International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Sentiment Expression via Emoticons on Social Media: Twitter

Mr Amritkumar Tupsoundarya[1], Prof. Padma S. Dandannavar[2]
[1,2]*Computer Science & Engineering, KLS Gogte Institute of Technology, Belagavi (An Autonomous Institution under Visvesvaraya Technological University, Belagavi)*

Abstract: ***Extensive use of emoticons such as :) :-) ;) :D :(in the "social media" have drawn attentions of researchers for using them widely in sentiment analysis and other "Natural Language Processing (NLP)" tasks as features entries of sentiment lexicons or to "machine learning algorithms". Although emoticons are a common and strong sign of expression of feelings/sentiments in social platforms, the relation between expression of sentiments and emoticons is not always clear & hence both must be included to draw nearly exact sentiment. Therefore, any such algorithm that deals with expressing the sentiment must take emoticons into account but one has to be extremely careful with the emoticons that are to be considered. In this work we first analyzed & observed the emoticons that are frequently being used in Twitter dataset records, these emoticons were then analyzed for their appearance in texts within the twitter datasets and later we examined the relations among the nature of the sentiment & emoticon used, and also the contexts in which the emoticons are used. Several analysis were then performed to analyze the effect of emoticons on tweets using machine learning techniques. Finally, the overall performance analysis of the classifier was computed using the confusion matrix.***

Keywords*: Sentiment; emoticon; sentiment analysis; social media; Twitter*

I. INTRODUCTION

Emoticons like :) ;) :-) and :(, are often used online in social media, instant messaging (e.g. Twitter, WhatsApp, and Skype), blogs, forums and other types of online social interactions and they are usually direct signals of the feelings. The emoticons in the text have been widely used by NLP researchers in such tasks as analyzing emotions as features of machine learning algorithms or as sentimental lexicon entries for rule-based approaches. Different tools and online communities may elicit varied degrees of usage of emoticon. Twitter, a micro-blogging site, is one of the most popular social media. Access to its vast amount of user-generated data is essential to understand the behavior of users and the expressed mood. Having authority to access over millions of tweets per day (via the Tweepy Twitter API), we thought it would be interesting to understand the emoticon evolving on Twitter nowadays, how users express and perceive the feelings/sentiments of emoticons, and whether emoticons are a reliable indicator of the class of the sentiment, if the polarity of feelings/sentiments can be used.

II. PREVIOUS WORK

There are ample studies carried on "sentiment analysis" in recent past years [1-11]. In particular, the applications of Sentiment Analysis has attracted many interests on social platforms, both in industry and academia. In industries, researches like [1] surveyed a group of participants for their perceived sentiments polarity of most frequently used emoticons and have also performed analysis to examine clustering of words and emoticons using word2vec and K-means algorithms, and compared the sentiment polarity of microblog posts before and after the emoticons were removed from the text. They have used Naïve Bayes classifier for classification, and have also tested the hypothesis that removing emoticons from text hurts the sentiment classification. The work conclude that classifier becomes less accurate when emoticons were removed with the caveat that classifier was trained with less positive/negative samples. Whereas in academics, the work [2] says that, it is very difficult task to analyze exact sentiments attached with that natural language. They state that Sentiment Analysis is a study of people's attitude, opinions, and emotions that helps classify them as 'positive', 'negative' or 'neutral'. In many other studies conducted in the past, emoticons have played an important role in training machine learning classifiers and the development of sentimental lexicons [3]. It has been assumed that emoticons are reliable indicators of feelings/sentiments. Several repeated attempts were made based on the emoticon usage to construct the corpus of sentiments [4]. However, previous studies do not directly examine the relationships between the polarity of sentiments and emoticons in social media and the role of emoticons in such context. Some of the previous work [5] are useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about their products before making the purchase(s), where they have used 'Parts Of Speech' (POS)-specific prior polarity features and 'tree-

kernel' for preventing the requirement of 'monotonous feature engineering'. Work [6] compares the sentence polarity before and after emoticons are removed, clustering words and emoticons, also performs sentiment analysis based on meaning obtained during clustering. Conclusion made was that, emoticons play a major role in sentiment analysis and clustering of words & emoticons could provide a valuable insight into the relationship between emoticons and sentiment polarity. Some work [7] proposes a framework for tweet sentiment classification that adopts a hybrid approach based on the Singular Value Decomposition, to reduce the dimensionality of the data. Once the original data is represented in a reduced dimensional space, the Extended Binary Cuckoo Search algorithm is applied to further reduce the matrix dimensionality. Other work [8] sets out to answer a series of questions regarding the nature and extent of cross-platform emoji misinterpretation, and to provide a solution that can help researchers and practitioners to overcome this platform-specific inconsistency in their analysis. The key contribution in some work [9] lies in validating the important role emoticon plays in conveying overall sentiment of a text in Twitter Sentiment Analysis (TSA) though a series of experiments, whereas few work [10] only compares various approaches like Machine Learning, Lexical and Hybrid approaches for “a-state-of-art” study on sentiment analysis. Some previous researches [11] says that, most text based methods of analysis may not be useful for sentiment analysis in various other domains. To make a significant progress, one still need novel ideas. Using twitter names and hashtags to collect training data can provide better results. Also adding symbol analysis using emoticons and emoji characters can significantly increase the precision of recognizing of emotions. They conclude that most successful algorithms will be probably integration of natural language processing methods and symbol analysis.

III.SYSTEM DESIGN

In this work, the area of focus is on how a machine learning algorithm uses supervised learning techniques to perform sentiment analysis on data that is either collected and stored in a comma separated values file or retrieved online. A broad view of this work can be understood by the System Block Diagram as shown in figure 1.

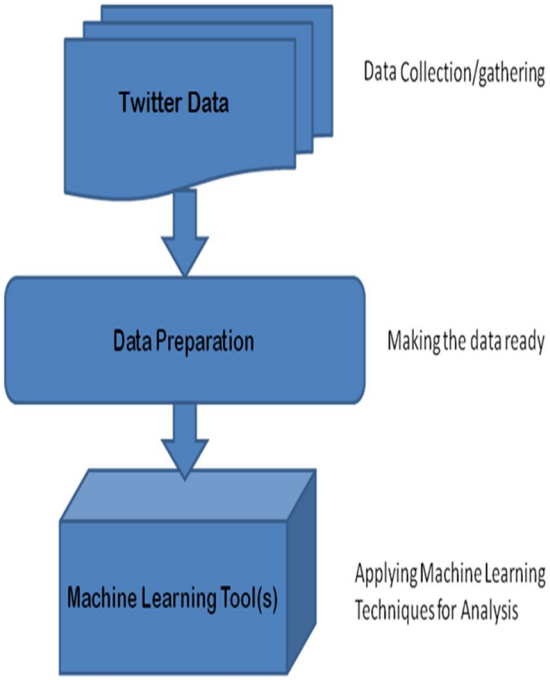

Figure 1. System Block Diagram

The "Supervised learning algorithms" are trained with the help labeled examples, like an input for which the output expected is known. Consider an example, a piece of equipment may have data points labeled "0" as "inactive" or "1" as "active". Set of inputs with the corresponding correct outputs are supplied to the learning algorithm, and the algorithm learns by comparing its correct output with the actual outputs to find errors, then it modifies the model accordingly. Using methods such as classification, regression, prediction, and gradient enhancement. The supervised learning uses models to predict tag values for additional unlabeled data. Assisted learning is often used in applications where earlier data or historical data predicts events likely to occur in future. For example, to classify feelings as positive, negative, neutral or not sure.

A. Module Design

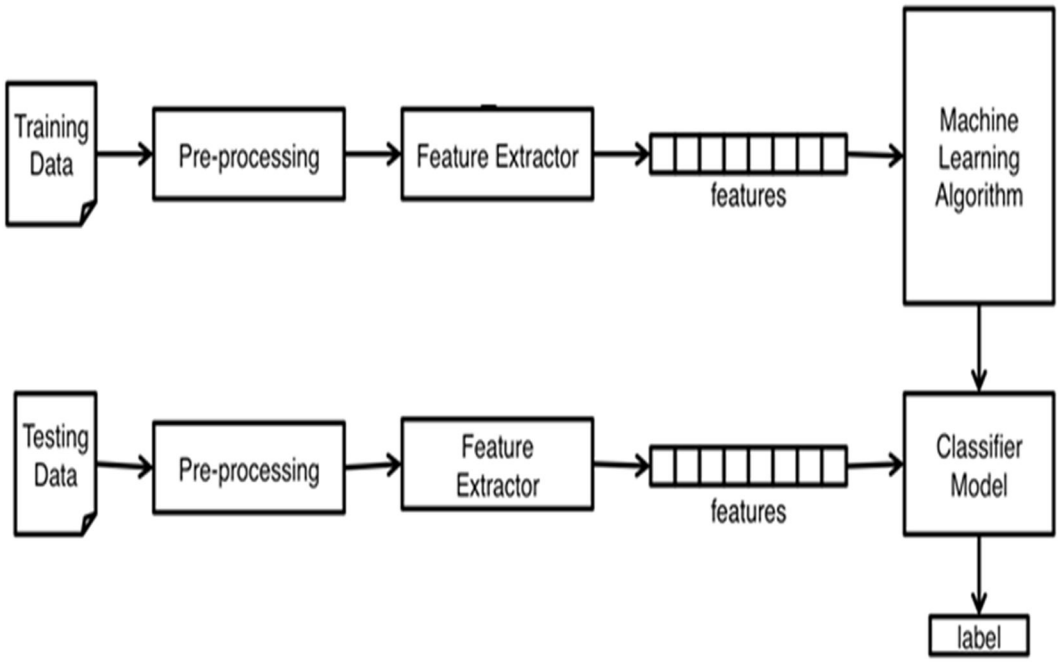

Figure 2. Module Design

Module design or modularity in design is a designing approach that splits a larger system in smaller parts known as modules. In this work important modules included are Training Data & Testing Data, Pre-processing, Feature Extractor, Features, Machine Learning Algorithm, Classifier Model and Label. Every module plays a significant role in overall implementation.

1) *Training Data:* Contains set of attributes and its instances which are given as input to classification algorithm to train model.
2) *Test Data:* Test Data is the dataset used to test the trained model which is in the similar form as that of training data.
3) *Preprocessing:* Initializing the data by performing specific tasks and make the data ready in order to process further.
4) *Feature Extractor:* Uses technique(s) to draw the important features possessed by the tweet.
5) *Features:* Features are the important numerical facts & figures drawn by the extractor
6) *Classifier Model:* Classifies the features extracted using classification algorithm.

B. Preprocessing Steps

The aim of the following preprocessing is to create a Bag-of-words data representation. The steps will execute as follows:

1) Cleansing
a) Removing URLs
b) Removing usernames (mentions)
c) Removing tweets with *Not Available* text
d) Removing special characters
e) Removing numbers (digits)
2) Text processing
f) Tokenization
g) Transformation to lowercase
h) Stem
3) Build word list for Bag-of-Words.

C. Data Flow Diagram

This DFD shows what kind of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored. It does not show information about the timing of process or information about whether processes will operate in sequence or in parallel.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

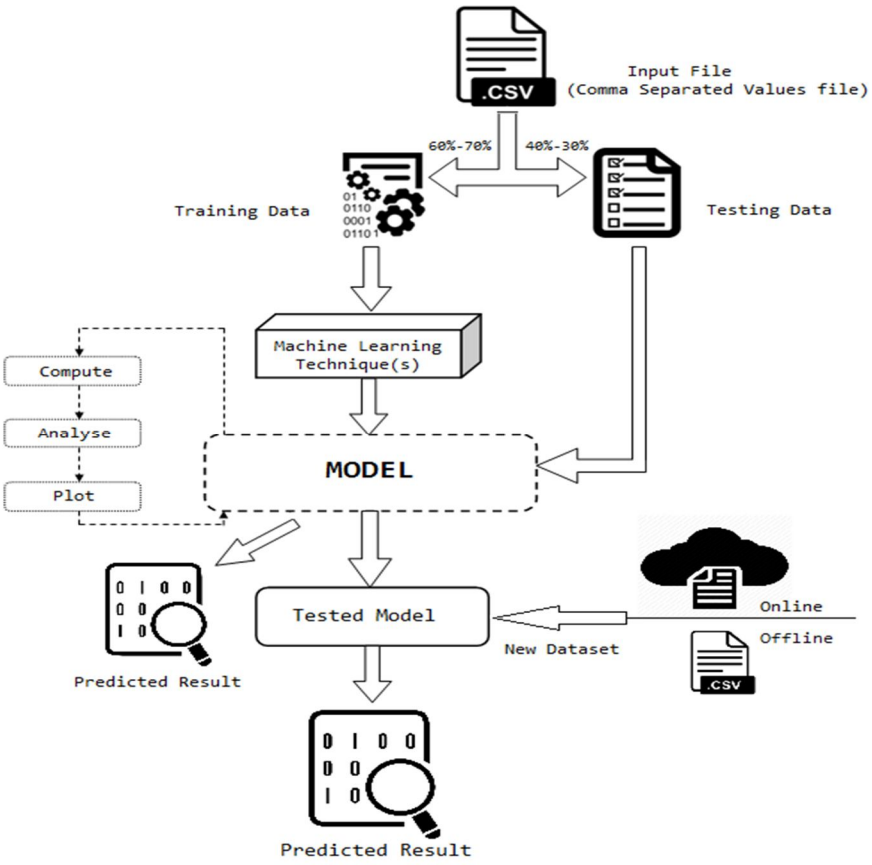

Figure3. Data Flow Diagram

IV. SYSTEM IMPLEMENTATION

A. Dataset

Datasets used for training and testing the example model is named as "Sentiment.csv" has many attributes but we will be focusing on only attributes that are significant for our work, namely "text" and "sentiment". Here "text" is called as the predictor variables and "sentiment" is the response variable. This dataset consists of two variables as stated below: "text" contains the twitter comments that includes text, hashtags, usernames, etc "sentiment": response variable that holds either 'positive', 'negative' or 'neutral' as values defining sentiments. The dataset can be divided to use 60% to 70% of the data for training purpose and the remaining 40% to 30% for testing purpose. In this case, the dataset is divided in such a way that 2/3 of the data is used for training and 1/3 of the data is used for testing. This ratio ensures that double of the testing data is used for training the model.

B. Sentiment Analysis without Emoticon

Procedure for sentiment analysis without emoticon are explained below:

1) *Dataset* is acquired.
2) Each *Tweet* is retrieved one at a time.
3) For each *Tweet* retrieved, perform data *pre-processing* tasks (*Cleansing*, *Tokenizing*, *Stemming*, *Lemmatization*, etc.).
4) *VADER* process is applied on result obtained from step (3), to get the *features* extracted.
5) *Features* extracted, as a result of step (4), are used to prepare *knowledge base (KB).*
6) After Classification, results are obtained.
7) Results of classification obtained in step (6) helps compute overall *Performance Analysis*.

C. Sentiment Analysis with Emoticon

Procedure for sentiment analysis with emoticon are explained below:

1) *Dataset* is acquired.
2) Each *Tweet* is retrieved one at a time.

3) For each *Tweet* retrieved, perform data *pre-processing* tasks (*Cleansing*, *Tokenizing*, *Stemming*, *Lemmatization*, etc.).
4) *VADER* process is applied on result obtained from step (3), to get the *features* extracted.
5) *Features* extracted, as a result of step (4), are used to prepare *knowledge base (KB)*, where emoticon used is assigned an emoticon type (E-type) as shown below.

Emoticons	Examples						E-type
EMOT_SMILEY	:-)	:)	(:	(-:			1
EMOT_LAUGH	:-D	:D	X-D	XD	xD		2
EMOT_NEUTRAL_NOT_SURE	=_=	-_-	:-|	:-\\	:O	:-!	0, 3
EMOT_FROWN	:-(	:(	(:	(-:			-1
EMOT_CRY	:,(	:'(	:"(	:((			-2

Table 1 List of Emoticons

6) After Classification, results are obtained.
7) Results of classification obtained in step (6) helps compute overall Performance Analysis.

V. RESULTS

Implementation of machine learning technique, using Python and machine learning algorithm, was performed after analyzing & observing emoticons that were frequently used in twitter dataset records, where first the tweets were extracted from the dataset available offline/online and later performed pre-processing task, like removing stop words etc., on text. The bag of words is obtained after performing preprocessing. Several analysis were then performed to analyze the effect of emoticons on tweets using machine learning techniques. Finally, the overall performance analysis of the classifier was computed using the confusion matrix.
The Confusion Matrix is obtained after the execution of Performance Analysis code named “Performce_analysis.py”. This result shows the total count of True Positive, True Negative, False Positive and False Negative, as shown in Table 1.

Table 1. Confusion matrix with emoticons

		Predicted				
		Positive	Negative	Neutral	Not Sure	Total
Actual	Positive	39	0	1	0	40
	Negative	0	30	10	0	40
	Neutral	1	0	39	0	40
	Not Sure	0	0	0	40	40
	Total	40	40	40	40	160

The Performance Analysis graph, as shown in Figure 4, is generated which shows the actual and predicted values for 'Positive', 'Negative', 'Neutral' or 'Not Sure' sentiments for analysis “with” emoticons.

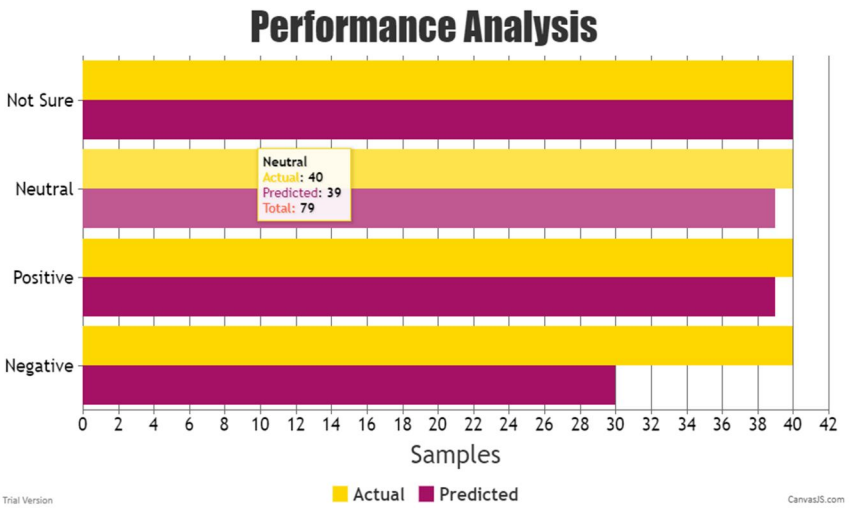

Figure 4. Performance Analysis Bar Graph

The Accuracy Analysis shows the result that analysis carried “with emoticons” are more accurate (i.e. 92.5%) when compared to analysis “without emoticon” (i.e. 79.37%), as shown in Figure 5.

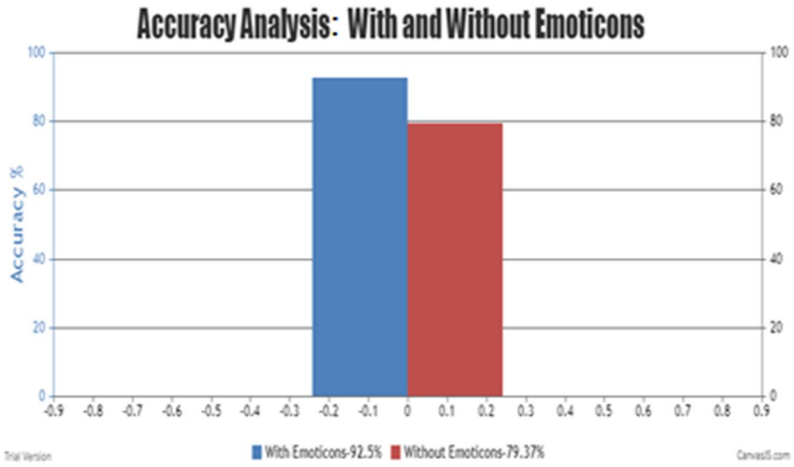

Figure 5. Accuracy Analysis: “With” and “Without” emoticons

VI. CONCLUSIONS AND FUTURE WORK

In this work, it has been shown that emoticons are widely used by Twitter users. The implementation that has been carried out illustrates the fact that the twitter data retrieved offline and online is put under performing pre-processing tasks such as removal of stop words etc. and only the text along with emoticons essential to draw the sentiment is examined further for feature extraction. This includes performing operations like stemming, using Porter-Stemmer, Lemmatizing, and removal of Punctuations. Several analysis were performed using machine learning techniques to analyze the effect of emoticons on tweets. Eventually, Performance Analysis is carried out, which computes overall performance of the testing data, from the datasets available offline, to evaluate their "sentiments", "confusion matrix" and "accuracy" based on 'Positive', 'Negative', 'Neutral' or 'Not Sure' values, which concludes that results obtained in analysis “with emoticons” are more accurate (i.e. 92.5%) when compared to analysis “without emoticon” (i.e. 79.37), as produced in graphical representation, i.e. bar graph, for the comparison and better understanding of the performed analysis and for better visual appearance. Hence, we can say that, in today’s social media trend, the emoticons play significant role in expressing the sentiment of a comment/tweet.

For future work, we plan to annotate updating the Python-Code with more optimized and efficient code. The current Machine Learning Technique can be replaced with more powerful methods/algorithms to compute, analyze and predict the results much faster and accurate with minimum error rate.

REFERENCES

[1] Hao Wang, Jorge A. Castanon Silicon Valley Lab, IBM, USA. "Sentiment Expression via Emoticons on Social Media".
[2] Ms. Payal Yadav, Prof. Dhatri Pandya "SentiReview: Sentiment Analysis based on Text and Emoticons" International Conference on Innovative Mechanisms for Industry Applications. (ICIMIA 2017)
[3] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, "Exploiting Emoticons in Sentiment Analysis".
[4] Waghode Poonam B, Prof. Mayura Kinikar, "Twitter Sentiment Analysis with Emoticons" International Journal of Engineering And Computer Science ISSN: 2319-7242 Volume 4 Issue 4 April 2015, Page No. 11315-11321
[5] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data" International Journal of Innovative Research in Advanced Engineering (IJIRAE). Issue 1, Volume 2 (January 2015)
[6] A. Beryl Joylin, Aswathi T, Nancy Victor, "Sentiment Analysis based on Word-Emoticon cluster" International Journal of Pharmacy and Technology (IJPTFI). ISSN: 0975-766X.
[7] Molly Redmond, Sadegh Salesi and Georgina Cosma, "A Novel Approach Based on an Extended Cuckoo Search Algorithm for the Classification of Tweets which contain Emoticon and Emoji", International Conference on Knowledge Engineering and Applications (ICKEA).
[8] Fred Morstatter, Kai Shu, Suhang Wang, Huan Liu, "Cross-Platform Emoji Interpretation - Analysis, a solution and Applications".
[9] Katarzyna Wegrzyn-Wolska, Lamine Bougueroua, Haichao Yu, Jing Zhong, "Explore the Effects of Emoticons on Twitter Sentiment Analysis".
[10] Harsh Thakkar, Dhiren Patel, "Approaches for Sentiment Analysis on Twitter: A State-of-Art study", Department of Computer Engineering, National Institute of Technology, Surat, India.
[11] Wiesław Wolny, "Sentiment Analysis of Twitter data using emoticons and emoji ideograms", University of Economics w Katowice.
[12] AnalyticsVidhya. (2016). A Complete Tutorial to Learn Data Science with Python from Scratch. Available: https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/
[13] Computer User: Emoticons. (2016). Available: http://www.computeruser.com/resources/dictionary/emoticons.html
[14] EUROPA Data. (2011). EU Open Data Portal, Credit Risk-2011. Available: https://data.europa.eu/euodp/en/data/dataset/data-related-to-credit-risk-2011
[15] Analytics Software & Solutions (SAS). (2018). Machine Learning. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html
[16] Python Software Foundation [US]. (2018). SciPy (ver. 1.1.0). Available: https://pypi.python.org/pypi/scipy
[17] Lena Kallin Westin, "Receiver operating characteristic (ROC) analysis Evaluating discriminance effects among decision support systems", Department of Computing Science, Umeå University, SE-90187 Umeå, Sweden.