***I.J. Information Engineering and Electronic Business*, 2019, 1, 36-49**
Published Online January 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2019.01.05

Business Decision Support System based on Sentiment Analysis

Stephen Opoku Oppong
Faculty of Computing and Information Systems
Ghana Technology University College, Ghana
Email: sopokuoppong@yahoo.com

Dominic Asamoah, Emmanuel Ofori Oppong
Department of Computer Science
Kwame Nkrumah University of Science and Technology, Ghana
Email: dominic_asamoah@yahoo.co.uk; eoforioppong@yahoo.com

Derrick Lamptey
Department of Computer Science
Kwame Nkrumah University of Science and Technology, Ghana
Email: derrick.lamptey@gmail.com

Received: 13 August 2018; Accepted: 18 October 2018; Published: 08 January 2019

***Abstract*—**Since organizational decisions are vital to organizational development, customers' views and feedback are equally important to inform good decisions. Given this relevance, this paper seeks to automate a sentiment analysis system - SentDesk- that can aid tracking sentiments in customers' reviews and feedback. The study was contextualised in some business organisations in Ghana. Three business organizational marketers were made to annotate emotions and as well tag sentiments to each instance in the corpora. Kappa and Krippendoff coefficients were computed to obtain the annotation agreement in the corpora. The SentDesk system was evaluated in the environment alongside comparing the output to that of the average sentiments tagged by the marketers. Also, the SentDesk system was evaluated in the environment by the selected marketers after they had tested the platform. By finding the average kappa value from the corpora (CFR + ISEAR), the average kappa coefficient was found to be 0.40 (40%). The results of evaluating the SentDesk system with humans shows that the system performed as better as humans. The study also revealed that, while annotating emotions and sentiments in the datasets, counsellor's own emotions influences their perception of emotions.

Index Terms—Emotion, Sentiment analysis, Classification, Annotation, Business intelligence

I. INTRODUCTION

Given the fast development of ICT in Ghana, most business organizations are now upgrading to embrace ICT for an improved service delivery. With this, business organizations are realizing the alternate of reaching out to potential customers or their customers in regards to services and products delivery [1]. Usually, this is achieved through the web and other digital media. Although not all business organizations have embraced ICT in Ghana, the practice is stealthily gaining root as many businesses are realising that they can better compete with competitors when business intelligence techniques are well implemented. Lately, it has become a common practice that many people, especially the youth, relies on the web for information regarding business services and products.

Before the widely acceptance and usage of the web by the business organisation for conducting businesses, the traditional practice had always been to take data from customers through questionnaires. Data from these questionnaires are often analysed manually with the intent of extracting opinions and sentiment for business decisions. This is still an ongoing practice of most of the activities of the business organizations in Ghana. The question is, after having to sought data or feedback from customers through the web, how do the business managers analyse the data content for sentiments and opinions? Are they able to analyse the data efficiently, especially with large dataset? This highlights the need for business intelligence (BI) system to complement the work of the business organizations in terms of data mining or analysis.

Aalderks [1] believe that extracting sentiments or opinions from customers' feedback helps in making good decisions. For this reason, it is essential to design and develop an appropriate tool that will automatically detect emotions and sentiments of customers, in order to aid in organizational decisions. Meanwhile, Shivhare & Khethawat [2] believe that detecting an emotional state of a person by analysing a text document written by him or her is quite challenging. Despite the challenges associated with emotion detection in text, Miner [3] believe that a considerable effort has been taken by researchers, in recent years, to reduce the challenges and produce a more accurate prediction of emotions in text. Emotions and sentiments expressed in text can be extracted using a Machine Learning or ontology-based approaches. Machine Learning (ML) approaches have long been an ideal computational technique to efficiently analysed large volumes of data. Several academic domains have used ML algorithms to perform these kinds of task, such as sentiment and emotion analysis. Most notable domains include Natural language processing, affective computing and human language interaction.

While several business organizations manually review and track sentiments from the text-based responses of their customers, this paper presents a digital approach of employing a Natural Language Processing (NLP) using a Na ïve-Bayes supervised machine learning algorithm.

This paper seeks to develop an automatic system (SentDesk) for detection, categorization and visualization of sentiments in text; to investigate the annotation agreement of emotions in text by humans and finally to evaluate the sentiment analysis system with collected humans, thereby ascertaining the efficacy of the system.

II. Related Works

In this section, review of the theoretical framework and related works that have gone into emotion and sentiment analysis are discusses.

A. Emotion and sentiment

Sentiments, also termed opinion, are used interchangeably with emotions, especially in the study of emotions. The term sentiment is referred to as an attitude, thought or judgement that give rise to feelings. Corse-grained analysis of emotions in NLP is termed sentiment analysis. In course-grained text analysis, corpus is classified into continuous emotions categories such as positive and negative. On the other hand, is fine-grained emotions analysis. This is where text content of corpus is classified based on basic emotions, such as Ekman [4] and Plutchik's [5] basic emotions. Solomon [6] further justified why the study of emotions is important for human development. He argued further on the relevance of the study of emotions, where he was emphatic about the fact that emotions as an academic discipline cannot be overlooked. With this, emotions need to be critically studied, thereby unravelling how humans are able to contain and manage their emotions (emotional intelligence). This goes to support the claim by Lerner *et al.* [7] who argued that good decisions on humans are better formed based on how well emotions and sentiments are understood.

Schalke [8] defined emotions as feelings characterized by the state of the mind. Emotions from the perspective of psychological domain are pervasively characterized by *love, hate, anger, trust, joy, panic, fear, and grief* whiles [6], on the other hand, saw emotions as a complicated phenomenon of which he tried to justify by making some categories of emotions. This, [6] recognise that "Emotion is indeed a heterogeneous category that encompasses a wide variety of important psychological phenomena. Some emotions are very specific, insofar as they concern a particular person, object, or situation. Others, such as distress, joy, or depression, are very general. Some emotions are very brief and barely conscious, such as a sudden flush of embarrassment or a burst of anger" [6]

Following the preliminary study from Darwin, Tomkins [9] first proposed eight basic emotions that represent human emotions*: surprise, interest, joy, rage, fear, disgust, shame* and *anguish*. Since then several researchers in the area have propose different basic emotions and the most notable ones are the [4] and [5] basic emotion. These two basic emotions have widely been employed. Kolog *et al* [10] recently employed Plutchik's basic emotions in the development of intelligent system for tracking emotions in text in his PhD dissertation. On the contrary, [2] believe that there are no standard emotion word hierarchy (basic emotions). However, they revealed that the most common emotion hierarchy are the emotions from the cognitive psychology domain, which comprises of the Ekman and Plutchik emotions. Parrot [11], in his book *"*Emotions in

Table 1. Basic emotions and the associated keywords

Happiness	**Sadness**	**Anger**	**Disgust**	**Surprise**	**Fear**
Great	abandon	Irritated	Dislike	Amazement	Nervous
Love	Kill	Hostile	Disgust	Astonishment	Fear
Satisfied	anguish	Aggressive	Hatred	Surprise	Alarm
Glad	Disappointed	Bitter	Distasteful	Dismay	Shock
Cheerful	Abduction	Offensive	Contempt	Suspicious	Horror
Elated	Abuse	infuriate	Appal	Accident	Terror
Jubilant	Abyss	Resentful	Degrade	Accolade	Panic
Merry	Defenceless	Fuming	Abhor	Alert	Distress
Thankful	Deformity	unpleasant	Abominable	camouflage	Mortification
Festive	Sad	Provoked	Infectious	Carnage	Tenseness
overjoyed	lost	aggressive	Infantile	Untimely	Uneasiness

Social Psychology*"* outlined and classified human emotions through an emotion hierarchy in six classes at primary level, which are *Love, Joy, Anger, Sadness, Fear* and *Surprise*.

Table 2 shows the basic emotions and their proponent

Table 2. Discrete emotions and their proponents

Proponents	**Basic emotions**
Plutchik [5]	Anger, disgust, fear, happiness, sadness and surprise, trust, anticipation
Levenson *et al.* [12]	Happiness, Sadness, Anger, Fear, Disgust
Ekman [4]	Anger, disgust, fear, happiness, sadness and surprise
Frijda [13]	Joy, Sadness, Anger, Fear, Surprise Regret, Relief, Hope
Izard [14]	Joy, Sadness, Anger, Fear, Surprise, Disgust Shame, Interest
Johnson-Laird & Oatley [15]	Happiness, Sadness, Anger, Fear, Disgust
Damasio *et al.* [16]	Happiness, Sadness, Anger, Fear
Vytal & Hamann [17]	Happiness, Sadness, Anger, Fear, Disgust
Phan *et al.* [18]	Happiness, Sadness, (Anger), Fear, (Disgust)

B. Sentiments Analysis in Text

In recent times, sentiment analysis from text has seen quite a considerable recognition and usage by co-operate and academic environments. Elder *et al.* [19] describe sentiment analysis as seeking “to determine the general sentiments, opinions and affective states of people reflected in a corpus of large text documents”. Unlike emotions analysis which uses emotional class, sentiments analysis uses classes of positive and negative. Some researchers include neutral as a class in automated sentiment analysis. In basic form, sentiment analyse, particularly in text, is a way of identifying or analysing positive and negative views, opinions, and emotions. Quite a lot of ideas goes into how textual data are being analysed. The most idea with regard to the data suitable to be analysed is the semantic and syntactic features in the text. For instance, *hate* found in a text sentence `I *hate* dogs` expresses some sort of fear in the texts and this triggers negative sentiment towards dogs. Similarly, love in *`I love dogs`* would trigger positive sentence given that love expresses positive sentiment towards dogs.

Though analysis of sentiments in text could be done manually, the growing number of world population and the gradual reluctance of human to analyse large volumes of text has practically made it a difficult task [10]. Computational analysis of sentiment in text do not only optimize time but it also ensures accuracy and consistency. Automatic sentiment analysis is also known as opinion mining which therefore refer to as the use of Natural Language Processing, text analysis and computational linguistics to identify and extract subjective information in source materials. The practice is often use in business intelligence to harness the possibility of making good decisions, thereby maximizing profit and reducing cost.

Several tools have been developed by researchers and other related field expert for commercial purpose and others are freely available for research purposes. Example for these systems are *google analytics, Tweetfeel and Twitrratr.* Examples of engine or framework developed to aid in Analysing sentiments in text for the purpose of research include Weka *(developed by Waikato University)* and *General Architecture of text engineering (GATE).* These systems are often developed to base on the objectives and the purpose it is meant for. For instance, some are developed purposely for Analysing tweets or social network in general and others are also developed to analysis opinions and customer feedback and reviews.

C. Related works in emotions and sentiment analysis

Regarding the annotations of emotions in text documents, [10] undertook a study on determining the influence of counselors’ emotional state on their emotion perception. Though the researchers also delved into the relevance of emotion detection system in counseling, they tasked selected counselors to annotate emotions in students’ life stories in text. They found that counselors agree weakly to each other when annotating emotions in the text corpus. As a result, they reported that counselors own emotional state influences their judgment when they analysis emotions in text. Similar to their work was [20] who rather found strong agreement after untrained participants were selected to annotate emotions in fairy tale corpora. Unlike the work of [10, 20], marketers were selected in this study to perform the emotional and sentimental annotation in reviews and their feedback collected from customers and ISEAR (International survey of Emotional Antecedents and Reaction) corpora. However, the rational for tasking the marketers to annotate the sentiments and emotions in the corpora is to aid in forming a gold standard thereby comparing o the SentDesk platform, which is developed purposely on the grounds of assisting and complementing the work of marketing management in the discharge of business decisions.

Munezero *et al.* [21] undertook a study to investigate the most efficient classifier among Multinomial Naïve Bayes (NBM), Support Vector Machines (SVM) and J48. The study experimented with ISEAR dataset, movie reviews and data from Wikipedia. In the end, the study found SVM and NBM as the most accurate and efficient classifiers given its impressive performance in detection of emotions.

In a related study, [21] designed and implemented an automated system for emotion extraction and visualization. Similar to [20, 21] experimented the system with a collection of students' learning diaries. The researchers employed Plutchik's [5] eight basic emotions viz. *anger, joy, trust, fear, surprise sadness, disgust and anticipation.* However, some of the emotions were combined to give secondary emotions of *anxiety* and *frustration*. In the end, the study revealed an impressive performance of the system. Experimenting with their system, they also reported a high positive polarity of emotions as compared to the negative polarity of emotions in the learning diaries. Altrabsheh *et al*. [22] investigated the learning performance of students based on feedback from the students. In contrast to the work of [21, 22] used real-time interventions of students' feedback rather than aftermath learning diaries of students. In the nutshell, while this study has considered the inter-annotation of emotions in the customers' reviews and the ISEAR corpora by selected Marketers and subsequent classification of the same corpora by using SentDesk system [21, 22] employed machine learning classifiers to extract sentiments in students learning diaries and real time feedback respectively.

Lu *et al.* [23] built a system which uses semantic labelling and a web mining engine technique for emotion detection and classification. While the semantic labelling tool implements semantic role labelling, the web mining engine built in the system, allows search by a specific keyword, and provide answers to lexical questions. The researchers built the system based on a manual procedural approach for creating an emotional model. But in this study, the emotional class was adopted from Plutchik. Recent attention to the development of systems to aid in business and other intelligence have widely adopted either Plutchik's eight basic emotions or Ekman's emotions.

D. Machine learning approaches for text classification

Several approaches of computational analysis of emotions and sentiment have been proposed. The underneath takes a systematic view of some proposed mechanism and algorithms that deals with emotions and sentiment analysis in text. Given the scope of this study, the researcher focused on some basic classifies or algorithms used in both supervised, semi-supervised and unsupervised learning algorithms.

- Supervised learning

Supervised learning is a Machine Learning task for text categorization or classification which infers a function from labelled training data. In supervised learning, input datasets (typically a vector) are labelled and use to train a Machine Learning classifier. The trained classifier or the machine is then used to analyse the test datasets (also called the *supervisory signal*) which is also labelled with salient or interested classes (known as categories). The class can be emotions such as *joy, sadness, and anger* or sentiment such as positive and negative. The algorithm will then determine the class labels for unseen instances. There are several algorithms that have been developed to aid in classifying text in supervised learning. Of the available algorithms and classifiers developed to detect and categorize emotions in supervised text, Naïve Bayes and Support Vector Machines (SVM) are the commonly used due to their simplicity and high efficiency. Particularly, SVM is recognized as the most effective classifier due to its higher efficacy and accuracy. This has been proven by several researchers in the field [21, 22, 24]

1. Naïve-Bayes Classifiers

Naïve-Bayes Classifiers are probabilistic classifiers which are based on Bayes' theorem. The theorem state that "the probability of an event occurrence is equal to the intrinsic probability (calculated from present available data) times the probability that it will happen again in the future (based on knowledge of its occurrence in the past)' [19] Therefore, this theory is used in the development of Naive-Bayes classifier for text classification only when the proportion of each class in a dataset reflects 'prior' probability that any new object will belong in one class or another[19]. Naive Bayes is a conditional probability model which has been adopted extensively in predictive studies. Building a model with Naïve-Bayes from a large dataset or corpus is easy. Regarding this, Ray [25] noted that "Naive Bayes is known to outperform even highly sophisticated classification methods". With this in mind, the researcher opted, in this study, to build the algorithms using Naïve-Bayes based on the intent that as more and more people use the web there will be the likelihood that large dataset would be required for analysis. Mathematically, by given a problem instance in text for classification, denoted by a vector X = (x_1, x_2,…….,x_n) representing some *n* features (independent variables), which assigns to this instance probabilities:

$$p(C_k \mid x_{1,\dots\dots\dots,} x_n) \qquad (1)$$

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c) as shown in the equation (2):

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)} \qquad (2)$$

Where P(*c*|*x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*)., *P*(*c*) is the prior probability of *class*, *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class* and the *P*(*x*) is the prior probability of *predictor*.

2. Support Vector machine (SVM)

SVM is a supervised learning discriminative classifier or algorithm, which is defined by separating hyperplane. What it means is that given labelled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. Support Vector machine (SVM) is often use to analyse data for classification and also for regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. Support vector machine classification algorithms has proven to robust and more efficient algorithm for text classification.

- Unsupervised learning

In an unsupervised learning, which is a machine learning task, infers a function to describe hidden pattern of structure from unlabelled text. Most characteristics of unsupervised learning is that neither test or train dataset is labelled, because of that there is no error or reward signal to evaluate the solution. Unsupervised learning allows one to approach a problem with no idea or whatsoever the solution might me. Several approaches are used in unsupervised learning. There approaches, most notable, are *Hierarchical clustering, k-means clustering and Gaussian mixture models.*

1. Hierarchical clustering: Hierarchical is unsupervised learning algorithm in statistics or data mining. Rokach et al. [26] described it as a method of cluster analysis, which seeks to build a hierarchy of clusters. This is often accomplished as a result of building a multilevel hierarchy of clusters by creating a cluster tree.
2. K-means clustering: K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
3. Gaussian mixture models: Gaussian Mixture Model (GMM) is defined as a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system.

- Semi-supervised Learning

Semi-supervised learning combines supervised and unsupervised learning algorithms. With this machine learning technique, two datasets are involved: test and training set. The training dataset require to train the algorithms (classifier) are labelled with defined themes (supervised) while the test set are unlabelled. Typically, a small amount of labelled data with a large amount of unlabelled data are used. It is mostly recognized as a class of supervised learning. This approach is the most widely used approach especially in the context of business intelligence. This is because it takes a colossal amount of resources for a large dataset to be labelled manually in order. With 10% labelled data from large dataset can be used to predict the unlabelled data. In this study, the platform was developed towards the concept of unsupervised learning. What this implies is that the algorithm used for this approach was trained using collected annotated corpora in line with customers' review and feedback. Meanwhile, proposed input data liable to be analysed by the system are unlabelled. The expectation is that the system would analysed any unlabelled data by categorizing and visualizing the results.

E. Evaluation of text classification algorithm

To establish the efficiency and efficacy of automatic detection of emotion or sentiment analysis system or perhaps a classification algorithm, there is the need to benchmark with human way of analysing emotions. This is usually carried out by tasking trained humans to first annotate emotions or sentiment in a corpus suitable for testing a classifier. Most often, two or three human annotators are chosen for such task. In the end, the level of agreement among the annotators are computed. Kappa, Krippendoff alpha and among many other measures are used in the computation of inter-rater agreement (IRA). And this is the first step in ascertaining the efficacy of any classification system. Aside the kappa measure, Krippendoff also proposes a multiple rater agreement measure [27]. In most cases, to achieve a higher accuracy two or more agreement measures are used for the computation of IRA.

By ascertaining the accuracy of Classification algorithms such as Naïve-Bayes and support vector machines, *F-measure, precision and recall* are the common constructs that are used. In addition to the kappa statistical measure, the section also takes a look at the *F-measure, precision and recall.*

- Kappa statistics

Cohen Kappa was the first to propose officially an Inter-rater agreement measure (IRA) [28]. The IRA measure by kappa was named after him; *Kappa*. Kappa is therefore a measure of chance-corrected agreement between two raters using nominal scale [28]. Cohen kappa was considered as the best method for computing IRA. This is because Cohen kappa computation considers the occurrence by chance. Cohen Kappa (k) for two raters is computed by finding the observed *(A_O)* and expected values *(A_e).* The observed values calculates how the raters agree on annotating or rating individual instances in a text document while the observed *(A_e)* calculates how much the raters are expected to agree based on what each assigns randomly to a sentence or the sort. Generally, based on A_O and A_e, Cohen kappa is calculated in the equation (3) below [20]:

$$k_c = \frac{A_o - A_e}{1 - A_e} \quad (3)$$

Cohen kappa was originally developed for two raters. Researchers, who are interested in this field, took Cohen's kappa further and this led to agreement measure that considers more than two raters. Most noble of these researchers is Light [29]. On the one hand, through the work of [10], unlike Cohen's measure, Fleiss's kappa allows fixed number of raters, but different items could be rated by different raters. However, Fleiss kappa was developed to be used only for nominal or binary scale ratings. On the other hand, Lights extension of Cohen's measure was on a disagreement between raters. [28] explored into the limitations of Light's measure of multiple agreement measure and subsequently published a paper in 1980 to criticized Light's approach on the IRA. [28] pointed out that Light`s kappa measure was unclear for computing more than 3 rater agreement. Nonetheless, [28] agreed that Light measure was most suitable for only 3 rater agreement measure, beyond that brings to the fore ambiguities.

- Precision, Recall and F-measure

Precision is the fraction of retrieved documents that are relevant to a query. It is computed in information retrieval to ascertain the level of accuracy and robustness of classification algorithms. In a plain term, precision considers all retrieved documents. It is however the intersection of relevant and the retrieved document divided by the retrieved document. Beneath is the mathematical representation of precision.

$$Precision = \frac{(relevant\ Document) \cap (retrieved\ Document)}{(retrieved\ Document)} \quad (4)$$

In using precision in classification tasks, *true positives (tp), true negatives (tn), false positives (fp),* and *false negatives (fn)* compares the results of a classification algorithms to that of external measure such as *gold standard* from human. Perruchet and Peereman [30] explained that by using positive and negative it means a classifier's prediction while the true and false refer whether that prediction corresponds to the external judgment. Based on this approach in testing a classifier, [30] defined Precision as:

$$Precision = \frac{(tp)}{(tp)+(fp)} \quad (5)$$

In information retrieval, Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. It is referred to as sensitivity in binary classification. In practice, recall is the number of correct results divided by the number of results that should have been returned. It is often combined with precision for F-measure in text classification. This is because obtaining a recall of 100% is quite trivial and not enough. Hence, the need to measure non-relevant documents also by computing for the precision. It is the intersection of relevant and retrieved documents divided by relevant document as shown in the formula below.

$$Recall = \frac{(retrieved\ Document) \cap (relevant\ Document)}{(relevant\ Document)} \quad (6)$$

In using precision in classification tasks, [30] defined Recall as:

$$Precision = \frac{(tp)}{(tp)+(fn)} \quad (7)$$

F-measure, also known as F-score, is a statistical analysis of binary classification in information retrieval, which is relevant in evaluating classification algorithms. F-measure is the harmonic mean of precision and sensitivity. Perruchet and Peereman [30] explained that "F-measure can be interpreted as a weighted average of the precision and recall, where an F_1score reaches its best value at 1 and worst at 0". Hence, F-measure considers the precision and recall. It also tells how precise and robust a classifier is.

$$F - measure = 2.\frac{Precision\ .\ Recall}{Precision + Recall} \quad (8)$$

$$F - measure = \frac{(tp)}{(2tp)+(fn)+(fp)} \quad (9)$$

III. METHODOLOGY

This section elaborates conceptual framework of the study, the architecture of the SentDesk system and finally evaluation of the SentDesk system. Fig. 1 shows the conceptual framework of the SentDesk system displaying the high-level view of a system.

A. Classification System

Naïve-Bayes algorithm was used as the learning classifier, which calculates the probabilities of each instances of the training data.

As seen in the SentDesk architecture in Fig. 2, the system comprises two classification phases: training and prediction. On the one hand, the datasets are first made to train the SentDesk classifier for a model to be created. The implication is that the multi-class Naïve-Bayes classifier learns from the training data to predict unlabelled or unseen text. The prediction phase, on the other hand, is where the classifier model extracts and classifies the emotions and sentiments from the input text according to defined emotion and sentiment categories.

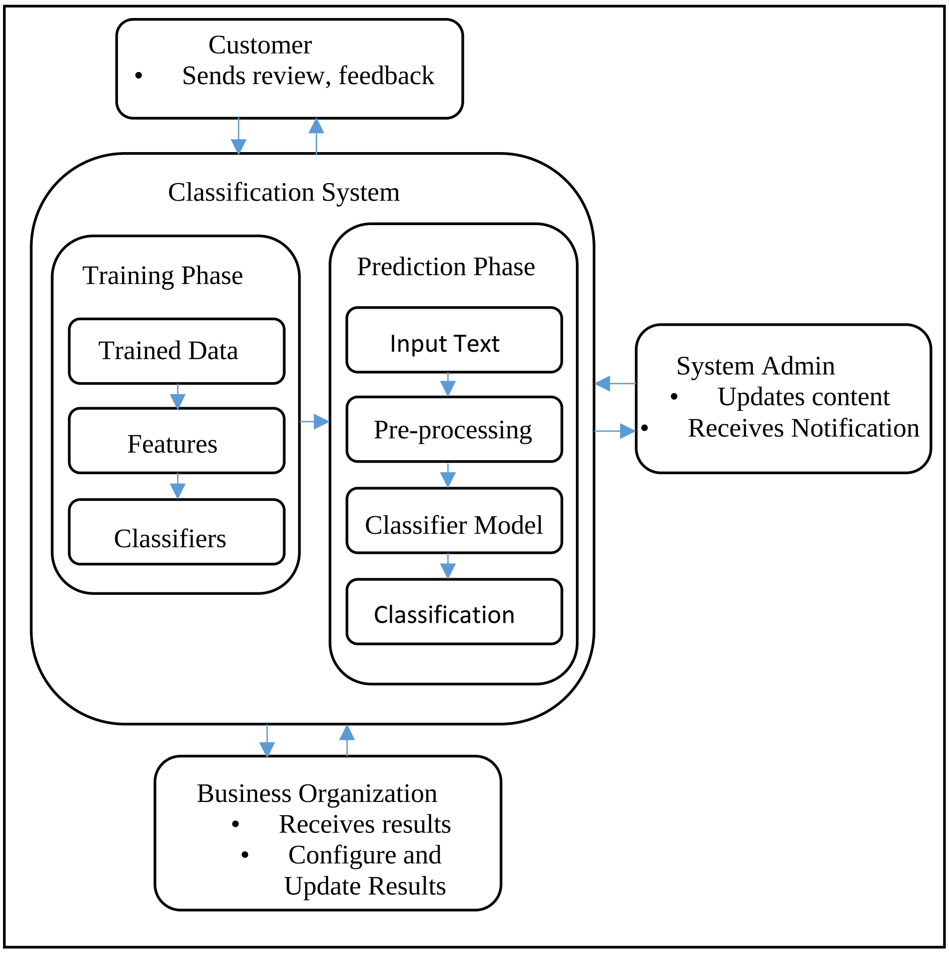

Fig.1.Conceptual Framework

From the architecture in Fig. 2 and the training phase, the system works by first tokenising the training data (life stories) into words. After that, the tokenised words are tagged by their parts of speech, which is accomplished by a POS tagger from the NLTK package. The POS tagging helps to determine the 'stopping words'; they are removed afterwards. To this end, the emotion features are extracted from the text after the removal of the stopping words. At the training phase, the feature words at this point are lemmatised before feeding them into the classifier. Lemmatisation refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. At this stage, the feature words are then fed into the classifier as a training feature set.

After the training, a classifier model is created, which then predicts unseen input text once it is fed into the classifier model. At the prediction phase, just like the training phase, the unseen input text goes through similar pre-processing stages where the unseen text is converted into feature sets. The feature sets are then fed into the classifier model, which generates the predicted labels (thus, emotions and sentiment). Emotional feature words are also spotted and output to the system interface. Fig. 3 shows the use case model of the SentDesk platform.

B. Classifier evaluation

The Na ïve-Bayes classifier, which has been adapted in this study, was evaluated with human way of tracking emotions in text. And this was achieved through inter-annotation exercise.

Human-labelled text is originally referred to as *annotation*. However, advance in technology has made it possible for automatic annotation, where digital tools have been developed for automatic annotation. As a matter of the research objectives, the study while working to determine the efficacy of the classifier based on the polarities (sentiments), this study also aimed to justify why human analysis of emotions in text could be influenced by the own emotions and perhaps other external factors. Quite a recognizable emotional class have been proposed by different researchers, of which Ekman and Plutchik's basic emotions stands out to be the most predominant. While emotions found from ISEAR respondents were used as the baseline emotions (emotion category) for the ISEAR corpus in this study, Plutchik's eight basic emotions were also use as the baseline emotions (emotion category) for the CRF(Customer Review and Feedback) in the annotation. Before the write-up of this paper, this researcher developed the SentDesk system based on the general DSR framework.

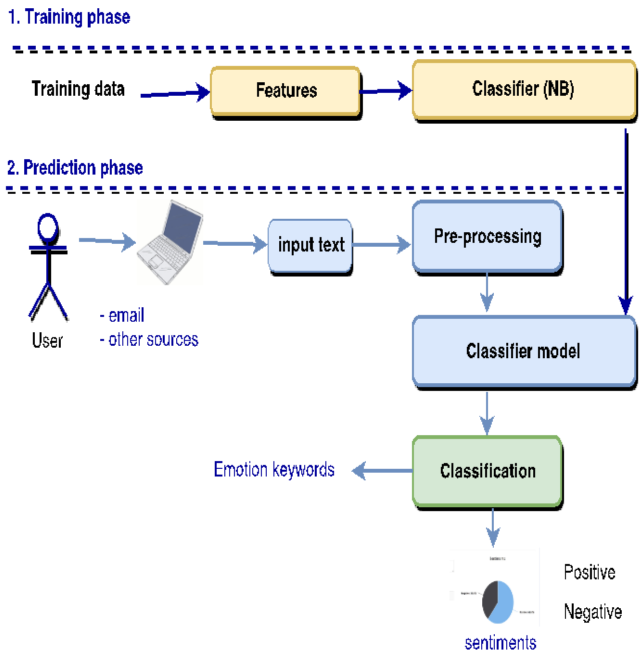

Fig.2. The classification process of the SentDesk

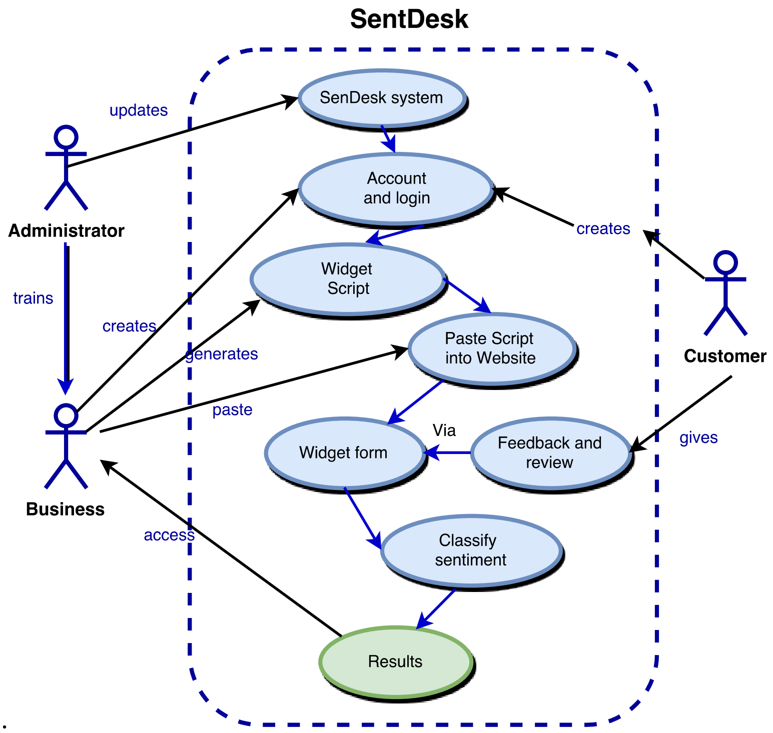

Fig.3. Use Case diagram for SentDesk system

All the three selected marketers, hereafter to be termed M_1, M_2 and M_3, were tasked to annotate and decide on the emotions and sentiments in the collected corpora (CRF and ISEAR). The idea was to compute for the inter-annotator's agreement in the corpora. In this paper, this researcher has used Marketers and annotators interchangeably to refer to the same entity. With the help of a colleague, training was provided to the annotators before starting to annotate the emotions and sentiment in the corpora. During the training, the following were carried out:

1. The objectives of the study were explained clearly to the annotators.
2. The reasons for the annotations to aid in achieving the study's objectives was clarified.
3. The researcher took them through how the annotation should be carried out.
4. Sample data from the corpora was given to the annotators to annotate as a way of ascertaining whether they have understood what is expected of them.

The collected corpora (CRF and ISEAR) were pre-processed into a form suitable for easy annotation by the marketers. In the end, the annotators were tasked to peruse semantically the content of the corpora and annotate the emotions in them. Also, the annotators were made to decide and tag sentiments (negative and positive) to each instances of the corpora based on the emotions expressed in them. In the end, Kappa and Krippendoff coefficients were computed.

Fig. 4 shows the annotation process with regards to the researcher and the marketers' roles.

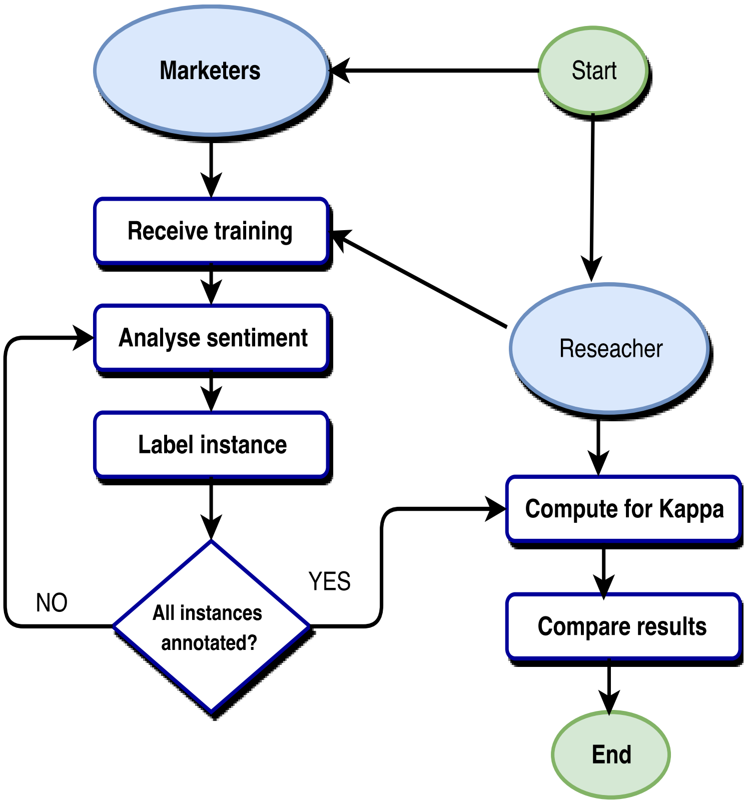

Fig.4. Experimental process for the annotation

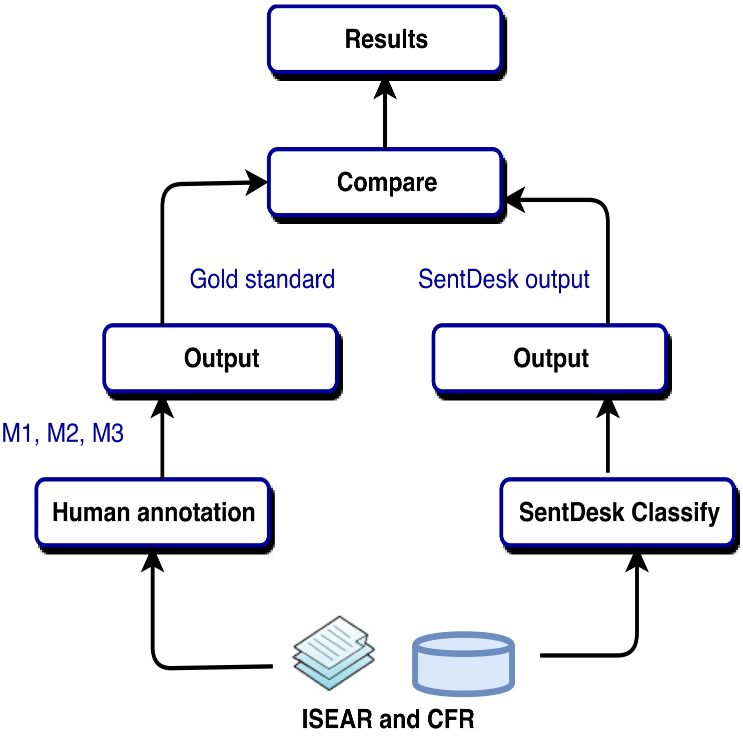

Fig.5. The research and evaluation process

To ascertain the efficacy of the system, the researcher compared the outcome of the SentDesk system, after testing with the corpora, to the outcome from what the marketers labelled. However, the inter-annotators' agreement of emotions in the corpora was meant to investigate the extent of human's level of consistencies and accuracies when tracking emotions in text.

Fig. 5 shows pictorial representation of processes involved in the classifier evaluation. From the Figure, the output from the human labelled is referred to as the gold standard which was made to compare with the SentDesk output.

IV. Results

The findings from the inter-annotation agreement from the annotators is presented in this section. The classifier performance based on the experiment and the contextual evaluation are reported in this section as well.

From the annotated corpora, Table 3 presents the frequencies of the emotions that was identified and labelled by the annotators in the customers' reviews and feedback (CRF), while Table 4 represents the frequencies of the emotions in the ISEAR corpus that was identified and labelled by the annotators. Since the same corpora were given to all the marketers, annotated emotions of each instances of the corpora were compared. Hence, the agreements in column five of Tables 3 and 4 were the agreements extracted based on what each marketer assigned an emotion category to (instance). Figure 6 and 7 represents the marginal distribution of the emotions annotated manually by the marketers. From the figure, it can be deduced that the agreements in both corpora has the least values.

Table 3. Class of emotional occurrence in the annotated CRF corpus

	Customers' Reviews and Feedback (CRF)			
Emotions	**Annotator (M_1)**	**Annotator (M_2)**	**Annotator (M_3)**	**Agreements**
Anger	2	1	0	**0**
Happiness	12	18	10	**8**
Sadness	2	0	0	**0**
Disappointment	4	1	3	**1**
Surprise	3	0	3	**0**
disgust	0	0	0	**0**
Trust	0	1	6	**0**
Fear	0	0	0	**0**
TOTAL	**23**	**20**	**10**	**9**

Table 4. Class of emotional occurrence in the annotated ISEAR corpus

	ISEAR			
	Annotator (M_1)	**Annotator (M_2)**	**Annotator (M_3)**	**Agreement**
Anger	40	44	42	**11**
Sadness	72	56	71	**36**
disgust	21	22	26	**12**
Fear	24	44	40	**12**
Shame	27	28	25	**07**
Guilt	24	34	28	**10**
Joy	63	43	39	**32**
TOTAL	**271**	**271**	**271**	**120**

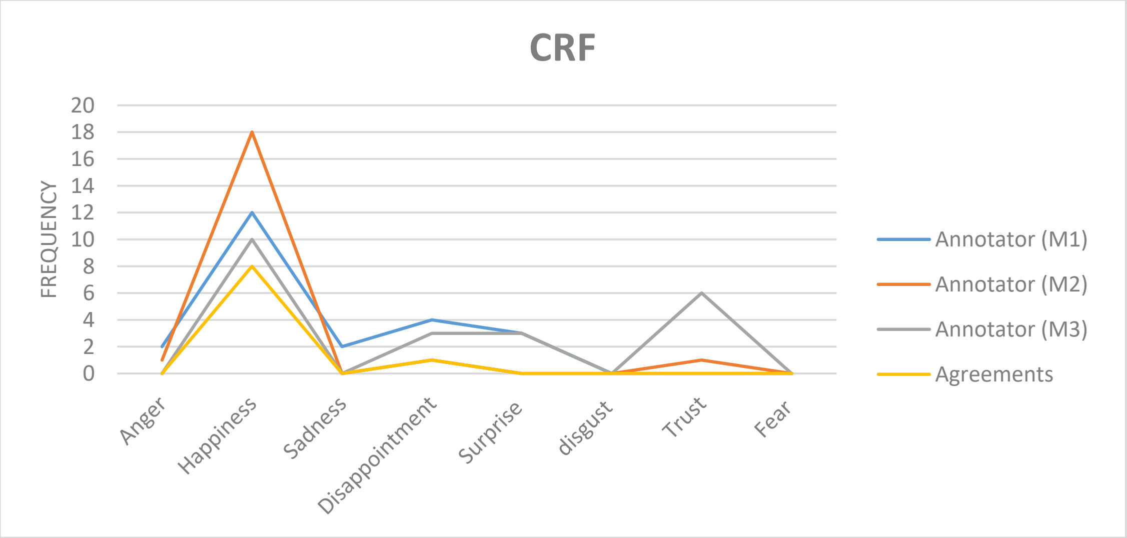

Fig.6 Marginal distribution of the emotions for CRF

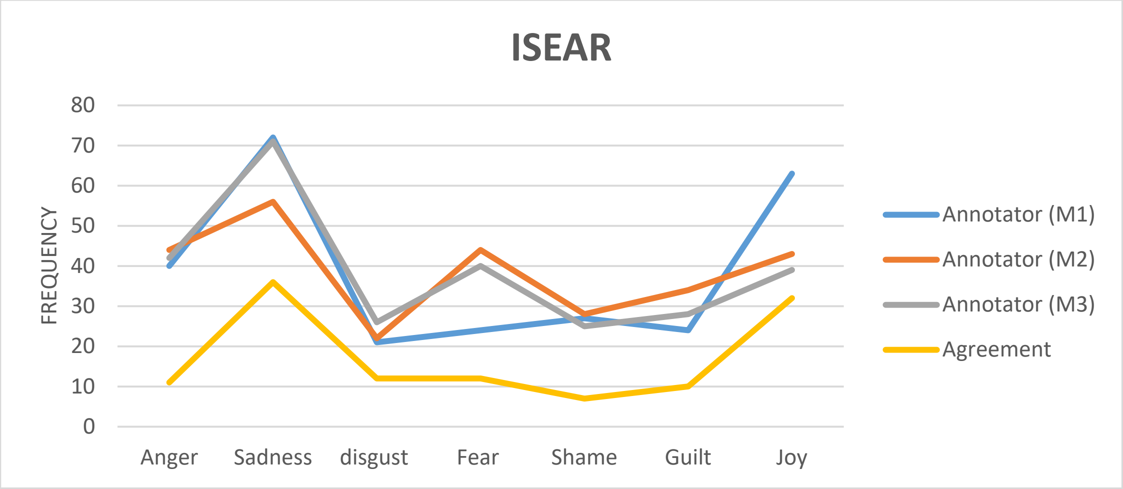

Fig.7. Marginal distribution of the emotions for ISEAR

A. *Findings from the inter-annotation agreement*

After the marketers had annotated the text corpora, this researcher computed for the kappa, Krippendoff and percentage agreements of emotions in the annotated corpora. The coefficients of the kappa values were in line with the ratings of [31] who noted that kappa *value < 0 has no agreement, 0 – 0.20 has weak agreement, 0.21-0.40 has fair agreement, 0.41-0.60 has moderate agreement, 0.61-0.81 has a substantial agreement and 0.81 – 1 has almost perfect agreement*. Hence, after computing for the kappa, the inter-annotation agreement of emotions in the CRF was found to be ***weak*** (Kappa = 0.305, Krippendorff's = 0.315 %Agreement = 30.5%), while in the ISEAR, the inter-annotation agreement of emotions was found to be *moderately agreed* (Kappa = .514, Krippendorff's = .515, % Agreement = 59.3). By finding the average kappa value from the corpora (CFR + ISEAR), a single average kappa value of 0.40 (40%) was realised. This implies that, the marketers, though they had been doing this kind of work manually for years, they *agreed fairly* when analysing emotions in both text corpora. In practice, the marketers, who are humans are not always right when perusing and tracking sentiments, emotions or opinions of their customers in text. This finding highlights the need for computational analysis sentiments/ emotions in text. The proposition is not to replace the work of the humans, in other words the marketers, but to complement their work, thereby ensuring high level of accuracy and consistency. This finding can be supported by [10] who also found *weak agreement* of emotions after giving counsellors opportunity to annotate dataset from students' life stories. Also, [32] had carried out annotation task for several times in order to arrive at an acceptable kappa value of 80%.

Table 5.Intra-annotator's agreement of emotions

Method	**CRF**	**ISEAR**
Fleiss' Kappa	0.305	0.514
Krippendorff's Alpha	0.315	0.515
Percent Agreement	**30.5%**	**59.3%**

Table 6a. Pairwise agreement of emotions among annotators in the data corpus

		CFR	
	M_1	M_2	M_3
M_1	1	0.241	0.404
M_2	0.241	1	0.32
M_3	0.404	0.32	1

Table 6b. Pairwise agreement of emotions among annotators in the data corpus

ISEAR			
	M_1	M_2	M_3
M_1	1	0.549	0.47
M_2	0.549	1	0.528
M_3	0.47	0.528	1

From Table 5, the average sentimental value extracted from the annotated CRF by the marketers yielded 62.5% positive and 34.7% negative, while in the same CRF corpus, SentDesk system yielded 76.9% positive and 23.1% negative sentiments. Also, from the CRF corpus, the sentiments from the marketers yielded 23.24% positive and 76.75.7% negative, while in the ISEAR corpus, the SentDesk system yielded 21.7% positive and 78.3 negative sentiments. Both the marketers and the system yielded higher positive sentiments to that of the negative sentiments in both corpora. However, the variations in the outputs by comparing both the machine (SentDesk) and the humans (marketers) are minimal. The idea is to find out the performance of the SentDesk classifier by comparing the performance with humans. By these results, this researcher believe that the SentDesk system is as capable as humans in tracking sentiments in text. Based on similar studies conducted by [32] and [10] humans are more likely to make changes to what they have already labelled as sentiments to the same corpora if given another opportunity to carry out the same annotation task. However, the SentDesk system was repeatedly used to evaluate the same corpora but produces the same results. From this, this researcher can deduce that machines are more consistent. This however highlight the need of complementing the work of humans with machines but not to replace completely.

Table 7. Comparing the polarities in the corpora with the SentDesk system

	CRF		ISEAR	
	Positive	**Negative**	**Positive**	**Negative**
M_1	65.2%	34.7%	23.24%	76.75%
M_2	78.6%	21.7%	15.86%	84.13%
M_2	82.6%	17.3%	14.39%	85.60%

The range of the best performing classifier is from the coefficient from 0.70 to 1.0.

From Table 7, the weighted average score of the classifier, for both the negative and positive polarities, yielded highly. The weighted recall is .90 (90%), which implies that 90% of the annotated gold standard was actually identified and tracked by the classifier. In the same vein, the weighted average precision score is .88 (88%), which implies that the identified sentiments by the classifier from the gold standard is only 88% was correctly detected and matched with the gold standard annotation. The individual polarities were computed as well.

Table 8. Sentimental output from the manual annotation and from the SentDesk system

	CRF		ISEAR	
	Positive	**Negative**	**Positive**	**Negative**
Inter-annotation agreement	65.3%	34.7%	23.2%	76.8%
SentDesk	76.9%	23.1%	21.7%	78.3%

B. *SentDesk classifier Performance*

The performance of the SentDesk classifier is reported in this section. In this paper, a fine-grained evaluation measures were used to ascertain the performance of the classifier. These measures are the *Precision, recall and f-measure*. *Recall,* also known as sensitivity, measures the fraction of labelled instances of the gold standard that were identified and extracted by the system (i.e., the coverage). *Precision* measures the fraction of the automatically extracted data that was labelled correctly in the gold standard (i.e., the accuracy). The f-measure is the harmonic mean of the recall and the precision.

The weighted f-measure is .80 (80%) and this implies the harmonic mean of the precision and recall. It can be seen that the value of the harmonic mean is beyond 70% which emphasise that the performance of the classifier is impressive. From the table, the best performance of the classifier is the detection of the negative sentiment. In nutshell, the Na ïve-Bayes classifier which was used in this study performed impressively and it is deem suitable a as a tool for analysing sentiments of customers.

Table 9. The evaluation performance of the SentDesk Classifier from combining the CRF and ISEAR corpora as the gold standard

Emotion	Recall	Precision	F-measure
Positive	0.87	0.85	0.89
Negative	0.89	0.92	0.90
Weighted avg.	**0.90 (90%)**	**0.88 (88%)**	**0.80 (80%)**

V. Conclusion And Recommendation

In this study, the researcher collected sample corpora from customers’ reviews and feedback (CRF) from three business organizations. The CRF was complemented with sample corpus from ISEAR. Together they formed the corpora for this paper. By this corpora, three marketers were selected from the three selected business organizations to annotate the emotions in them. Also, they were tasked to tag in each of the instances of the corpora the polarities (negative and positive). After computing for inter-annotation agreement of emotions in the corpora, this researcher found that the marketers (participants) disagreed fairly when the output was compared with the annotated corpora of each marketers (called gold standard). With these results, this researcher has shown that humans (marketers) are likely to disagree when given the opportunity to annotate emotions and sentiments in text. By comparing the output from the SentDesk system to that of the annotated corpora by the marketers (humans), the SentDesk system performed as accurate as the marketers in terms of the sentiment classification in text. The other evaluation was to collect and analysed data from the marketers while they tested the platform in the end. The marketers were enthusiastic about the functionalities of the SentDesk platform and agreed that the system performed impressively, and they believe that the system will serve its purpose for tracking sentiments in text.

Several classification algorithms have been developed, of which a considerable research has been conducted to establish which among the available classification algorithms is efficient. Though a lot of findings have found *support vector machine* and *Na ïve-Bayes* classifiers as the most efficient classifies, there are other emerging classifier that needs to be compared with the existing ones. In this study, Na ïve-Bayes was employed in the development of SentDesk system. However, in the future, the researcher intends to employ support Vector machine in building similar system. The researcher would then compare its output to that of the current system which uses Na ïve-Bayes algorithms.

References

[1] Alderks, D. (2012). Sentiment Analysis: Analysing the polarity of Communication. Retrieved March 3, 2016, from https://web.njit.edu/~da225/media/Assignment7&SentimentAnalysis.pdf

[2] Shivhare, S. N., & Khethawat, S. (2012). Emotion detection from text. arXiv preprint arXiv:1205.4944.

[3] Miner, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

[4] Ekman, P. (1999). Basic emotions. In Handbook of cognition and emotion. pp. 45–60.

[5] Plutchik, R. (1980). Emotion: Theory, research, and experience: Theories of emotion. New York: Academic, 1, 399.

[6] Solomon, R. (2015). Emotions. Retrieved June 23, 2016, from http://global.britannica.com/topic/emotion

[7] Lerner, J., Li, Y., Valdesolo, P., & Kassam, K. S. (2014). Emotion and decision making. Annual Review of Psychology, 66, pp 799-823.

[8] Schalke, M. (2014). Approaches of emotion detection from text. International Journal of Computer Science and Information Technology, 2, pp. 123 - 128.

[9] Tomkins, Silvan S. (1962), Affect Imagery Consciousness: Volume I, The Positive Affects. London: Tavistock.

[10] Kolog, E., Montero, S. C., & Sutinen, E. (2016). Annotation Agreement of Emotions in Text: The Influence of Counselors' Emotional State on their Emotion Perception. In proceeding of of International Conference on Advanced Learning Technologies (ICALT) (pp. 357-359). IEEE.

[11] Parrott, W. G. (2001). Emotions in social psychology: Essential readings.

[12] Levenson, R. W., Ekman, P., Heider, K., & Friesen, W. V. (1992). Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. Journal of Personality and Social Psychology, 62, 972 – 988.

[13] Frijda, N.H. (1987). Comment on Oatley and Johnson-Laird's 'Towards a cognitive theory of emotions'. Cognition and Emotion, I, 51-59.

[14] Izard, C. E. (1994). What develops in emotional development? Intersystems connections. In P. Ekman & J. Davidson (Eds.), The nature of emotions (pp. 356–361). Oxford, UK: Oxford Univ. Press.

[15] Johnson-Laird, P.N. & Oatley, K. (1989). The meaning of emotions: Analysis of a semantic. Cognition and Emotion, 3, 81-123.

[16] Damasio H., Adolphs R, , Tranel D, Cooper G, Damasio AR. (2000), A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. Journal of Neuroscience.;20:2683–2690.

[17] Vytal, K. & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxelbased meta-analysis. Journal of Cognitive Neuroscience, 22, 2864-2885.

[18] Phan K. L., Wager T., Taylor S. F., Liberzon I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage 16331–348. 10.1006/nimg.2002.1087

[19] Elder, J., Miner, M., Delen, D., Fast, A., Hill, T., & Nisbet, B. (2012). Practical text mining and statistical analysis for non-structured text data applications. Waltham, MA: Academic Press.

[20] Volkova, E., Mohler, B. J., Meurers, D., Gerdemann, D., & Buultho, H. H. (2010). Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text., (pp. 98-106).

[21] Munezero, M., Montero, C., Mozgovoy, M., & Sutinen, E. (2013). Exploiting Sentiment Analysis to Track Emotions in Students' Learning Diaries. In Proceedings of Koli calling (pp. 145-152.). Koli: ACM.

[22] Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2004). Learning Sentiment from Students' Feedback for Real-Time Interventions in Classrooms. In proceedings of the third International Conference, ICAIS, (pp. 40-49.). Bournemouth, UK.

[23] Lu, C.-Y., Hong, J.S. & Cruz-Lara, S., 2006. Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques. Third Taiwanese-French Conference on Information Technology – TFIT. Nancy-France.

[24] Lloret, E., Saggion, H., & Palomar, M. (2010). Experiments on Summary-based Opinion Classification. Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, (pp. 107–115). Los Angeles, California.

[25] Ray, S. (2015). "6 Easy Steps to Learn Naive Bayes Algorithm with code in Python. Retrieved June 12, 2016, from http://www.analyticsvidhya.com/blog/2015/09/naivebayesexplained/

[26] Rokach, L., & Oded, M. (2005). Clustering methods. US: Springer

[27] Krippendoff, K. (1967). Content Analysis: An Introduction to its Methodology. Sage Publications.

[28] Conger, A. (1980). Integration and generalization of kappa's for multiple raters. Psychol Bull. , 88, pp. 322–328.

[29] Light, R. (1971). Measures of Response agreement for qualitative data: some generalizations and alternatives. Psychological bulletin, 76(1), pp. 365-377.

[30] Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. J. Neuro-linguistics, 17, pp. 97–119.

[31] Landis, J., & Kouch, G. (1977). The measurement of Observer agreement for categorical data. Biometrics, 33(1), pp. 159–174.

[32] Crowston, K., Liu, X., Allen, E., & Heckman, R. (2010). Machine Learning and rule-based automated coding of qualitative data. In proceedings of 2010 ASIST conference. Pittsburgh, PA, USA.

Authors' Profiles

Stephen Opoku Oppong received his BSc degree in Actuarial Science and MPhil degree in Information Technology from KNUST, Ghana. He is a Lecturer in the Faculty of Computing and Information Systems at Ghana Technology University College, Ghana. His research areas include Mathematical and Statistical Modeling, Algorithms and Machine Learning.

Dominic Asamoah received his BSc MPhil and PHD Degree in Computer Science from Kwame Nkrumah University of Science and Technology (KNUST), Ghana. He is a Lecturer in the Department of Computer Science, KNUST. He has an extensive career of over 15 years teaching experience in Computer Science. Research areas include Image processing, Data Structures and Algorithms and Computer Systems Architecture.

Emmanuel Ofori Oppong received his BSc in Computer Science and MPhil Degree in Industrial Mathematics from Kwame Nkrumah University of Science and Technology (KNUST), Ghana. He is a Lecturer in the Department of Computer Science, KNUST. Research areas include Optimization and Programming.

Derrick Lamptey had his BSc in Computer Science from Presbyterian University College, Ghana and is also a researcher with the Department of Computer Science, KNUST, Ghana. Research areas include Artificial Intelligence and Programming.

How to cite this paper: Stephen Opoku Oppong, Dominic Asamoah, Emmanuel Ofori Oppong, Derrick Lamptey, "Business Decision Support System based on Sentiment Analysis", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.1, pp. 36-49, 2019. DOI: 10.5815/ijieeb.2019.01.05