# Sentimental Analysis on Cognitive Data Using R

**Ramachandra Rao Kurada and Karteeka Pavan Kanadam**

**Abstract** Internet is now vested with new form of societal interactive activities like social media, online portals, feeds, reviews, ratings, posts, critics etc., where people are able to post their expression-of-interest as tweets. Sentiment Analysis (SA) is used for better understanding of such linguistics tweets, extracting features, determine subjectivity and polarity of text located in these tweets. SA inherits text mining approach to process, investigate, and analyze idiosyncratic evidences from text. Now a days, SA was screamed as one of a predictor tool for improvement in knowledge management, revenue generation and decision-making in many businesses firms. The purpose of this work is to leverage a constructive tactic for SA towards dispensation of cognitive information, and seed pragmatic alley to researchers in cognitive science community. This study uses machine learning packages of R language over cognitive data to gain knowledge, discover sentiment polarity and better prediction over the data. To carry out a semantic study over cognitive data we thrived text from numerous numbers of social networking sites. This data was articulated in form of unstructured sentences, words and phrases in a document. Suitable linguistic features are captured to engender dissimilar sentiment polarity and analyze expression-of-interest of user. One of the most prevalent text classification method, Naïve bayes is applied over the text corpus to pinpoint the sentiment and assign its polarity. The connotation in this approaches are evaluated in terms of statistical measures precision, recall, f-measure, and accuracy, thereby these substantial outcomes help to arcade user behavior and predict future trends using SA.

**Keywords** Sentiment analysis · Text mining · Natural language processing Cognitive data · Cognitive science · Machine learning · Artificial intelligence Data mining · Classification · Classification

R.R. Kurada (✉)
Department of Computer Science & Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India
e-mail: ramachandrarao.kurada@gmail.com

K.P. Kanadam
Department of Information Technology, RVR & JC College of Engineering, Guntur, India



R.B. Korrapati et al., *Cognitive Science and Health Bioinformatics*,
SpringerBriefs in Forensic and Medical Bioinformatics,
https://doi.org/10.1007/978-981-10-6653-5_2

## 1 Introduction

Now a day, society is profited with many cutting edge technologies like big data, Internet-of-Things, cloud computing, mobile computing, social networking and semantic web applications. These consequences led to disseminate tremendous cognitive data in the form as the text, images, audio, video etc. at various repositories. This cognitive data stored in such repositories are so huge in volume, and with a variety of attributes, where people are enforced to rely on artificial intelligence tools to process and use or predict for further usage in their perspective business domains. This made machine learning (unsupervised and supervised) methods to combine with computer science, neuroscience and computing techniques to extract knowledge and hidden patterns in cognitive data by realizing in the way a human process previse, thinks and learns.

Buckwalter and Schaffer in 2015 reported cognitive science as a fundamental psychological procedure which influences people's thought with "knowledge", "realization" and "learning" [1]. Later in 2016, Knobere constructed this theory by adding a precise "always" before the phrase "knowledge" and established a feedback mechanism between knowledge and realization and concluded with learning mechanism to quantifying people thinking according to situations [2]. Both these theories intuition people's conceptual knowledge over their study of thought, psychology, linguistics, memory attention, reasoning, artificial intelligence, neuroscience and computer programming.

The cognitive data are available in variety of forms, it needs human intelligence to transform a way to represent, process and examine function of cognition. With this motivation to rehearsal such goals, cognitive data is convoluted with machine learning techniques over the computers to concord with the way human thinks in way of understanding the problem, decision-making and solving problem.

Most of the text data floated over the web via social media or networking sites like Facebook, twitter, LinkedIn are unstructured. Hence it is a complex task to gain deeper understanding of cognitive data and even to analyze. People's knowledge relevant or irrelevant over the subject, and his judgment state count intuition towards SA or opinion mining (OM) [1].

## 2 Motivation Towards Sentiment Analysis

Online marketing is purely dependent on the customers review or rating. Such reviews are accepted as inputs for sentiment analyses. The methods practiced in this analysis reviews the sentiment, analyze and generate the score of a sentiment by Hussein in 2016 [3], Tawunrat and Jeremy in 2015 [4], Matthew et al. in 2015 [5]. Basant et al. [6] in his study expressed SA or OM is wrapped with concepts and techniques from cognitive sciences, artificial intelligence, text mining, natural

language processing and with machine learning primitives like clustering and classification to extract, model, review and use the sentiment. Sentiments are classified into three types (a) structured (b) semi-structured (c) unstructured. Structured reviews which are highly structured and organized in levels as information in a relational database and is dependent on data model. Such reviews have easiness to gain access, store, query, analyze and readily available for prediction. Semi structured reviews are in custom with structured reviews but does not maintain a formal structure or does not fit in any data model associated in the database. In fact they have their own advent to self-describe its own structure. Unstructured reviews are unorganized information which does not fit into any data model or database is arranged in a pre-defined order. This unstructured data cannot be readily classified and used for analysis. Such data are available on internet in the form as text data as tweets, posts, blogs, web pages, PDF files, emails, wikis, documents, video, pictures and graphic images [7, 8]. The biggest challenge of sentiment analysis is to discover and manage knowledge by estimating the sentiment disseminated in unstructured raw data and establish sentiment polarity into various class labels. Sentiment polarity is assessed by the evaluation and detection of sequence of sentiments [9, 10].

Data mining involve four major steps before applying its primitives (clustering, classification, association) over the structured data. The four major steps involved in classical data mining are as follows: (a) Identifying appropriate data (b) cleaning the data (c) selecting relevant features in data for user specific application (d) analyzing distribution of data. Text mining techniques supplements one more step to the existing four steps before applying data mining primitives for knowledge discovery in text related data sets. In prospection of text mining the existing steps of data mining are restructured as follows: (a) Identifying appropriate data (b) cleaning the data (c) Extracting features in data (d) selecting relevant features in data for user specific application (e) analyzing distribution of data. The purpose of adding this additional step "extracting features of data" in text mining is to process unstructured data. Hence this step is used to convert unstructured data to structured data before applying the set of data mining primitives [11].

Viewing these advantages, natural language processing adopted text mining tactics over unstructured data for reviewing the sentiment structure and analysis [12]. Sentiments can be classified in various levels. They are divided into classes with labels as positive, negative and neutral [13]. The major challenge in sentiment analysis is to choose an appropriate algorithm to categorize a sentiment into one the labels positive, negative and neutral with high accuracy.

Hence, in this work we have created a facility for the system to understand the cognitive data by converting the unstructured data available on the web into structured data (data cleaning). This data has been segregated by extracting its appropriate features relevant for sentiment analysis. Thereby to analyze, recognize and propagate the distribution of data, we used supervised and unsupervised techniques. This work exhibits the domain dependence relationship of sentiments with high accuracy results.

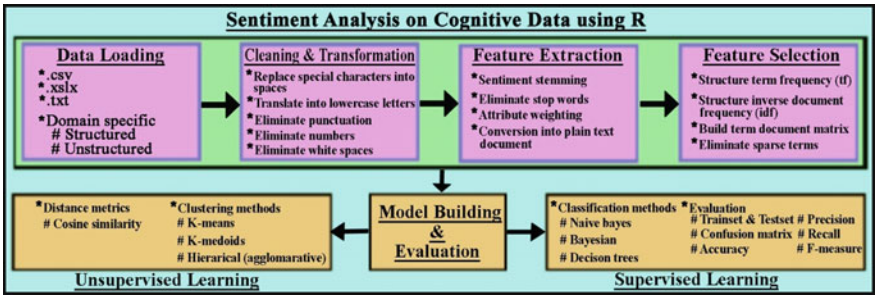


**Fig. 1** Workflow of sentiment analysis on cognitive data

Figure 1 describes the above mentioned stratagem as SA work flow on cognitive data by underlining its changeovers from raw data to enriched data and finally lessening acquaintance towards making strategic business pronouncements.

## 3 Model Formulation and Evaluation

SA uses text mining to study people's expression or emotion in the form as text towards a context, thereby classes its polarity as positive, neutral or negative. Sentiments are extracted from web data as text tweets, and are used for making decision making to determine acceptance or improve quality of the relevant context. Sentiments are posted online in social media, portals etc. as comments, feedback or critique. These tweets are now used as indicators for knowing the pulse of the public. The polarity of tweets are signified in form of positive, negative, neutral or in n-point scale poor, average, good, excellent etc. SA is used to interpret and classify these sentiments into one the categories like positive, negative or neutral. The substantial mechanisms of SA i.e. machine learning and lexicon based techniques are used to uphold such task by estimating or predicting the sentiment. It is an acceptable judgment in SA that if the adapted learning mechanisms derive 70% accuracy over cognitive data, the end outcomes are impressive.

### *3.1 Methodical Approach for Commissioning Sentiment Analysis*

Input: -*Online data with sentiments*

Output:-*Discovery of knowledge, sentiments, polarity and patterns for prediction*

(a) Text transformations and cleaning
*Replace special characters in sentiment with spaces*
*Translate sentiment into lower case letters*
*Eliminate punctuations from sentiment*
*Eliminate numbers from sentiment*
*Eliminate white space from sentiment*
(b) Feature Extraction
*Eliminate stop words from sentiment*
*Stemming the sentiment*
*Transfigure sentiment into plain text document*
(c) Feature Selection
*Structure term frequency (TF), inverse document frequency (IDF)*
*Build term document matrix with weighted (TF-IDF)*
*Eliminate sparse terms with a threshold value*
(d) Model erection and evaluation
*Unsupervised: - compute cosine similarity distance measure*
*Model k-means/k-medoids/hierarchical clustering*
*Supervised: - decompose data into training and testing set*
*Model Naive Bayes/Bayesian classification*
*Evaluation: - Construct confusion matrix*
*Estimate prediction with accuracy, precision, recall and f-measure*

Most of the web data in social networking cites is unstructured, to process SA with supervised learning techniques, this unstructured data have to be converted to structured data. Hence data cleaning, feature extraction techniques are used to identify appropriate attributes for analysis. The expression of SA is done in two levels. (a) Document level: categorize the sentiment as positive or negative, presuming the entire content in the document confining towards one specific topic. (b) Sentence level: The scope of the sentiment is restricted to a single sentence either positive or negative. These procedures are used to deduct using either lexicon-based methods or statistical methods. Statistical methods are automated procedures readily available in the tool and whereas lexicon-based methods needed human interaction. Hybrid approach is now popular by using both lexicon and statistical methods to discover the sentiment polarity. These hybrid methods use supervised and unsupervised learning mechanisms to analyze the sentiment and its polarity. In this work we confined our scope to one of a supervised learning technique naïve bayes classification.

## *3.2 Naïve Bayes*

Naïve Bayes is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. It assumes that the presence or absence of a

particular feature of a class is unrelated to the presence or absence of any other feature. Based on this advantage, this classifier widely used in text mining domain. In text mining this model extract features from bag of words, extract useful features, model posterior probability for a class, based on distribution of words in the document.

## 3.3 Inverted Index

Inverted Index is a data structure central to text corpus. Text is organized in the form of key-value pairs. Key maps the words as tokens, depending on the granularity of the index and a value is in the map as list of postings. This collection of documents in form of text is referred as text corpus [14].

## 3.4 Term Frequency—Inverse Document Frequency (Tf-Idf)

This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance of words increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Tf–idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

## 3.5 Classifier Evaluation Metrics

Classifier evaluation metrics are used to understand and assess how the classification model performs when applied to a dataset [15]. The following statements contain the concise description of classification evaluation metrics adapted in this work. The confusion matrix provides a tabular summary of the actual class labels vs. the predicted ones. Overall classification accuracy is defined as the fraction of instances that are correctly classified. Precision is defined as the fraction of correct predictions for a certain class, whereas recall is the fraction of instances of a class that was correctly predicted and f-measure is defined weighted average of precision and recall.

# 4 Results and Discussion

In this section we practice the systematic methodology described in Sect. 4 to experience the outcomes in an effective way. We used R programming language to tender sentiments from cognitive data sets on a core i3 processor with 4 GB RAM and 64-bit windows 8.0 operating system. The enriched cognitive data sets used in this work are employed from https://www.crowdflower.com/data-for-everyone. To appraise the accomplished consequences of text classifier, we employed the most recurrently used statistical metrics, i.e. precision, recall, f-measure and prediction accuracy. The high values of precision, recall and F-measure regulate the accuracy of the results. The more the value of accuracy the better the results of abstraction.

## *4.1 Dataset 1 (Global Warming)*

This dataset is contributed with three attributes sentiment text, sentiment confidence and sentiment polarity with class labels positive, negative and neutral. The sub-sized tweets credence the extant of global warming or climate change. The transitions in Table 1 denote size of data sets in rows and columns at three stages original, document term matrix and after elimination of sparse terms. Row 2 in Table 1 describes the size of cognitive data after construction of document term matrix of size 6090*12885. This matrix is attained after computation of IF-IDF values. The sparsity of terms in this term document matrix is reduced by eliminating the sparse terms in document vector. Elimination of spare terms are done by setting a threshold value with a numeric between 0 and 1. An important remark here was the value of sparsity is smaller as it approaches i.e. towards bigger zero to small one, hence it eliminates lot of terms occurring 0 times in the text corpus. This was replicated in the third row of Table 1, by exhibiting the size of data set after

**Table 1** Transitions in dataset 1 (global warming)

| Size of dataset 1 (global warming) | Rows | Columns |
|---|---|---|
| Initial | 6090 | 3 |
| Document term matrix | 6090 | 12885 |
| elimination of sparse terms | 6090 | 6 |

**Table 2** Confusion matrix of both training set and test set on dataset 1 (global warming)

| Predictions | Trainset | | | Testset | | |
|---|---|---|---|---|---|---|
| | Negative | Neutral | Positive | Negative | Neutral | Positive |
| Negative | 169 | 11 | 208 | 79 | 3 | 90 |
| Neutral | 347 | 528 | 782 | 134 | 248 | 314 |
| Positive | 787 | 230 | 1201 | 349 | 94 | 516 |

**Table 3** Classifier evaluation metrics on training set and test set of dataset 1 (global warming)

| Sentiment polarity | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| Negative | 0.1642 | 0.4628 | 0.2424 | 0.0497 | 0.148 | 0.4381 | 0.2213 | 0.0465 |
| Neutral | 0.7091 | 0.3294 | 0.4499 | 0.1332 | 0.7476 | 0.3214 | 0.4495 | 0.128 |
| Positive | 0.5260 | 0.5487 | 0.5371 | 0.2678 | 0.5425 | 0.5635 | 0.5528 | 0.2791 |

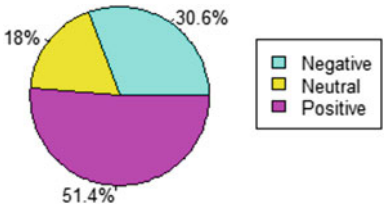


**Fig. 2** Sentiment polarity of dataset 1 (global warning)

elimination of sparse terms with 6090*6. We have partitioned the rows of dataset as trainset and testset with a ratio of 70:30. Table 2 shows the confusion matrix of dataset 1 after using naïve bayes classifer. The samples are classified into 3 class labels "positive", "neutral" and "negative", by establishing a relationship concerning actual and predicted values in form of true positive, false positive, true negative, false negative values. The outcomes of this model is derived form the knowledge supplied to the predictor.

Table 3 includes classifier statistical evaluation metrics as prediction accuracy, precision, recall, f-measure. It is noteworthy from Table 3, that accuracy value of positive sentiment polarity is high in both trainset and testset with a values 0.5371 and 0.5528. The value of accuracy computed by the classifier both at trainset and testset exhibits high values in positive sentiment polarity as 0.2678 and 0.2791. These implications ratify the model was accurate in producing the polarity of sentiment as positive. Further, this accuracy value exhibits credibility by showing more than 50% of people believe in a fact, that the effects of rising temperatures aren't waiting for some far-flung future. Figure 2 denotes trainset sentiment polarity and highlights 51.4% of people in their tweets express their belief in existence of global warming or climate change.

## *4.2 Step Wise Implementation of Dataset 1 (Global Warming) Using R*

```
library(e1071)
library(tm)
setwd(``E:/Sentiment Analysis/SAonCognitiveData'')
tweet_polarity = read.csv(``tweet_global_warming.csv'',
stringsAsFactors=FALSE)
nrow(tweet_polarity)
ncol(tweet_polarity)
gw_dataframe<- DataframeSource(as.data.frame(tweet_polarity[,1]))
gw_Corpus<-Corpus(gw_dataframe)
inspect(gw_Corpus[1:4])
gw_Corpus <- tm_map(gw_Corpus, tolower)
gw_Corpus <- tm_map(gw_Corpus, removePunctuation)
```

```
gw_Corpus <- tm_map(gw_Corpus, removeNumbers)
gw_Corpus <- tm_map(gw_Corpus,stemDocument)
gw_Corpus <- tm_map(gw_Corpus, removeWords, stopwords(``english''))
gw_Corpus <- tm_map(gw_Corpus, stripWhitespace)
gw_Corpus<- tm_map(gw_Corpus, PlainTextDocument)
gw_Tdm<- DocumentTermMatrix(gw_Corpus, control = list
(weighting = weightTfIdf, stopwords = TRUE, minWordLength=3))
dim(gw_Tdm)
gw_Tdm<- removeSparseTerms(gw_Tdm, 0.90)
dim(gw_Tdm)
gw_Tdm_temp <- as.matrix(gw_Tdm)
trainset<- sample(1:nrow(gw_Tdm_temp), trunc(0.7*nrow(gw_Tdm_temp)))
nb_classifier <- naiveBayes(gw_Tdm_temp[trainset, ], as.factor
(tweet_polariy[trainset, 2]))
trainpredict<- predict(nb_classifier,gw_Tdm_temp[trainset, ])
trainset_gw_cm<- table(trainpredict,tweet_polarity[trainset, 2])
trainset_gw_cm_diag=diag(trainset_gw_cm)
num_of_objects=sum(trainset_gw_cm)
trainset_gw_accuracy=trainset_gw_cm_diag/num_of_objects
num_of_classes=nrow(trainset_gw_cm)
num_of_objects_per_class_row_sums = apply(trainset_gw_cm, 1, sum)
num_of_predictions_per_class_col_sums=apply(trainset_gw_cm, 2, sum)
trainset_gw_precision=trainset_gw_cm_diag/num_of_predictions_
per_class_col_sums
trainset_gw_recall=trainset_gw_cm_diag/num_of_objects_per_class_
row_sums
trainset_gw_f_measure=(2*trainset_gw_precision*trainset_gw_recall)/
(trainset_gw_precision+trainset_gw_recall)
data.frame(trainset_gw_precision, trainset_gw_recall, train-
set_gw_f_measure)

sentiments<-diag(trainset_gw_cm)
colors<-c(``cyan'', ``magenta'', ``yellow'')
sentiment_labels<-round(sentiments/sum(sentiments)*100,1)
sentiment_labels<-paste(sentiment_labels, ``%'', sep=``'')
pie(sentiments, col=colors, labels=sentiment_labels, cex=0.8)
legend(1.5, 0.5, c
(``Negative'', ``Neutral'', ``Positive''), cex=0.8, fill=colors)
testpredict = predict(nb_classifier,gw_Tdm_temp[-trainset, ])
testset_gw_cm=table(testpredict,tweet_polarity[-trainset, 2])
testset_gw_cm_diag=diag(testset_gw_cm)
num_of_objects=sum(testset_gw_cm)
testset_gw_accuracy=testset_gw_cm_diag/sum(testset_gw_cm)
num_of_classes=nrow(testset_gw_cm)
num_of_objects_per_class_row_sums = apply(testset_gw_cm, 1, sum)
```

```
num_of_predictions_per_class_col_sums=apply(testset_gw_cm, 2, sum)
testset_gw_precision=testset_gw_cm_diag/num_of_predictions_
per_class_col_sums
testset_gw_recall=testset_gw_cm_diag/num_of_objects_per_class_row_sums
testset_gw_f_measure=(2*testset_gw_precision*testset_gw_recall)/
(testset_gw_precision+testset_gw_recall)
data.frame(testset_gw_precision, testset_gw_recall, testset_
gw_f_measure)

sentiments<-diag(testset_gw_cm)
colors<-c(``red'', ``green'', ``blue'')
sentiment_labels<-round(sentiments/sum(sentiments)*100,1)
sentiment_labels<-paste(sentiment_labels, ``%'', sep=``'')
pie(sentiments, col=colors, labels=sentiment_labels, cex=0.8)
legend(1.5, 0.5, c
(``Negative'', ``Neutral'', ``Positive''), cex=0.8, fill=colors)
```

## 4.3 Dataset 2 (Judge Emotion About Products)

This dataset reviews the sentiments of web users on brands or products like ipad, iphone, android app etc. The tweets are taken into confidence for understanding the trend of sentiment, and there by improve the business. The dataset contain 3 columns as sentiment text to hold the tweet shared by user, product column to describe the product and third column to represent possible category of sentiment confidence. Table 4 shows the row and column sizes of cognitive data during its

**Table 4** Transitions in dataset 2 (judge emotion about products)

| Size of dataset 2 (judge emotion about products) | Rows | Columns |
|---|---|---|
| Initial state | 9093 | 3 |
| Document term matrix | 9093 | 10404 |
| After elimination of sparse terms | 9093 | 10 |

**Table 5** Confusion matrix of trainset and testset of dataset 2 (judge emotion about products)

| Predictions | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Negative | Neutral | No emotion | Positive | Negative | No emotion | Neutral | Positive |
| Neutral | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Negative | 15 | 78 | 218 | 140 | 7 | 31 | 99 | 61 |
| No emotion | 67 | 201 | 2922 | 1367 | 29 | 86 | 1241 | 581 |
| Positive | 27 | 122 | 638 | 570 | 11 | 52 | 271 | 259 |

**Table 6** Classifier evaluation metrics on trainset and testset of dataset 2 (judge emotion about products)

| Sentiment polarity | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| Neutral | 0.0000 | 0.0000 | 0.0000 | 0.1743 | 0.0000 | 0.0000 | 0.0000 | 0.0109 |
| Negative | 0.1743 | 0.1617 | 0.1678 | 0.7723 | 0.1910 | 0.1675 | 0.1785 | 0.4563 |
| No emotion | 0.7723 | 0.6393 | 0.6995 | 0.2915 | 0.7522 | 0.6384 | 0.6907 | 0.0839 |
| Positive | 0.2915 | 0.4352 | 0.3492 | 0.3412 | 0.2605 | 0.3829 | 0.3100 | 0.3310 |

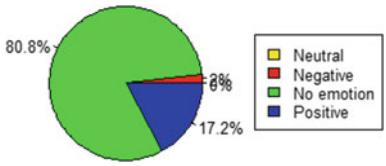


**Fig. 3** Sentiment polarity of dataset 2 (judge emotion about products)

transition. The row 2 of Table 4 exhibits the term document matrix size as 9093*10404. Scarcity on document matrix is applied to eliminate zero terms and to reduce its dimensions, and its values are shown in row 3 of Table 4. Table 5 shows the confusion matrix of both trainset and testset set over the cognitive data when applied over the Naive Bayes classifier. This classifier uses 70% of data in trainset and rest of the 30% of data in testset. The outcomes of this prediction model from Table 5 are analyzed and shown in Table 6.

The sentiment polarity "no emotion" label attains high values of precision with 0.7723, recall with 0.6393 and f-measure with 0.6995. The implication from these tweets reveal that most of the people have no emotions or concerns regarding any products. The values produced by prediction accuracy both in trainset and testset are maintained consistently. Figure 3 shows the spread of sentiment polarity in all four categories of trainset, over dataset 2. Figure 3 indicates 80.8% of people have no emotions towards any product and only 17.2% of people expressed positive tweets about the product.

## *4.4 Dataset 3 (Airline Twitter Sentiment)*

This dataset archives 14640 tweets from 7700 users. The data was scraped with 15 attributes like tweet_id, airline_sentiment, airline_sentiment_confidence,

**Table 7** Transitions in dataset 3 (Airline twitter sentiment)

| Size of dataset 3 (Airline twitter sentiment) | Rows | Columns |
|---|---|---|
| Initial | 14640 | 20 |
| Document term matrix | 14640 | 15095 |
| Elimination of sparse terms | 14640 | 6 |

**Table 8** Confusion matrix of trainset and testset on dataset 3 (Airline twitter sentiment)

| predictions | Trainset | | | Testset | | |
|---|---|---|---|---|---|---|
| | Negative | Neutral | Positive | Negative | Neutral | Positive |
| Negative | 5921 | 1558 | 1137 | 2526 | 646 | 515 |
| Neutral | 287 | 396 | 288 | 119 | 181 | 122 |
| Positive | 238 | 219 | 204 | 87 | 99 | 97 |

negative reason, negativereason_confidence, airline, airline_sentiment_gold, name, negativereason_gold, retweet_count, text, tweet_coord, tweet_created, tweet_location, user_timezone. Contributors or airline travelers in this dataset record their emotions on the topics they tend discuss [16]. We build a model to understand the drift, verify systematic variations, that exists in tweets and satisfaction levels of contributors given in form of sentiment as positive, negative, neutral.

The transitions in Table 7 denotes the size of datasets in form of rows and columns at three stages. Column 3 of Table 7 shows the refined document term matrix after elimination of sparse terms with 14640 rows and 6 columns. The confusion matrix with predictions over dataset 3 is shown in Table 8. The observations related to this dataset are 84.6% of samples in trainset confusion matrix and 92% of samples in testset confusion matrix is polarized with negative-negative polarity. This inference indicate the level of customer satisfaction is very low with the airlines services. Table 9 displays the outcomes of model with assessment metrics prediction accuracy, precision, recall and f-measure. The accuracy metrics renders high values in both trainset and testset by revealing an interesting pattern, that US airline services mentioned in dataset 3 attains low positive polarity and high negative polarity. This was justified in Fig. 4 by reporting about 90% sentiments holding negative divergence. All these implications originate to an inference that, US airlines must focus more on reforms to improve existing services, to capture clients attention and have their gratification.

### *4.5 Dataset 4 (Drug Relation Database)*

This cognitive data is constructed with 2020 records and 16 attributes. The attributes include unit_id, golden, unit_state, trusted_judgments, last_judgment_at, human_relation, human_relation:confidence, human_relation_type, human_relation_type:confidence, documentid, gold, human_relation_gold, human_relation_-gold_reason, human_relation_type_gold, human_relation_type-gold_reason, text. The text tweets supplied to the model is in 16th column, an independent variable, used accept tweets from contributors, the prediction is located in 8th column, used as dependant variable, used to estimate the sentimentality between drug and symptom or disease. In this cognitive data the polarity of the sentiment is classified into 5 class labels as (a) causes side effect, (b) is contraindicated in, (c) is prescribed for a certain disease, (4) others, (5) was effective against a certain disease, and the naïve bayes model classifies the records into one of these classes. Table 10 shows the change overs of size in dataset 4 from actual size to to size of dataset after elimination of sparsity in the document term matrix. The confusion matrix of the model when applied over the dataset 4 is shown in Table 11. It was a mere

**Table 9** Classifier evaluation metrics on trainset and testset of dataset 3 (Airline twitter sentiment)

| Sentiment polarity | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| Negative | 0.9185 | 0.6872 | 0.7862 | 0.5777 | 0.9245 | 0.6851 | 0.7870 | 0.5751 |
| Neutral | 0.1822 | 0.4078 | 0.2519 | 0.0386 | 0.1954 | 0.4289 | 0.2685 | 0.0412 |
| Positive | 0.1252 | 0.3086 | 0.1751 | 0.0199 | 0.1321 | 0.3427 | 0.1907 | 0.0220 |

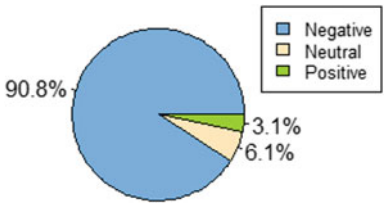


**Fig. 4** Sentiment polarity of dataset3 (airline twitter sentiment)

**Table 10** Transitions in dataset 4 (drug relation database)

| Size of dataset4 (drug relation database) | Rows | Columns |
|---|---|---|
| Initial state | 2020 | 16 |
| Document term matrix | 2020 | 5957 |
| Elimination of sparse terms | 2020 | 1 |

**Table 11** Confusion matrix of both train set and test set on dataset 4 (drug relation database)

| Confusion matrix | Trainset | | | | | Testset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (a) | (b) | (c) | (d) | (e) | (a) | (b) | (c) | (d) | (e) |
| (a) | 915 | 27 | 174 | 159 | 139 | 411 | 10 | 71 | 61 | 53 |
| (b) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (c) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (d) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (e) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

observation that naïve bayes model predicts all the instances of both trainset and testset to only one polarity i.e. actual class label "cause side effects". Table 12 exhibits the model evaluation metrics used on dataset 4. It is shown in Table 12 that the values of prediction accuracy quotes acceptable values in both trainset and testset. This infer there exist an exact likelihood among actual and predicted values of model, since the values are close to high values of 1. The depiction of knowledge from all these implications was the drug specified to cure a disease would cause side effects.

## *4.6 Dataset 5 (Do Chemical Contribute to a Disease)*

This dataset is available with 5713 rows and 21 columns. It includes attributes unit_id, golden, unit_state, trusted_judgments, last_judgment_at, comment_box,

**Table 12** Classifier evaluation metrics on trainset and testset of dataset4 (drug relation database)

| Sentiment polarity | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| (a) | 1 | 0.6690 | 0.8016 | 0.6690 | 1 | 0.6270 | 0.7707 | 0.6270 |
| (b) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (c) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (d) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (e) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 13** Transitions in dataset 5 (do chemical contribute to a disease)

| Size of dataset 5 (do chemical contribute to a disease) | Rows | Columns |
|---|---|---|
| Initial state | 5713 | 21 |
| Document term matrix | 5713 | 12536 |
| After elimination of sparse terms | 5713 | 3 |

**Table 14** Confusion matrix of both trainset and testset on dataset 5 (do chemical contribute to a disease)

| predictions | trainset | | | testset | | |
|---|---|---|---|---|---|---|
| | No-relation | Yes-direct | Yes-indirect | No-relation | Yes-direct | Yes-indirect |
| No-relation | 1400 | 144 | 0 | 584 | 62 | 0 |
| Yes-direct | 67 | 34 | 0 | 29 | 29 | 0 |
| Yes-indirect | 1872 | 475 | 7 | 770 | 236 | 4 |

verify_relationship, verify_relationship:confidence, orig__golden, chemical_id, chemical_name, disease_id, disease_name, form_sentence, original_job_id, pmid, relation_pair_id, sentence_id, uniq_id, verify_relationship_gold, verify_relationship_gold_reason. Users contribute by giving their opinion as tweet text (column 14) in dataset and verdicts their opinion in (column 7) whether both chemical and disease were present or not. They even extend their subscription to determine whether if the chemical directly or indirectly contributed to the cause of disease. Naïve Bayes model is imposed on this cognitive data, to ascertain the nature of dataset. This model makes arrangement for sentences into three categories: no-relation, yes-direct and yes-indirect.

The changeover of dataset 5 from its actual size to document term matrix and after elimination of sparse term is shown in Table 13. The confusion matrix of trainset and testset is shown in Table 14. It was revealed from the confusion matrix that 56.6% of trainset samples and 42.2% of trainset samples are hoarded between the actual (Yes-direct) and predicted (No-relation) sentiment polarities. This unveils a fact that the people's expression of interest was the chances of curing a disease by the chemicals used in the dataset is low. Table 15 shows the predicted values attained after applying the model in form of evaluation metrics. The values of f-measure and accuracy exhibits high values at (No-relation) sentiment polarity in both trainset and testset, showing better results of abstraction than the other labels. Figure 5 reveals a clear extrapolation, that 97.2% of samples in trainset fall under the category of no-relation sentiment polarity, thereby indicates chemicals used in dataset 5 does not contribute to a disease.

**Table 15** Classifier evaluation metrics of dataset 5 (do chemical contribute to a disease)

| Sentiment polarity | Trainset | | | | Testset | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure | Accuracy |
| No-relation | 0.4192 | 0.9067 | 0.5734 | 0.3407 | 0.4222 | 0.9040 | 0.5756 | 0.3500 |
| Yes-direct | 0.0520 | 0.3366 | 0.0901 | 0.0169 | 0.0886 | 0.5000 | 0.1506 | 0.0023 |
| Yes-indirect | 1.0000 | 0.0029 | 0.0059 | 0.0085 | 1.0000 | 0.0039 | 0.0078 | 0.0017 |

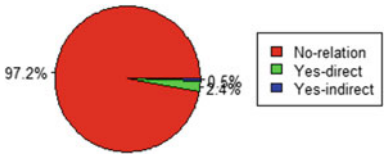


**Fig. 5** Sentiment polarity of dataset 5 (do chemical contribute to a disease)

## 5 Conclusion

In this work, we have applied sentiment analysis over cognitive data reposited in public domains, which are readily accessible via internet. These datasets are available in the form as tweets, posts, reviews, critique, opinions etc. in social media. Understanding the context of these tweets and user preferences has become a challenge in this era. In this work we derived a methodical approach to extract and mine useful knowledge in the form as sentiments to business analytics, in turn helps them to take sensible decisions.

Text preprocessing is commissioned over unstructured data, with a motto to covert it as structured data by using feature extraction and feature selection techniques. As a part of model building over this refined data, one of the most popular supervised learning technique, naive bayes is used to mine knowledge and perform SA on enriched cognitive datasets. Furthermore exploration is made to validate the model by adapting statistical evaluating metrics. The results of these evaluation metrics consistent legalize the accuracy of results by tendering high values across all the cognitive datasets used in this work.

## References

1. Buckwalter W, Schaffer J (2015) Knowledge, stakes, and mistakes. Noûs 49(2):201–234
2. Knobe J (2016) Experimental philosophy is cognitive science. A companion to experimental philosophy
3. Hussein DMEDM (2016) A survey on sentiment analysis challenges. J King Saud Univ—Eng Sci. http://dx.doi.org/10.1016/j.jksues.2016.04.002
4. Tawunrat C, Jeremy E (2015) Chapter information science and applications, simple approaches of sentiment analysis via ensemble learning, volume 339 of the series lecture notes in electrical engineering, DISCIPLINES Computer Science, Engineering SUBDISCIPLINESAI, Information Systems and Applications-Computational Intelligence and Complexity
5. Matthew JK, Spencer G, Andrea Z (2015) Potential applications of sentiment analysis in educational research and practice. In Proceedings of Society for Information Technology & Teacher Education International Conference 2015. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA
6. Basant A, Namita M, Pooja B, Garg S (2015) Sentiment analysis using common-sense and context information. Hindawi Publishing Corporation Computational Intelligence and Neuroscience
7. Landge MA, Rajeswari K (2016) A survey on chemical text mining techniques for identifying relationship network between drug disease genes and molecules. Int J Comp Appl 146 (1):0975–8887

8. Poria S, Cambria E, Howard N, Huang GB, Hussain A (2016) Fusing audio, video and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59. doi:10.1016/j.neucom.2015.01.0950925-231
9. Khairullah K, Baharum B, Aumagzeb K, Ashraf U (2014) Mining opinion components from unstructured reviews: a review. J King Saud Univ Comput Inform Sci 26(3):258–275
10. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
11. Russell S, Norvig P (2003) [1995] Artificial intelligence: a modern approach (2nd ed). Prentice Hall, Upper Saddle River. ISBN:978-0137903955
12. Arjun M, Vivek V, Bing L, Natalie G (2013) What yelp fake review filter might be doing. In: Proceedings of The International AAAI Conference on Weblogs and Social Media (ICWSM-2013), Boston, USA
13. Doaa ME (2016) Enhancement bag-of-words model for solving the challenges of sentiment analysis. Int J Adv Comput Sci Appl 7(1)
14. Ramos J (2003) Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning
15. Bleik S, Gauher S (2016) Computing classification evaluation metrics in R. http://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html
16. Wan Y, Gao Q (2015) Ensemble sentiment classification system of twitter data for airline services analysis. In: IEEE 15th International Conference on Data Mining Workshops, 978-1-4673-8493-3/15. doi:10.1109/ICDMW.2015.7

http://www.springer.com/978-981-10-6652-8

Cognitive Science and Health Bioinformatics
Advances and Applications
Korrapati, R.B.; Divakar, C.H.; Devi, G.L.
2018, IX, 121 p. 59 illus., 30 illus. in color., Softcover
ISBN: 978-981-10-6652-8