.. Represents input documents from Grimm's test dataset used for testing. DG1, DG2,.. Represents input documents based on general topics such as facebook posts, news, movie reviews, comments, etc. C1, C2,C9 - Represents the class of emotions such as Anger, Astonish, Disgust, Energy, Fear, Mirth, Pleasure, Serenity and Sorrow respectively, in which the sentences are classified. PC1, PC2, ……., PC9 Represents the predicted class of emotions such as Anger, Astonish, Disgust, Energy, Fear, Mirth, Pleasure, Serenity and Sorrow respectively, in which the sentences are classified.

1) Accuracy:

Accuracy is defined as the ratio of number of sentences for emotion labels are extracted correctly to the total number of sentences in the document.

$$\text{Accuracy} = \frac{number\ of\ sentences\ for\ which\ emotion\ labels\ are\ extracted\ correctly}{total\ number\ of\ sentences\ for\ which\ emotion\ labels\ are\ extracted\ by\ the\ algorithm} \times 100$$

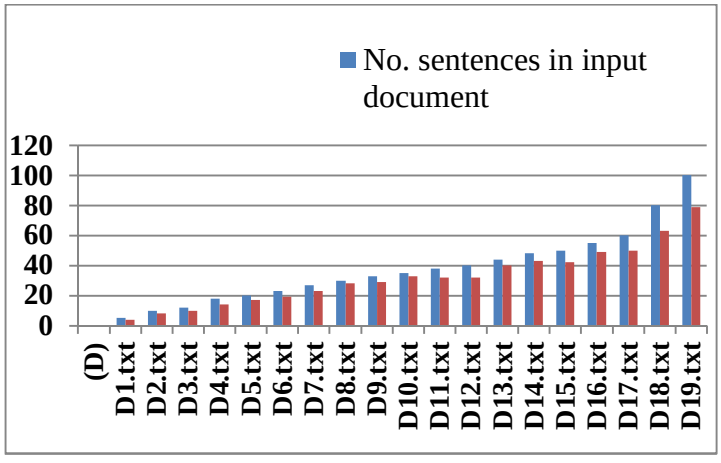


Figure 14. Evaluation Graph of the system for documents based on General Topic test dataset

2) Efficiency:

To evaluate the efficiency of the system, the accuracy of the existing system is compared with the accuracy of the current system. Table 5.2 describes the efficiency of the current system i.e. Rule-based + SVM-KNN.

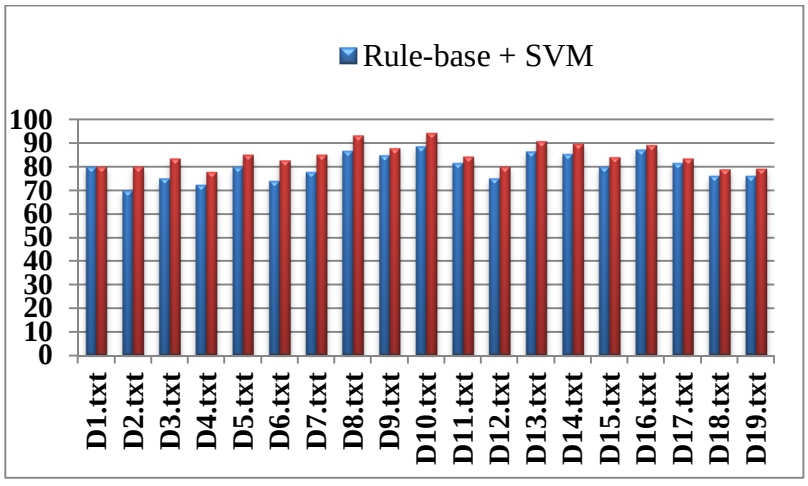


Figure 15. Efficiency comparison graph

3) Precision and Recall:

As the system deals with multi-class classification problem, the only way to evaluate the performance of the system is for each class label compute the precision and recall and analyze the individual performance on those class labels or average on class labels or average the values to get the overall precision and recall. For evaluating precision and recall for the system, 10 sample sentences are considered for each class of emotion.

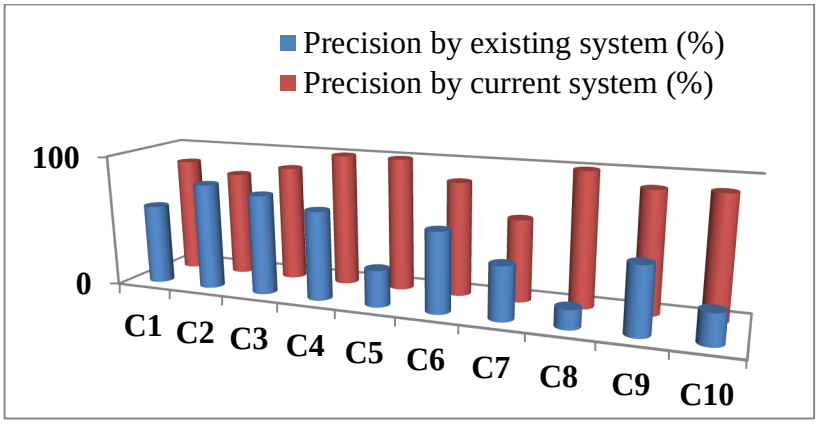


Figure 16. Precision comparison graph

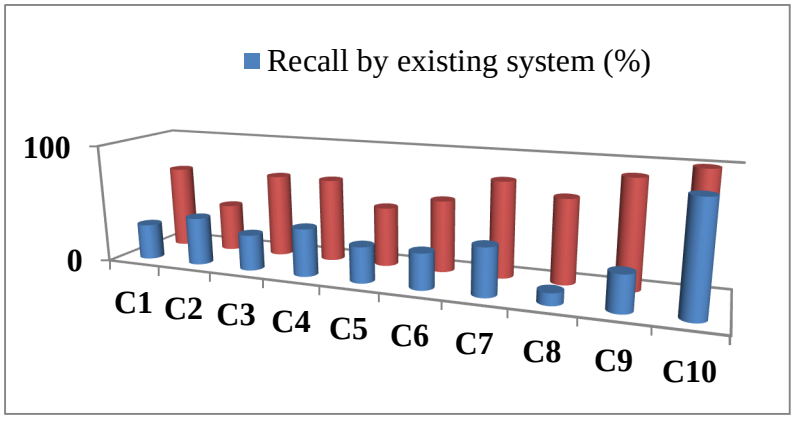


Figure 17. Recall comparison graph

Conclusion

The system has performed extremely well in terms of accuracy while showing the signs that it can be further scaled to much bigger and different dataset with better performance. In this research, the main focus was on analyzing emotions at sentence level which can further explored to paragraph level and topic level. The work can also be expanded in future by introducing methods that increase the accuracy by handling problems like change of emotions when the personal pronoun changes which still needs to be evolved properly.

Acknowledgment

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

References

[1] Ruchika Sharma and Amit Arora, ―Improve Sentiment Analysis Accuracy using Multiple Kernel Approach‖ International Journal of Computer Applications (0975 – 8887) Volume 71– No.20, June 2013

[2] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, "Opinion Mining and Analysis: A Survey" – International Journal of Natural Language Computing (IJNLC), Vol 2, No.3, June 2013

[3] Jayashri Khairnar, Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification" – International Journal of Scientific and Research Publications, Volume 2, Issue 6, June 2013.

[4] V.S.Jagtap, Karishma Pawar, Analysis of different approaches to Sentence-level Sentiment Classification – International Journal of Scientific Engineering and Technology, Volume 2, Issue 3, April 2013.

[5] Parvesh Kumar Singh, Mohd Shahid Husain, "Analytical Study of Feature Extraction Techniques in Opinion Mining" – CS & IT – CSCP 2013.

[6] Azadeh Nikfarjam, Ehsan Emadzadeh and Graciela Gonzalez, "A Hybrid System for Emotion Extraction from Suicide Notes", Biomedical Informatics Insights 2012:5 (Suppl. 1) 165–174, doi: 10.4137/BII.S8981

[7] G. Vinodhini and RM. Chandrashekharan, "Sentiment Analysis and Opinion Mining: A Survey" – International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[8] Neviarouskaya A, Prendinger H, Ishizuka M. EmoHeart: Conveying Emotions in Second Life Based on Affect Sensing from Text. Advances in Human-Computer Interaction. 2010;(2):1–13.

[9] T. Mullen and N. Collier, ―Sentiment analysis using support vector machines with diverse information sources,‖ in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04), pp. 412–418, Barcelona, Spain, July 2004.

[10] W.G. Parrot, ―Emotion in Social Psychology,‖ in Psychology Press, Philadelphia 2001, p.p. 102-105.

[11] P. Ekman, (1992), ―An Argument for Basic Emotions‖ International Journal of Cognition and Emotion, Vol. 6(3), published by Lawrence Associates Ltd, US, Jan 1992, p.p. 169- 200.

[12] Daniel Dichiu, Ana Lucia Pai, Sunita Andreea Moga, Catalin Buiu, ―A Cognitive System for Detecting Emotions in Literary Texts and Transposing them into Drawings‖, IEEE, 2010.

[13] Teng, F. Ren, and S. Kuroiwa, ―Recognition of Emotion with SVMs‖, in Lecture Notes of Artifical Intelligence 4114, D-S. Hung, K. Li, and G. W. Irwin, EDS. Springer, Berlin Heidelberg, 2006, p.p. 701-710. 33

[14] N. Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches – A Review" – International Journal of Innovations in Engineering and Technology, Volume 3, Issue 1, October 2013.

[15] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment Analysis Algorithms and Application: A Survey", 2090-4479, 2014 Production and hosting by Elsevier B.V. on behalf of Ain Shams University.

[16] SentiWordNet 3.0, http://sentiwordnet.isti.cnr.it/ (Last accessed on Spetember 13, 2015 at 3.05 P.M).

[17] Stanford Parser, http://nlp.stanford.edu:8080/parser/index.jsp, June 15, 2012. (Last accessed on September 15, 2015 at 4.00 P.M).

[18] http://www.gutenberg.org/files/2591/old/grimm10.pdf (last accessed on September 15, 2015 at 4.10 P.M).

[19] http://nlp.stanford.edu/downloads/ (last accessed on October 10, 2015 at 8.10 P.M).

[20] http://www.redbooks.ibm.com/redpapers/pdfs/redp3937.pdf (last accessed on October 10, 2015 at 10.0 P.M)