International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 5, May 2016

Classification of Textual Information for Predicting Emotional Traits and Sentiments

[1] Sumit Gupta, [2] Santanu Mandal
[1] Assistant Professor, [2] PG Student,
[1][2]Department of Computer Science & Engineering,
University Institute of Technology, The University of Burdwan
[1] sgupta@uit.buruniv.ac.in, [2] santanuit12@gmail.com

Abstract: **Our big world has seemingly become very small as humans across the globe can explore the immense possibilities owing to the revolutionary technology in terms of Virtual Community. Interactions among online users over social networking sites have increased manifold and the social being has left no stone unturned to bank upon the different offerings of a virtual world. The last decade has witnessed a colossal growth in the research arena on how to comprehend and analyze the human mind, behavior and thought-process by using information from such online communities. To do so, one such interesting tool being created by different researchers is Sentiment Analyzer. The textual information in the form of tweets, posts, messages, blogs, reviews, comments, opinions etc are being considered as sources of data and are used to understand the sentimental needs and requirements of an individual. But using texts for sentiment analysis is a very challenging task that includes intricate computational techniques. Through this paper, we aim to bring to the fore the different types of sentiment analysis being performed nowadays and the different classifiers used to classify text. We have also designed a system and fed textual inputs to it to observe how two of the classifiers yield results.**

***Keywords*: Opinion, Sentiment Analysis, Text Mining, Classifiers**

I. INTRODUCTION

Sentiment is a view, opinion, attitude, thought, belief and/or judgment prompted by feelings or emotions. It is an amalgamation of our autonomic responses, behavior as well as cultural or societal meaning. Emotion, which is simply a way of expressing oneself in life, is instinctive whereas sentiment is logical. Sentiments are brought out through emotions. As humans express their emotions to portray their sentimental side, thus, both emotions and sentiments are the essential components of any human interaction.

The emotion and sentiment analysis and recognition mechanisms have been implemented in many research areas like mood detection, facial expressions, speech recognitions, image classifications, linguistics, psychology, textual data analysis and so on. Among these, textual data is of great importance to the researchers. The Internet harbors a huge pool of unstructured text data; that is increasing in leaps and bounds. Some of the most popular virtual communities are the great sources of such unstructured text data and if we could manage to decipher human qualities from such data, we would be able to get closer to human beings and create a consummate environment.

Online users' use virtual communities to communicate with each other in the form of tweets, posts, messages, blogs, reviews, comments, opinions etc which are utilized by some companies to understand the mindset of its consumers and the thus helps them rethink on any product manufacturing policies and plan their marketing strategies accordingly. Social media has thus become an emerging phenomenon due to the rapid advances in the IT field with people sharing their opinions with each other about a wide variety of subjects, products and services. This has made it a rich resource for text mining and sentiment analysis. For example, even criminals have such accounts on social networking sites. Analyzing their accounts on web can be of immense help to investigating agencies.

The basic task in sentiment analysis is to classify the polarity of a given text at three levels- the document level, the sentence level and the feature/aspect level. The polarities associated with the text are positive, negative, or neutral. Moreover, there has been a recent emergence in analyzing and classifying texts based upon the emotional states of the users. This paper mainly deals with the latter approach to use words from texts that represent emotions and classify them to predict the sentiment of the user-

whether the person is an optimist (with positive emotional traits), a pessimist (with negative emotional traits) or an ambivalent (with mixed emotional traits).

Our proposed framework firstly identifies the sentiment by extracting the emotional words in the textual data such as chat messages or posts and subsequently classifies each word in a linear fashion as well as probabilistic fashion to analyze the pros and cons of these two approaches. We have calculated the polarity levels and polarity emotion level percentages so that we could predict the sentiment of the online user.

The rest of the paper is organized as follows: Section II discusses the previous related works by different researchers. Section III presents an overview of Sentiment Analysis highlighting the different types of classification approaches associated with it with a brief tabular comparison among some well-known Sentiment Analysis Classifiers. In Section IV we have modelled our work through a proposed architecture and an algorithm to perform a simplified linear classification on textual data from the online community. We have further tried to run our algorithm so as to perform the probabilistic classification as well. The results are showing the performance level comparison between our linear as well as probabilistic approaches. Finally, we have concluded this paper by suggesting some future research possibilities in this research area.

II. PREVIOUS RELATED WORK

A lot of work has been done in the area of sentiment analysis. This section discusses a few of the popular works related to this domain.

In paper [1], the authors had proposed a system of emotion extraction from real time chat messenger by building a lexicon of emotion conveying words. The proposed system was encoded by using nine emotion labels and was implemented in C#. The system used the rule based approach called vector space model (VSM), in which each document was represented as a vector and each dimension corresponded to a separate term.

Paper [2] proposed a machine learning approach called Decision tree to classify the given Digg dataset into two known classes of data. Affective Text was used in a three month complete crawl. These datasets were designed to have six different emotional strength, positive-negative sentimental strength and emotional intensity. Finally, the root mean square and mean of each dimension in emotion class were evaluated.

Paper [3] dealt with automatic generation of emotions from texts obtained from Social networking websites using Text Mining. The authors developed a visual image generation approach that generates images according to the emotions in the text. The authors used Naïve Bayes classifier for training the input data.

In [4], the authors discussed about the sentiment analysis for language learning using Naïve Bayes Classifier. Their proposed system was designed to classify an opinion using sentence-level classification. The system is trained to accept inputs in the form of status updates.

Paper [5] proposed a hybrid system that used two methods- keyword based method and machine learning method. Keyword based method used emotional keywords to determine the emotional state in text. If there were no emotional words then the proposed system used knowledge based artificial neural network (KBANN) which used approximate domain knowledge and high level features. The system handled eight kinds of emotions.

Ana C.E.S Lima and Leandro N.de Castro [6] presented a hybridized emotional-based and word-based approach for automatic sentimental analysis of Twitter. They have used the basic text mining techniques and Naïve Bayes classification algorithm.

In [7], the authors had proposed a text based framework for predicting human personality in three phases- Text Extraction, Text cleaning and Text Analysis. They have used Hierarchical clustering to cluster semantically similar words.

Paper [8] presented a method to assess fixed set of emotions in a tweet or a set of tweets using opinion mining. The authors had divided their framework into two main modules- Pre-processing module and Scoring module. A corpus-based method was used to find the emotion value vector of adjectives and the dictionary based method was employed to find the semantic orientation of verbs and adverbs. Once the orientation scores of adjectives, adverbs & verbs were determined, the average emotion value of tweets was calculated using a linear equation in the scoring module.

III. SENTIMENT ANALYSIS

Text generally means words, sentences or paragraphs. Text, on one hand is the input from social media that can be used to understand the mindset and emotional state of the online user because of which he or she has formed an opinion. Opinion, on the other hand is the crux for performing sentiment analysis. Whenever we use any product or an item, associate with an organization, interact

with another person or simply observe our surrounding, we tend to form an opinion about that entity. Let us first understand what the terms opinion and sentiment analysis mean in particular.

Opinion- An opinion is basically an expression (or a text) that consists of two fundamental components- target (or topic) and sentiment. A target is the entity under consideration e.g., any noun such as book, movie or person. A sentiment is the judgment on the target. For instance, if the text is "I love your never-say-die attitude", the noun "attitude" is the target and the sentiment (as conveyed in the text by the verb "love") is positive. For the text-"This movie is boring", "movie" is the target and the sentiment (as conveyed in the text by the verb "boring") is negative.

Sentiment Analysis- Sentiment Analysis (or opinion mining) is defined as the task of finding the opinions of authors about specific entities [10]. It refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. So, we see that if our objective is to perform textual-based sentiment analysis, then we need to look for the opinions in the online content and pick the sentiment within those opinions. It is widely applied to social media for a variety of applications, ranging from marketing to customer service with an aim to determine the attitude of a user or a consumer with respect to some item or the overall item-specific polarity.

Extracting emotions and sentiments from text is a classification task which comes under text mining. A large number of statistical and machine learning classifier techniques have been developed for affective computing, which includes Naive Bayes, vector-space-model, Support vector machines, decision tree and so on [9]. So, let us know what basically Text mining is.

Text Mining- Text Mining is the process of retrieving useful and interesting information, knowledge, features or patterns from the unstructured text that are collected from different sources. It usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. Text analysis involves the retrieval of information, lexical analysis to study the word frequency distributions, recognition of features/patterns, tagging/annotation, extraction of information, data mining techniques including link and association analysis, visualization, and predictive analytics. The main purpose is to turn text into data for analysis.

IV. CLASSIFICATION OF SENTIMENT ANALYSIS TOOLS:

We can classify the sentiment analysis tools into two classes:

1. Machine learning approach (or Automated) – An automated sentiment analysis system is one where an algorithm processes the text string and determines its overall sentiment without any human intervention; generally flagging comments as either positive, negative or neutral. The issue with automated sentiment analysis is its inability to differentiate between subtle nuances, usually only detectable through verbal communication, such as sarcasm.

2. Lexicon based approach (or Manual/Human) – A manual or human sentiment analysis system requires the intervention of a human element into the analysis and is required to dissect abbreviations, sarcasm, emoticons, slang etc and determine the true expressed sentiment. The downfall of human sentiment analysis is that it can be extremely time consuming. It is inherently more accurate but with less mass text analysis than the automated approach.

Based upon the nature of learning and approaches, the above two types of sentiment analysis tools are further classified [11]. The figure (Figure 1) shows the varied sub-categories of Sentiment Analysis techniques.

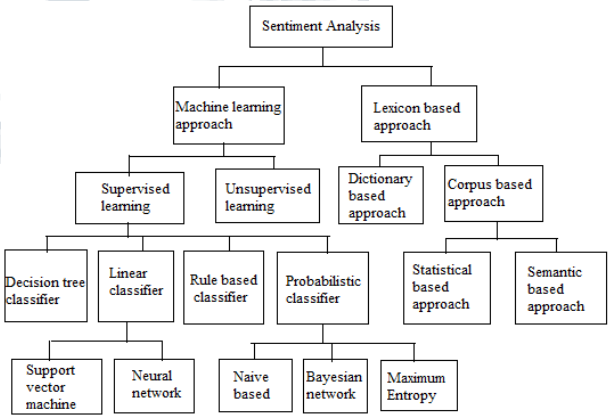

Figure 1. Sentiment Analysis Classification Techniques

Our main focus is to classify texts so that sentiments can be analyzed, rather than discussing in details about all the different techniques, we have shown in Table I, a comparison of different classifiers.

Comparison of different Classifiers:

As we are concerned with the classification of text to predict the sentiment of an individual, we must know

about the different advantages and disadvantages of different classifiers.

Classifier	Advantages	Disadvantages
Neural Network	Produce good results in complex domains. Suitable for both discrete and continuous data. Testing is very fast.	Training is relatively slow. Learned results are difficult for users to interpret. It may lead to over fitting.
Bayesian Classifier	Work well on numeric and textual data. Easy to implement. Easy computation.	Conditional independence assumption is violated. Performs very poorly.
Naïve Bayes	Very simple. Better performance. Good accuracy. Model is easy to interpret. Efficient computation. Less time required for training classifiers.	Assumptions of attributes being independent, which may not be necessarily valid.
Maximum Entropy	Good performance on experimental results and consistent result accuracy over a period of time.	Moderate time required for training classifiers.
Support Vector Machine	Very good performance on experimental results. Low dependency on data set dimensionality. Capture the inherent characteristics of the data better. Global minima vs. local minima.	Parameter tuning kernel selection. Difficult interpretation of resulting model.
Decision Tree	Easy to understand. Easy to generate rules. Reduce problem complexity.	Training time is relatively expensive. One branch. Once a mistake is made at a higher level, any sub tree is wrong. Does not handle continuous variable well. May suffer from over fitting.
K-nearest neighbor	Training time is relatively expensive. One branch. Once a mistake is made at a higher level, any sub tree is wrong. Does not handle continuous variable well. May suffer from over fitting.	Classification time is long. Difficult to find optimal value of k.
Rocchio's	Easy to implement. Very fast learner. Relevance feedback mechanism.	Low classification accuracy. Linear combination too simple. Various spelling correction techniques used.

Table I. Comparison of different classifiers

V. OUR PROPOSED METHODOLOGY

The overall architecture of the proposed system is shown in Figure 2. It consists of extracting the raw data from various texts available on social media and arranging them as per polarity traits. The oriented data is then fed to the database processing section where the words are matched with the content of Emotional Database. The Emotional Database comprises emotional tables containing emotional words (based on emotional dictionary). We have initialized the value of emotional traits as +1 for positive, -1for negative and 0 for neutral then finding out the emotion features of the sentences matching with emotional databases. If the sentence contains any emotion word then the extraction of emotion features in the sentence is done by analysing the sentence structure. After that by using text classifier, classification is done based on the emotion features. We have used Naïve Bayes classifier and Support Vector Machine (SVM) classifier for text classification purpose. The last step is the polarity classification to generate the type of emotional polarity- positive, negative or neutral.

Our Proposed Architecture:

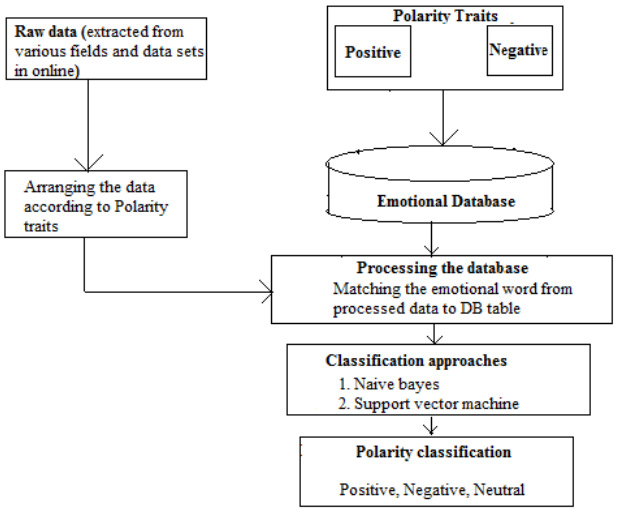

Figure 2. Our Proposed Architecture

Our Proposed Algorithm:

We have proposed an algorithm in which we have tried to incorporate the linear classification approach of Support Vector Machine Classifier in a more simplistic manner so that we could extract text from social media and predict the sentiment and emotional state of an online user. The proposed algorithm is as follows:

Input: Unstructured text $T := (m_1, m_2, \ldots, m_n)$
[where m_i represents the i-th message]

Database: Training Dataset
Positive Dataset: *pword* : = $(p_1, p_2, \ldots, p_n)$
[where p_j represents the j-th positive emotional word]
Negative Dataset: *nword* : = $(n_1, n_2, \ldots, n_n)$
[where n_j represents the j-th negative emotional word]

Output: Polarity Levels: *posp*, *negp*, *neup*
Polarity Emotion level Percentage: $posp_{per}$, $negp_{per}$
[where *posp*, *negp* and *neup* represents the positive polarity level, negative polarity level and neutral polarity level respectively and $posp_{per}$ *and* $negp_{per}$ represent the positive and negative emotion level percentages respectively]

Step 1: [Apply text mining process to T to produce structured words pertaining to each i-th message and set variable count to the message count that counts the number of messages in T]
count : = msg_count(T); $m_i := (w_{i1}, w_{i2}, \ldots, w_{in})$ [where w_{ik} is the k-th word of the i-th message]

Step 2: [Initialize all emotional counter variables to zero]
pos : = 0, neg : = 0, total1 : = 0; total2 : = 0;

Step 3: [Take each i-th message $\boldsymbol{m_i}$ for i= 1 to count and set variable *total1* to the word count of $\boldsymbol{m_i}$]
total1 : = msg_word_*count(* m_i);

Step 4: Load into the application all the positive emotional words from the Positive dataset *pword*

Step 5: Fetch each word $\boldsymbol{w_{ik}}$ of the message for j = 1 to *total1* from the application and proceed for matching with the already loaded dataset *pword*

Step 6: If there is any positive word matching, increment the pos counter value by 1.
pos : = *pos* + 1;

Step 7: Increment the value of j and repeat steps 3 to 5 until all the words of the i-th message are checked for positive word matching. The iteration stops when j becomes greater than *total1*.

Step 8: Similarly, load into the application all the negative emotional words from the Negative dataset *nword*

Step 9: Fetch each word $\boldsymbol{w_{ik}}$ of the message for j = 1 to *total1* from the application and proceed for matching with the already loaded dataset *nword*

Step 10: If there is any negative word matching, increment the neg counter by 1.
neg : = neg + 1;

Step 11: Increment the value of j and repeat steps 8 to 10 until all the words of the i-th message are checked for negative word matching. The iteration stops when j becomes greater than *total1*.

Step 12: Calculate the total number of words that are matched.
total2 : = total2 + total1;

Step 13: Increment the value of i and go to step 3 to continue the iteration process for the next i-th message until all the messages are taken as input. The iteration stops when i becomes greater than count.

Step 14: [Compute the polarity levels for each type of polarities viz. positive, negative and neutral]
posp : = pos/total2; negp := neg/total2;

neup : = posp - negp;

Step 15: (a) If *posp* and *negp* result in zero, then return a prompt message to the user that the message has no emotion word as its content and exit.

(b) If *neup* results in zero, then return a prompt message to the user that the user has mixed emotional state and is an ambivalent person and exit.

(c) Otherwise compute the percentage of emotions.

$posp_{per}$ *: =* *posp**100; $negp_{per}$ *: =* *negp**100;

Step 16: Predict the sentiment as "Positive" if $posp_{per}$ is greater than $negp_{per}$ and prompt a message that the person is an optimist, or else predict the sentiment as "Negative" and print that the person is a pessimist.

Moreover, we have modified the same algorithm to perform probabilistic classification using our simplified version of Naïve Bayes Classifier by altering the Step 14 of this algorithm to involve the probabilistic essence. The change in the algorithmic step proposed by us is as follows:

Step 14: [Compute the polarity levels for each type of polarities viz. positive, negative and neutral]

posp : = log(pos)/total2; negp := log(neg)/total2;

neup := posp - negp;

VI. IMPLEMENTATION & RESULT

Datasets: The training datasets applied in Sentiment Analysis are a relevant item in this field. Recently several evaluation datasets from blogs, tweets, chat, posts, movie reviews and product reviews have been made publicly available. Some of the well-known text-based datasets available online are the SemEval test set, the ISEAR dataset, the Digg dataset from Cyber emotion, the SemEval Affective Text-2007, the standard Sentiment140 dataset to name a few. In this paper we have used few portions of the UMICH S1650-Sentiment Analysis dataset [12] comprising around 7086 lines of training data and 33052 lines of test data to form the positive and negative word database. The Twitter dataset [13] has been also used by us. Moreover we have considered several online sources like Twitter, Facebook, Review forums etc and collected data to build our database with approximately 7000 positive and negative emotional words. Then we have run 200 sentences (containing both positive as well as negative words) on the system as test cases to calculate the performance and accuracy of our proposed versions of linear and probabilistic classifiers. The datasets are made available for non-commercial and research purposes only.

Result & Analysis: We have implemented our algorithm and fed different test cases to analyze the difference in performance characteristics between the linear and probabilistic classification approaches. Metrics such as Accuracy, Error rate, Recall, Precision and F-score (or harmonic mean of precision and recall) are calculated. This is the normal option to compute these indexes which are depending on the confusion matrix. Confusion matrix table is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

The following tables- Table II shows the confusion matrix for the classes emotion=positive and emotion=negative, Table III shows the confusion matrix with totals for Positive and Negative tuples and Table IV is the performance comparison table showing the different metrics calculated by us to compare the linear and probabilistic classification approaches.

		Predicted class		
		Emotion= positive	Emotion= negative	Total
Actual class	Emotion =positive	84	16	100
	Emotion= negative	7	93	100
	Total	91	109	200

		Predicted Class		
		Yes	No	Total
Actual class	Yes	Tp	Fn	P
	No	Fp	Tn	N
	Total	P'	N'	P+N

Table II. Confusion Matrix for the Classes Emotion=Positive and Emotion=Negative

Table III. Confusion Matrix shown with totals for Positive and Negative tuples

Measure	Formula	Linear Classifier	Probabilistic Classifier
Accuracy	$\frac{Tp + Tn}{P + N}$	88.50%	97.69%
Error rate	$\frac{Fp + Fn}{P + N}$	11.50%	2.31%
Recall	$\frac{Tp}{P}$	84%	96.21%
Precision	$\frac{Tp}{Tp + Fp}$	92.30%	98.22%
F-score	$\frac{2 * Precision * I}{Precision + Re}$	87.91%	96.70%

Table IV. Performance Comparison of Linear and Probabilistic Classifiers

Hence we can conclude that for huge amount of data, probabilistic classifiers yield best output over linear classifiers. As we know Naïve Bayes classifier is a simple and powerful probabilistic classification technique that we use for testing and problem classification purpose. Due to its simplicity, understanding enhances and better results are produced. Thus, probabilistic models are easy to build and make predictions faster than their linear counterparts.

V. FUTURE SCOPE & CONCLUSION

Sentiment Analysis has been a research interest for recent years and also involves practical applications in various fields. In this paper we have discussed a method to calculate the emotions through a text. Sentiment Analysis deals with evaluating whether this expressed opinion about the entity has positive, negative or neutral polarity which in turn helps us to understand the sentimental demeanor of an individual. For the future endeavor, we can definitely see a lot of scope in terms of including other factors of texts like the use of sarcasm, ambiguous grammar and contextual deviation etc so that the sentiment analysis system tends to become more realistic and is able to predict the subjective side of human beings objectively and in a more precise and perfect manner.

REFERENCES

[1] 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.

[2] Sriram, Sivaraman, and Xiaobu Yuan, "An enhanced approach for classifying emotions using customized decision tree algorithm", 2012 Proceedings of IEEE Southeastcon, 2012.

[3] Tejasvini Patil and Sachin Patil, "Automatic generation of emotions for social networking websites using Text Mining", IEEE 4th ICCCNT – 2013, July 4 - 6, 2013.

[4] Christos Troussas, Maria Virvou, "Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning", 2012.

[5] Yong-Soo Seol and Dong-Joo Kim, "Emotion recognition from text using Knowledge-based ANN", ITC-CSCC, 2008.

[6] Ana C.E.S Lima and Leandro N.de Castro, "Automatic Sentiment Analysis of Twitter Messages", IEEE Fourth International Conference on Computational Aspect .of Social Networks (CASoN), p.52-57, 2012.

[7] Ramya Sharada K, Arti Arya, Ragini S,Harish Kumar & Abinaya G, "A Text Analysis Based Seamless Framework for Predicting Human Personality Traits from Social Networking Sites". I.J. Information Technology and Computer Science, 2012.

[8] Akshi Kumar, Prakhar Dogra and Vikrant Dabas, "Emotion Analysis of Twitter using Opinion Mining", IEEE, 2015.

[9] Diksha Sahni, Gaurav Aggarwal, "Recognizing Emotions and Sentiments in Text: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering 5(5), pp. 201-205, May- 2015.

[10] Ronen Feldman, "Techniques and Applications for Sentiment Analysis", Communications of the ACM, pg-82-90, 2013.

[11] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093–1113.

[12] https://inclass.kaggle.com/c/si650winter11/data

[13]Https://Www.Mpi-Inf.Mpg.De/~Smukherjee/Data/Twitter-Data.Tar.Gz