Journal of Ambient Intelligence and Humanized Computing (2021) 12:5709–5720
https://doi.org/10.1007/s12652-020-02373-5

ORIGINAL RESEARCH

Automatic classification of emotions in news articles through ensemble decision tree classification techniques

S. Godfrey Winster[1] · M. Naveen Kumar[1]

Received: 28 December 2019 / Accepted: 21 July 2020 / Published online: 3 August 2020

Abstract
Emotions form a major role in human life. As human interactions with online systems have increased drastically, emotion prediction from online text, which otherwise can be monotonous, would help to provide a better environment to the users. Identification of emotions from a normal text itself is very complicated while news text that does not explicitly convey emotions adds more intricacy to it. Data mining methods can be utilized in this context. In this work, the potential of decision tree classifiers in emotion classification is explored. The advocated methodology incorporates two segments towards emotion identification. The first segment deals with data preparation and involves dataset elicitation, translation, HTML tag removal, stop word elimination and stemming. The second segment that implements data mining takes the output of the first segment as its input and applies feature vector formulation, correlation based feature selection, building of bagged Grafted C4.5 learning model and performance evaluation. Based on the evolved classification rules, the emotions are categorized into joy, surprise, fear, sadness, disgust, neutral and mixed kind. Experiments have been conducted to analyse the effect of feature selection methods and ensemble methods in generating efficient rules. The accuracy is compared against eight other decision tree classifiers and also the support vector machine learning model. The proposed methodology achieves the maximum accuracy of 87.83% justifying its utilization in the real time applications.

Keywords Emotion classification · News · Ensemble · Decision tree · Grafted C4.5

1 Introduction

With significant surge in human–machine interactions in the recent times, identification and classification of various kinds of emotions revealed by online texts is gaining high importance (Lee et al. 2012). Emotions can generally be stated as a subjective experience associated with mood, temperament, personality and disposition (Gray et al. 2001) Emotions are revealed by voice, body language and usage of words in the context. Emotions add more information to the message. During a verbal communication, one will be able to understand the emotion even if the language is unknown. But during non-verbal communication, emotion identification becomes a little more complex. Even in that case, the choice of words can depict the emotions. But, this is not the case in News articles, making the scenario highly complicated (Lin et al. 2007, July). In the news articles, journalists refrain themselves to explicitly provide the vocabulary that reveal positive or negative emotions. They express it indirectly through spot lighting a few information while not emphasizing a few others. They provide the opinion and statement of others rather than their own. However identification of emotions in a text becomes highly challenging when the feeling and sentiments are not explicitly and lexically expressed. Another complicating factor about news articles is that it targets a variety of domains and characterizes complicated event descriptions. Hence, automated emotion identification and classification in news text is an area of research that poses many challenges that are yet to be addressed.

Data mining (Han et al. 2011) techniques have proven to yield beneficial outcomes in many applications. These methods extract interesting and non-trivial patterns from a huge deluge of data, thus endowing with rich information from raw data. From the given data, the data mining

✉ M. Naveen Kumar
naveenzion@gmail.com2

S. Godfrey Winster
sgodfreysvee343@gmail.com

[1] Department of Computer Science and Engineering, Saveetha Engineering College, Chennai 602105, Tamil Nadu, India

techniques build a training model, through which the forth coming new instances are handled. Building of learning model can be carried out through supervised and unsupervised learning. Supervised techniques (Sadhana et al. 2017; Kotsiantis et al. 2007) include classification, wherein a training model is built through labeled training samples. Unsupervised techniques (Bandyopadhyay et al. 2012) include clustering, wherein training model is formulated through unlabeled training samples. Supervised classification techniques, though demand for a ground truth to build the training model, it has proved to be highly effective in providing solutions for complex real world scenarios.

In this work, supervised data mining classification techniques have been put to use to identify multi-class emotions from the news articles. This paper focuses on multi-emotion classification through data mining ensemble decision tree classification techniques. The methodology firstly incorporates steps related to data preparation with the view to single out the words that do not reveal emotion based information. After that, ensemble of supervised learning models is built from the available data in order to produce classification rules that identify and classify various kinds of emotions from the news text.

The remaining of the paper is presented as follows: Sect. 2 presents the literature survey related to identification of emotions in textual content with specific attention to news articles. Section 3 explains the methodology advocated towards emotion classification in news articles. Section 4 discusses the various experiments conducted and reports the associated results. Section 5 concludes the paper and provides insights for the future enhancements.

2 Background

Recognition of emotions in news articles is gaining popularity in the recent times. A brief review on the existing works pertaining to prediction of emotions in news articles is provided here.

In 2007, classification of emotions and valence has been carried out in news headlines with the view to analyse the associations between the emotions and its lexical semantics (Strapparava et al. 2007). The dataset is split such that it has 250 annotated headlines as training samples and 1000 annotated headlines as test samples. Six teams namely UPAR7, SWAT, CLaC, CLaC-NB, UA and SICS have taken part in this task for either classification of emotion or classification of valence or classification of both using different techniques. UPAR7 utilizes a rule based linguistic approach to classify the emotions and achieves an accuracy of 93.60%, 95.30%, 87.90%, 82.20%, 89%, 88.60% on annotating anger, disgust, fear, joy, sadness and surprise categories of emotions. SICS adopts a word space model, in which two high dimensional points are created to identify the positive and negative valence, enabling valence annotation. It has been able to achieve only 29% of accuracy. ClaC incorporates knowledge based domain independent unsupervised method where three kinds of knowledge are utilized namely list of sentiment bearing words, list of valence shifters and a set of rules that defines the scope and result of combing the other two lists to annotate the valence and yields an accuracy of 55.10%. ClaC-NB involves supervised Naive Bayes approach for detection of valence but has (Kanagaraj et al. 2020) been able to provide only 31.20% accuracy. UA determines the kind of emotion through statistics collected from various web search engines. The distribution of nouns, verbs, adjectives and adverbs is utilized to predict the emotions. It reports an accuracy of 86.40%, 97.30%, 75.30%, 81.80%, 88.90%, and 84.60% respectively on identifying anger, disgust, fear, joy, sadness and surprise kinds of emotions. SWAT adopts a supervised unigram model to build a learning model to predict the emotional content and yields an accuracy of 92.10%, 97.20%, 84.80%, 80.60%, 87.70% and 89.10% on categorizing the respective emotions.

In 2009, emotion classification from a reader's point of view is analyzed (Jia et al. 2009). Initially, the news headlines and the reader's emotion information are extracted from the web. Various features such as Chinese characters, character unigram, character bigram, word unigram, word bigram, parts of speech, news domain etc., are elicited. Then, Support Vector Machine (SVM) classification is performed to classify the reader's emotion. Experiments demonstrate that certain feature combinations provides higher performance than the others.

Another work in the same year 2009 has been put forth to identify emotions from news sentences from reader's point of view (Bhowmick et al. 2009). Initially words, polarity of subject, verb and object and semantic frame features are extracted from the provided sentences. Then, feature selection is carried out and the combination of polarity and top 80% semantic frame features forms the optimal feature set. The significant features are then provided to multi-label classification through Random K-Label Set (RAKEL). The system achieves an accuracy of 88.20% in classifying 1305 news sentences into four categories namely disgust, fear, happiness and sadness.

In 2010, emotion classification has been attempted to compare it with human classification (Bhowmick et al. 2010). For this experiment, various combinations of different kinds of features are considered. Also, various combinations of the corpus are considered. Best results are accomplished only when the combination of word features and polarity features are taken into account. Again, on exploring the combinations of the corpus, the anger and disgust classes are combined and surprise category is removed. An average precision of 79.50% is obtained through this setup using Alternating Decision Tree (ADTree).

In 2011, emotion identification has been attempted on Czech news headlines (Burget et al. 2011). The methodology involves data pre-processing followed by multi-class classification. SVM classifier has been put to use for this purpose. The system achieves a prediction rate of 71.60% for joy, 81.32% for fear, and 87.30% for anger, 87.30% for disgust, 95.01% for sadness and 75.40% for surprise on 1000 Czech news headlines.

Again in 2013, the dataset published by SemEval 2007 has been evaluated through SVM (Kirange 2013). The data is initially transformed into a matrix and provided as input to SVM. The results achieved are higher when compared to that of the participated teams. The system achieves an average precision of 89.95% with 250 training instances and 1000 testing instances to be categorized into six emotions.

In 2015, sentiment analysis on news comments has been embarked (Mukwazvure and Supreethi 2015). Initially, sentiment lexicons have been utilized to identify the polarity. The polarity can be positive, negative or neutral. The output of the lexicon based method is provided for sentiment classification into SVM and K-Nearest Neighbor (KNN) classifiers. The system achieves the best performance with SVM.

In 2016, event based emotion classification of news articles has been attempted (Li et al. 2016a, b). Initially, the anchor word related to an event is searched and identified through conditional random field algorithm. Then, based on these anchor words, features are derived and provided as input to SVM classification. The another research work discusses about the sentimental analysis of given sentence by extracting the opinion with its verbal context specification in the specific sentence (Sulthana et al. 2018).

Yet another attempt in the same year utilizes the emotional entropy to estimate the weight and handle the noisy training instances (Li et al. 2015). A Reader Perspective Weighted Model is proposed with its roots based on the Naive Bayes theorem. Thus, a mathematical model has been formulated to estimate the conditional probability of the reader's emotion. The model achieves a micro-averaged F1 measure of 36.47% on SemEval dataset and 56.12% on SinaNews dataset.

In 2017, a research has been carried out to analyze the impact of word class and position of news content in identifying the emotion of the news content (Hui et al. 2017). The word class has been divided into noun, verb and adjective. The position is categorized into headlines, first paragraph, last paragraph and various combinations of these. It has been found that polarity based methods achieve better results. The impact of word class and text position have also been revealed which says that adjective expresses the emotion better than verb and noun. Also, in the context of text position, headlines and content of first paragraph express more emotion relevant content than the other positions.

The extensive research that has been carried out in the context of emotion classification in news articles validates the significance of the current work. From the exhaustive earlier studies, it can be realized that exploitation of various classifiers is minimal in the context of news classification (majority of the works adopt only SVM) and Multi-class classification has yielded only lower results when compared to that of the binary classifications. The proposed methodology, involving various data preparation steps and ensemble classification for multi-class categorization of emotions, is explained in the following section.

3 Proposed methodology

The proposed methodology is depicted in Fig. 1. The methodology comprises of two phases namely the data preparation phase and the data mining phase.

The data preparation phase involves dataset collection, conversion of Chinese to English, HTML tag removal, stop word elimination and stemming. The feature vector is formulated through the stemmed words and other common words. The formulated feature vector is provided as input to the data mining phase. The data mining phase incorporates feature selection, ensemble classification, performance evaluation to evolve efficient rules and prediction of emotion categories. The modules in the proposed framework are described below.

3.1 Steps involved in data preparation

Data preparation involves transforming the data appropriately to make the data suitable for processing. In this work, data preparation phase targets at extraction of emotionally significant words and formulation of a feature vector with the extracted words. The steps involved in this context are detailed in the following sub-sections.

3.1.1 Dataset collection

The dataset is composed of 2924 news articles of April 2007 and February 2008 from the society channel of Sina [www.sina.com.cn/society]. The news articles reveal varying kinds of emotions such as positive, negative, neutral and mixed. The positive emotions are those that reveal happiness. It includes Joy and Surprise kind of feeling. The negative emotions trigger a feeling of negativity and melancholy. It comprises of Disgust, Fear and Sadness sorts of emotions. The neutral emotion reveals neither positive nor positive feelings and belongs to a normal stable condition. The mixed emotions signify the state when

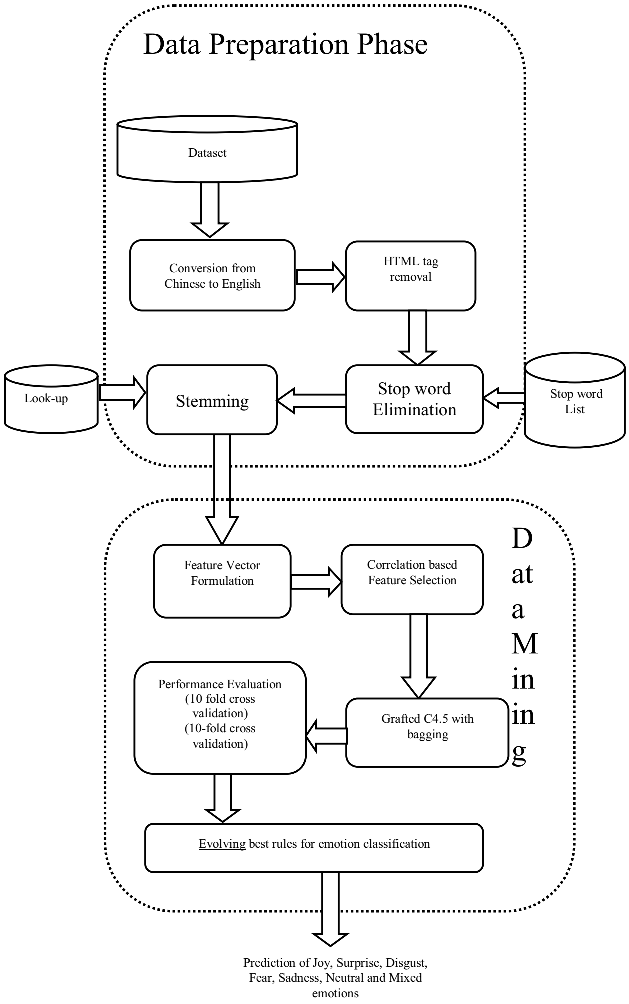

Data Preparation Phase
Dataset
Conversion from Chinese to English
HTML tag removal
Look-up
Stemming
Stop word Elimination
Stop word List
Data Mining
Feature Vector Formulation
Correlation based Feature Selection
Performance Evaluation (10 fold cross validation) (10-fold cross validation)
Grafted C4.5 with bagging
Evolving best rules for emotion classification
Prediction of Joy, Surprise, Disgust, Fear, Sadness, Neutral and Mixed emotions

Fig. 1 Proposed framework for emotion classification in news articles

more than one kind of sentiments pops up. They can either be conflicting emotion of positive and negative or mixture of various positive feelings or mixture of various negative emotions. Illustrations of mixed emotions can be quoted as Joy & Fear, Joy & Sadness, Joy & Sadness & Fear, Sadness & Fear, Surprise & Joy and Surprise & Joy & Sadness & Disgust & Fear. The online users have annotated the news articles based on the emotions that they feel after reading it. They have chosen one of the emotions mentioned above. In this work, seven kinds of emotions are considered. It includes Joy, Surprise, Sadness, Fear, Disgust, Normal and Mixed kind of emotions. The cardinality of instances for each group and the percentage of instances in each category are presented in Table 1.

The dataset is originally in Chinese language. To begin with, the dataset is converted into English, as described in the following sub-section.

Table 1 Details of experimental data distribution

Emotion type	Cardinality	Percentage (%) of instances
Joy	1034	35.48
Surprise	311	10.67
Sadness	231	7.93
Fear	278	9.54
Disgust	202	6.93
Normal	308	10.57
Mixed	550	18.87

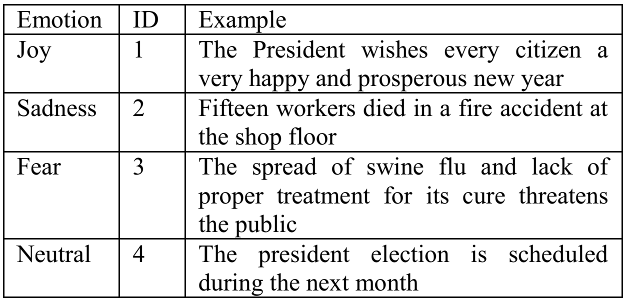
3.1.2 Conversion of Chinese language into English

All languages express emotions effectively. To generalize the methodology, the news articles are translated into the universal language, English. In this research, Google Translator has been used for the purpose of translation. After translating the entire content into English, the unnecessary and insignificant information in the context of emotion identification is removed. Initially, the html tags are removed, as discussed in the subsequent sub-section.

3.1.3 HTML tag removal

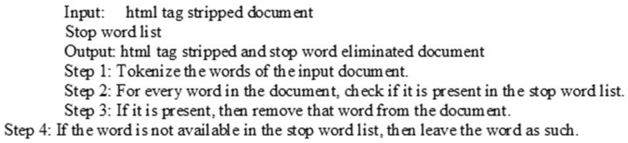
Online content will have html tags embedded to it. They are in no way related to the emotional context of the content. Hence, these tags can be completely removed. HTML parser is utilized for this purpose. The punctuation marks are also eliminated through this process. Four examples are considered to explain each process clearly. The sample sentences are presented in Fig. 2.

Subsequently, the words in the sentences that do not have emotional significance are removed. In this regard, firstly, the stops words are discarded from the sentence. The procedure for this process is explained in the following sub-section.



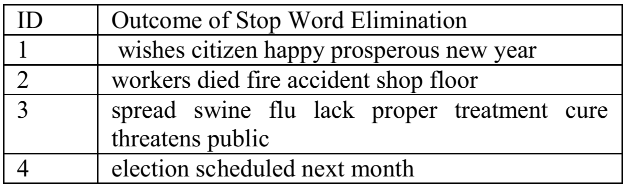
Emotion	ID	Example
Joy	1	The President wishes every citizen a very happy and prosperous new year
Sadness	2	Fifteen workers died in a fire accident at the shop floor
Fear	3	The spread of swine flu and lack of proper treatment for its cure threatens the public
Neutral	4	The president election is scheduled during the next month

Fig. 2 Examples of sentences revealing various emotions



```
        Input:   html tag stripped document
        Stop word list
        Output: html tag stripped and stop word eliminated document
        Step 1: Tokenize the words of the input document.
        Step 2: For every word in the document, check if it is present in the stop word list.
        Step 3: If it is present, then remove that word from the document.
Step 4: If the word is not available in the stop word list, then leave the word as such.
```

Fig. 3 Stop word elimination procedure



ID	Outcome of Stop Word Elimination
1	wishes citizen happy prosperous new year
2	workers died fire accident shop floor
3	spread swine flu lack proper treatment cure threatens public
4	election scheduled next month

Fig. 4 Illustration of stop word elimination

3.1.4 Stop word elimination

Stop words are those words or terms that occur very commonly in all sentences. (Vijayarani et al. 2015). These words have no or very little value in the context of determination of emotions. Common examples of stop words include proper nouns, pronouns, prepositions, conjunctions, articles, etc. These words can be discarded from the sentence to save computational time and resources (Manning et al. 2008). Common strategies to eliminate stop word include (i) obtaining frequency of all words in the documents and eliminating words with very high frequency and (ii) utilizing stop word lists such that if the words in the list and the document match, they are discarded. Many recent researches adopt the stop word lists to eliminate stop words. In this work also, stop word list has been employed to remove the stop words. The list is populated with 22,249 words which include articles, pronouns, proper nouns such as name of the countries, name of the personalities etc. If the sentence contains words that are present in the stop word list, then they are eliminated. The stop word removal algorithm is depicted in Fig. 3.

As per the procedure mentioned in Fig. 3, the stop words are removed from the news articles. An illustration of this procedure on sample sentences revealing various kinds of emotions is provided in Fig. 4.

The outcome of stop word elimination process is provided as input to the subsequent process namely stemming. Details of this process are presented in the next sub-section.

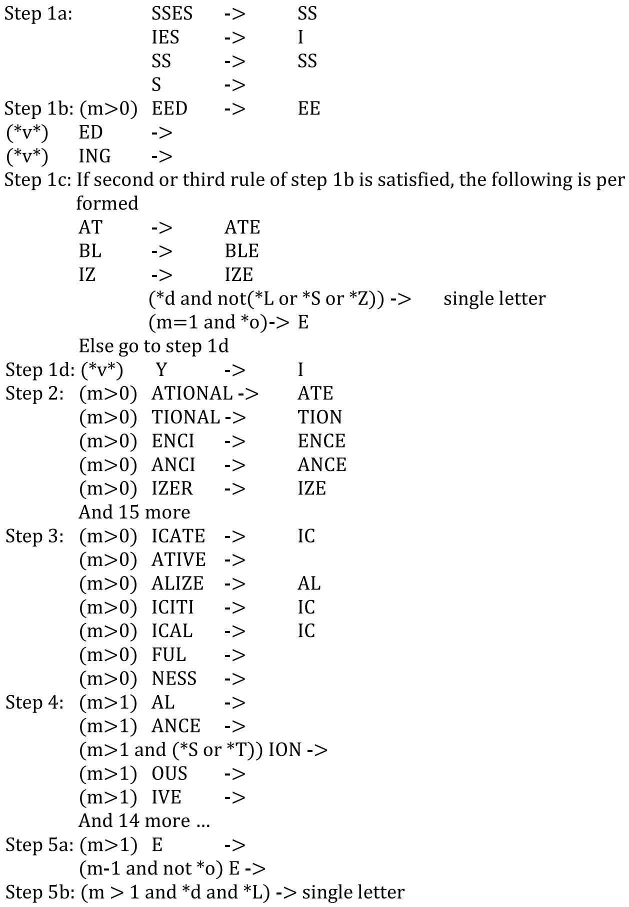
3.1.5 Stemming

In English language, words can be inflected. This means that each word can take many forms to depict tense, number, gender, case, etc. Also, there are families of derivationally related words with similar meaning. In order to facilitate text processing efficiently, these words are converted into their root forms or base forms. In other words, the words are transformed into their stem words in order to improve the efficiency of the further processes. This process can be referred to as stemming or lemmatization. The motive of stemming and lemmatization is to reduce inflectional forms and derivationally related forms of a word to a universal base form. Stemming refers to a crude heuristic process that chops off the ends of words with the expectation of achieving this root word correctly most of the time and often includes the removal of derivational affixes. On the other hand, Lemmatization refers to identifying the root word through the utilization of a vocabulary and morphological analysis of words thereby end up removing inflectional endings only and returning the base or dictionary form of a word.

In this work, the process of stemming is performed. It is carried out through Porter Stemming algorithm (Porter 1980). This algorithm is the most opted one for performing stemming in English language as a part of text processing, information retrieval systems or natural language processing. The algorithm derives the root word through suffix stripping. It considers complex suffixes as a combination of simple suffixes and handles these simple suffixes in a number of steps. In each successive step, the syllable obtained from the previous length is processed based on the length of the current syllable. The algorithm is depicted in Fig. 5.

Some basic notations which are followed by the algorithm are explained here.

i. Consonants refer to those letters other than A, E, I, O, U and Y preceded by a consonant.
ii. Vowels refer to those alphabets that are not regarded as consonants.
iii. A consonant is represented by c and a vowel is represented by v. List of consonants ccc is denoted by C and similarly list of vowels vvv is referred to as V. Hence any word take one of the following four forms namely (a) CVCV… C, (b) CVCV… V, (c) VCVC… C or (d) VCVC… V. The afore-stated four forms are depicted in a single form as [C]VCVC…[V], where the square brackets indicate the arbitrary presence of the content. It can also be denoted as [C] (VC)m[V]. In this representation, (VC)m denotes the repeated occurrence of VC for m times. m is referred to as the measure of the word or word part.
iv. The rules in the algorithm for stripping the suffix will take the form (condition) S1—> S2. This rule signifies that if a word ends with suffix S1 and the syllable before S1 satisfies the condition, then S1 is substituted with S2.
v. The condition part many contain single or multiple expressions with and, or and not.
vi. The implication of various conditions is as follows:



```
Step 1a:          SSES     ->     SS
                  IES      ->     I
                  SS       ->     SS
                  S        ->
Step 1b: (m>0)    EED      ->     EE
(*v*)    ED       ->
(*v*)    ING      ->
Step 1c: If second or third rule of step 1b is satisfied, the following is per
         formed
         AT       ->       ATE
         BL       ->       BLE
         IZ       ->       IZE
                  (*d and not(*L or *S or *Z)) ->     single letter
                  (m=1 and *o)-> E
         Else go to step 1d
Step 1d: (*v*)    Y        ->     I
Step 2:  (m>0)    ATIONAL ->      ATE
         (m>0)    TIONAL ->       TION
         (m>0)    ENCI     ->     ENCE
         (m>0)    ANCI     ->     ANCE
         (m>0)    IZER     ->     IZE
         And 15 more
Step 3:  (m>0)    ICATE    ->     IC
         (m>0)    ATIVE    ->
         (m>0)    ALIZE    ->     AL
         (m>0)    ICITI    ->     IC
         (m>0)    ICAL     ->     IC
         (m>0)    FUL      ->
         (m>0)    NESS     ->
Step 4:  (m>1)    AL       ->
         (m>1)    ANCE     ->
         (m>1 and (*S or *T)) ION ->
         (m>1)    OUS      ->
         (m>1)    IVE      ->
         And 14 more …
Step 5a: (m>1)    E        ->
         (m-1 and not *o) E ->
Step 5b: (m > 1 and *d and *L) -> single letter
```

Fig. 5 Porter stemming algorithm

a. (m = 0), (m > n), (m < n) are conditions based on the measure of word part.
b. *S refers to the condition that the stem terminates with S or similarly other alphabets.
c. *v* denotes the condition that the stem comprises of a vowel
d. *d denotes the condition that the stem should end with double consonant
e. *o represents the state where the stems terminate with cvc and the second c is not W, X or Y.
f. In each phase, with a set of rules written beneath each other, only one is obeyed and this is the rule the longest matching S1 for the given word.

The algorithm is formulated with five primary phases with each phase having sub-phases. These are clearly indicated in Fig. 5.

The rules formulated in Fig. 5 evidently present the stemming process. These rules are coded to execute the process of stemming. The rules in step 1 mainly focus on dealing with plurals and past participles. The subsequent steps take the outcome of the previous step for processing. In step 4, almost all the suffices are stripped away and step 5 deals with the cleaning up if required. An illustration of the outcome of the algorithm is presented for sample sentences in Fig. 6.

The outcome of the stemming process is the root form of the words that help in identifying the emotion of the article. It can be noticed that a few common words representing emotions effectively have not been stemmed to their correct forms. Hence, a look-up has been formulated such that around 13,500 words have been selected and their various forms are populated. The output of the stemming algorithm is provided to this look-up and if the words match any of the words in the look up, then, the root word is chosen appropriately. The output after traversing the look-up is provided in Fig. 7.

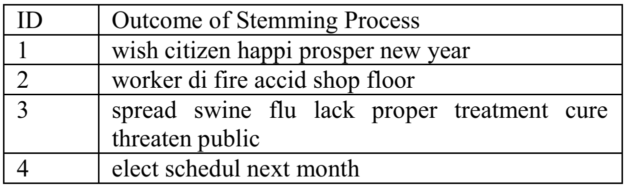
This forms the output of the data preparation phase. With these words, the algorithm enters into the data-mining phase, which is explained in the following section.

3.2 Data mining phase

Data Mining is the process of discovering non-trivial patterns from the huge deluge of data. In this work, identification of emotions forms the non-trivial patterns. The data mining phase involves feature vector formulation, feature selection, evolving of ensemble classification rules and performance evaluation. Discussed in steps below.

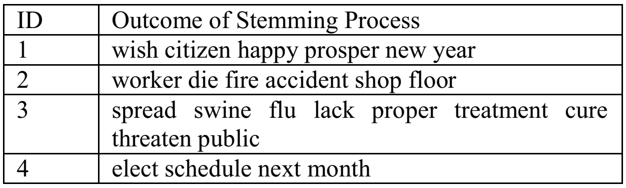
3.2.1 Feature vector formulation

The feature vector formulation plays a very significant role in the performance of the forthcoming classification process. The expressiveness of the features aid in better classification of the instances. In this work, the feature vector takes a binary representation. In this representation, a 1 represents the presence of a particular word while a 0 denotes the



ID	Outcome of Stemming Process
1	wish citizen happi prosper new year
2	worker di fire accid shop floor
3	spread swine flu lack proper treatment cure threaten public
4	elect schedul next month

Fig. 6 Result of stemming process



ID	Outcome of Stemming Process
1	wish citizen happy prosper new year
2	worker die fire accident shop floor
3	spread swine flu lack proper treatment cure threaten public
4	elect schedule next month

Fig. 7 Outcome after performing the look up

absence of the word. WordNet-Affect (Strapparava and Valitutti 2004) has been utilised to formulate the words in the feature vector. It is a collection of a list of words belonging to various categories such as emotion, cognitive state, trait, behavior, attitude and feelings. In this work, the words from the category of emotions are taken into account. The words specifically reveal emotions such as Joy, Surprise, Anger, Fear, Disgust and Sadness. Thus, a total of 1536 words have been elicited from Wordnet-Affect belonging to the emotion category with varying cardinality of words under each kind of emotion. After that, the feature vector is augmented with the root words that are meaningful English words obtained as an outcome of the stemming and look-up process on, provided they are not a part of the feature vector list already. IN this work, 5006 such words are identified and appended to the feature vector. This leads to a high dimensional feature vector characterizing 6542 dimensions.

Having constructed the feature vector, the input to the modules in data mining segment can be visualized as each row indicating a news article and each column denoting a word that expresses the emotion of the article significantly. Illustration of the input table is provided in Fig. 8.

As the dimensionality of the feature vector is very high, it increases the computational complexity of the problem. Therefore, only significant features are retained. This is done through a feature selection process, explained subsequently.

3.2.2 Feature selection

Feature selection (Kim et al. 2003) forms a key process in data mining with respect to many applications. It involves choosing of significant features from the entire set of extracted features. This in turn results in reduced computational complexity as well as improved performance in terms of accuracy (in most cases) as the irrelevant features which drop the performance levels are eliminated.

Many feature selection techniques have been proposed in the literature. For this work, Correlation based Feature Selection (CFS) (Hall et al. 2000) performs well in accomplishing enhanced performance. It is rooted on the notion that good feature sets comprise of features that are highly correlated with the class, yet uncorrelated with each other. The methodology of CFS for selecting a subset of features from the entire feature set is explained here. The subset of features is initially set to be empty. The features are then augmented one by one through the best-first approach. In this approach, the features are added by means of forward search. The search is terminated when five consecutive fully expanded subsets do not improve the performance over the current subset. The performance of each subset is evaluated in the terms of a merit metric which is based on feature-class correlation and feature-feature correlation. The correlation measure is calculated through symmetrical uncertainty technique as given in Eq. 1.

accident	cure	…	die	Elect	…	happy	Proper	Class
0	0	…	0	0	…	1	1	Joy
1	0	…	1	0	…	0	0	Sadness
0	1	…	0	0	…	0	0	Fear
0	0	…	0	1	…	0	0	Neutral

Fig. 8 Illustration of the feature vector

$$\text{Symmetrical uncertainty coefficient } (SUC_{XY}) = 2 * \frac{gain}{H(Y) + H(X)} \tag{1}$$

where,

$$gain = H(Y) - H(Y|X)$$

$$H(Y) - H(Y|X)$$

$$H(Y|X) = \sum_{y \in Y} p(y) \log(p(y))$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y) \log(p(y))$$

X and Y are features of a subset.

With the symmetric uncertainty measure, the merit metric is calculated as shown in Eq. 2.

$$\text{merit} = \frac{\mathrm{k} * \mathrm{ASUC_{cf}}}{\sqrt{\mathrm{k} + \mathrm{k}(\mathrm{k} - 1) * \mathrm{ASUC_{ff}}}} \tag{2}$$

In Eq. 2, k denotes the number of features in the subset $\mathrm{ASUC_{cf}}$ represents the average feature-class correlation calculated through symmetric uncertainty measure and $\mathrm{ASUC_{ff}}$ denotes the average feature-feature correlation.

When CFS is applied to the formulated feature vector with 6542 dimensions, a compact set of features is yielded. These significant features are then fed as input to the classification process. The classification process is explained in the following section.

3.2.3 Classification

The feature vector comprising of important features obtained from feature selection is given for classification. Decision trees have demonstrated good performance in many applications for solving real world problems. In addition to this, ensemble of classification trees (Dietterich et al. 2002) has exhibited higher performance when compared to that of the individual trees in majority of the applications. In this work, many bagged decision trees have been employed while the grafted C4.5 has yielded highest prediction accuracy. Grafting is the process of adding additional cuts to the leaf regions if it proves to improve the accuracy of the classification.

```
Bagging(D,A,numModels)
Step 1: For I = 1 to numModels do
Step 1a: Create Bootstrap samples $D_i$ by sampling D with replacement.
Step 1b: Call C4.5($D_i$, A)
Step 2: End
Step 3: A label of a new instance is predicted by each tree as follows
        For i = 1 to nt
           For j = 1 to num_classes
                   weighted_vote(j) = weightage(i) * prediction(i)
           End for
           maxprob = max(weightedvote)
           maxindex = find(weightedvote == maxprob)
           classprediction(i) = maxindex
       End for
Step 4: The class that has the maximum number of entries in classprediction
        is chosen as the final prediction from the forest.
```

Fig. 9 Bagging procedure

```
C4.5(D,A)
Step 1: Create a node N.
Step 2: If all instances in D belong to the same class C, then
                        Return N as the leaf node labeled with class C.
Step 3: If A is empty, then
                        Return N as a leaf node labeled with the majority class in D.
Step 4: For all attributes a in A, compute gain ratio as follows:
                        $GainRatio(a) = \frac{Gain\ (a)}{SplitInfo\ \ (a)}$
                        Where $splitinfo_a(D) = -\ {}_{j=1}^{v} \frac{|D_j|}{|D|} * log_2(\frac{|D_j|}{|D|})$
Step 5:Assign $a_{best}$ = attribute with maximum gain ratio
Step 6: Label node N with $a_{best}$ and let it test the splitting criterion.
Step 7: For each outcome j of the splitting criterion,
             $D_j$ = data instances in D satisfying outcome j.
Step 8: If $D_j$ is empty, then
                        Attach a leaf labeled with the majority class in D to node N.
Step 9:   Else, attach the node returned by recursively calling C4.5(D,A).
Step 10: call grafting(N,D,A)
Step 11: Return N
```

Fig. 10 C4.5 algorithm for tree construction

The Grafted C4.5 decision tree is implemented using the concept of All-Test-But-One-Partition (ATBOP). The ATBOP for a leaf refers to that region formed by eliminating all and only decision surfaces that enfold the leaf. The advantage of this procedure includes less computational requirements, insensitiveness towards the order in which decision boundaries are formed by the original C4.5 algorithm and reduced number of alternative cuts thereby reducing the probability of over fitting.

In order to execute the bagged grafted C4.5 decision tree, initially, the original C4.5 decision tree is executed followed by which the ATBOP grafting is performed as a post-processing step. Once a grafted C4.5 decision tree is formed, many such similar trees are formulated with different sets of training data and the majority of their predictions are considered to arrive at the final prediction. The procedure of bagged grafted C4.5 in the view of deriving classification rules with reduced error rates is presented in Figs. 8, 9 and 10. The input to the algorithm includes the dataset D that contain N instances with the set of attributes A along with their class label. The ensemble procedure that implements the bagging concept is presented in Fig. 8.

Figure 8 depicts the bagging procedure, where many decision trees are constructed from the bootstrapped samples of the training instances. Once the required number of decision trees are built, then the forest of trees are utilized for making the prediction. Every single tree provides its prediction and the final prediction is the class which has the majority number of votes. Every single tree is constructed through the C4.5 algorithm along with All-Test-But-One-Partition grafting as post-processing as shown in Figs. 9 and 10. The conventional C4.5 algorithm is presented in Fig. 9 along with a call to the grafting procedure.

C4.5 algorithm is one of the most widely used decision tree algorithm that has proved its potential in many applications. This algorithm uses gain ratio evaluation metric to assess the split attributes. The attributes with the highest gain ratio is assigned as the split attribute, the tree construction algorithm is a recursive process that iterates again and again building sub-trees until it arrives at leaves in which a leaf contains only one class or majority of the instances belong to the same class. Once the tree is constructed, it is subjected to grafting in order to increase its performance. The grafting procedure is presented in Fig. 10.

In the procedure mentioned in Fig. 10, cases(n) is the set of all training instances that can reach node n. value (a,x) is the value of attribute a belonging to the set of attributes A for a training instance x belonging to D. Pos(D,c) refers to the number of objects belonging to class c in the set of training examples D. Class (x) depicts the class label of the training instance x. Laplace of (D,c) can be computed as $\frac{\text{pos(D,c)+1}}{\text{|D|+2}}$. upperlim(n,a) signifies the minimum value of a cut c on attribute a for an ancestor node x of n with respect to which n lies below the a ≤ c branch of x. If there is no such cut, upperlim(n,a) = ∞. This determines an upper bound on the values for that may reach n. Similarly, lowerlim (n,a) depicts the maximum value of a cut c on attribute a for an ancestor node x of n with respect to which n lies below the a > c branch of x. If there is no such cut, lowerlim(n,a) = − ∞. This determines a lower bound on the values for a that may reach n. prob(x,n,p) be the probability of obtaining x or more positive objects in a random selection of n objects if the probability of selecting a positive object is p.

The grafting procedure investigates the leaves and places a cut on it only if the cut can improve the differentiability and provide enhanced performance. As the ATBOP considers only a subset of training instances that cover the leaf, the computational complexity is less. Though there are many other variants of grafting procedure, the ATBOP procedure is justified for its better performance and has been used in this work. The rules obtained be the classification algorithm is evaluated through performance evaluation techniques, mentioned in the subsequent sub-section.

3.2.4 Performance evaluation

The potential of the evolved rules is assessed through performance evaluation techniques. Broadly used techniques include train test method and k fold cross validation method. This work adopts tenfold cross validation for evaluating the performance. In tenfold cross validation, the dataset is divided into 10 subsets. During the first iteration, the first nine subsets are used for training while the tenth subset is utilised for testing. During the second iteration, the subsets 2 to 10 are utilised for training while the first subset is dedicated to testing. The process continues for 10 iterations such that every subset acts both as the training subset and the testing subset. The quantitative metrics used for evaluation include accuracy, presion, recall and f-measure. In order to define these measures, four quantities namely true positive, true negative, false positive and the false negative are obtained. The relevant class is considered positive and the other classes are considered negative. For instance, if the relevant class is ‚joy‘, then joy is considered a positive class while the other emotions such as anger, surprise, mixed, neutral etc., are considered as negative classes. Therefore, true positive (tp) indicates the positive instances that is correctly identified as positive, false positive (fp) denotes the negative instances that are wrongly identified as positive, false negative (fn) refers to the positive instances that are wrongly recognised as negative and the true negative (tn) signifies the negative instances that are correctly detected as negative.

Based on this context, accuracy is defined as the ratio of correctly predicted instances to the total number of instances. It is mathematically represented as $\frac{\text{tp+tn}}{\text{tp+tn+fp+fn}}$. Precision is defined as the ratio of true positives to the total number of positively prdicted instances. It is mathematically defined as $\frac{\text{tp}}{\text{tp+fp}}$. Recall denotes the ratio of true positive to the total number of actual positive instances, given by $\frac{\text{tp}}{\text{tp+fn}}$. Another metric namely f-measure combines both the precision and recall and is given by $2 * \frac{\text{precision}*\text{recaall}}{\text{precision+recall}}$.

The performance measures justify the performance of the proposed methodology in identifying the emotions from news articles. The following section discusses the related experiments and results.

4 Results and discussion

This section presents on the experimental set up and the experimental results obtained. Various experiments have been conducted to improve the performance of the proposed framework. The framework consists of two phases namely the data preparation phase and the data mining phase. During the data preparation phase, the data is collected from sina.com. Then the subsequent data processing steps have been implemented in Java. These steps include the HTML tag removal, stop word elimination, stemming (Porter stemming algorithm and the look-up). During, stop word elimination, the stop word lists are carefully chosen such that majority of the stop words are covered. Amny lists have been augmented for this purpose. After that, during the process of stemming, initially only Porter stemming algorithm have been incorporated. But, many commonly used words that revealed emotion expressively have been stemmed incorrectly. Hence, a look up is formulated to ensure that common words related to emotions are stemmed to their root forms correctly. This greatly impacts the performance of classification as only these root words forms the basis of the feature vector. The output of the data preparation step is provided as input to the data mining phase. The data mining phase is executed in Weka 3.6.11, an open source data mining software. During, feature vector formulation, the words from Wordnet-Affect have been considered appropriate to populate the feature vector. In addition, the meaningful words obtained from the stemming process is also appended to the feature vector. As the dimensionality increased highly, feature reduction is adopted followed by evolving classification rules with bagged classifiers.

Initially, the impact of feature reduction is assessed on whether it improves or atleast retains the performance or degrades the performance. For this purpose, Correlation based feature selection is employed. The effect of ensemble is studied in the context of classification accuracy. Then, for the best performing classifier, the precision, recall, f-measure values of every class is presented.

4.1 Impact of feature selection

In this work, CFS algorithm has been adopted to select the significant features. Out of 6542 features, features have been selected as important ones. Nine decision tree classifiers have been used for evolving the classification rules. The decision tree classifiers include Random Tree (RT) (Quinlan 2014), C4.5 (Quinlan 2014), Grafted C4.5 (Fürnkranz and Widmer 1994), Reduced Error Pruning Tree (REPT) (Fürnkranz and Widmer 1994), Naive Bayes Tree (NBT) (Kohavi 1996, August), Best First Tree (BFT) (Shi 2007), Functional Trees (FT) (Gama 2004), Logistic Model Tree (LMT) (Landwehr et al. 2000) and Classification and Regression Tree (CART) (Crawford 1989), In addition to this, the commonly used Support Vector Machine (SVM) (Adankon and Cheriet 2009) also has been employed for comparison purposes. Table 2 reports the classification accuracy (%) obtained from the classifiers with the entire set of features and with only the selected set of features.

Table 2 exhibits the impact of feature selection with respect to classification accuracy. It can be viewed that

Table 2 Performance of various classifier with and without feature selection

Classifier	Accuracy (%)	
	Without feature selection	With feature selection (CFS)
RT	74.01	82.49
C4.5	81.91	85.19
Grafted C4.5	84.23	84.30
REPT	83.07	85.67
NBT	78.27	86.05
BFT	77.12	85.77
FT	78.32	86.25
LMT	79.13	87.21
CART	78.26	86.73

irrespective of the classifiers, all classifiers have provided an improved performance with the selected set of features. Thus, feature selection not only reduces the computational complexity but also provides enhanced performance in the context of emotion prediction. The following sub-section presents the impact of ensemble classifiers.

4.2 Impact of ensembles

Having identified that feature selection yields better results, the next experiment is to assess the impact of ensemble classification. In order to implement the ensemble, bagging technique is adopted. Table 3 presents the classification accuracy obtained through individual classifiers and ensemble of ten classifiers on the selected set of features.

Table 3 portrays the effect of ensemble on the classification accuracy. Again it is apparent that all classification procedures offer an improved performance when ensemble that offered by the individual classifier. Grafted C4.5 has accomplished the maximum possible accuracy when compared to all other decision tree classifiers. It has also achieved a higher performance when compared to the accuracy attained through the widely adopted SVM, which could achieve only 87.48%. Figure 11 graphically represents the impact of feature selection and ensemble on all classifiers.

Table 3 Performance of various classifier with and without ensemble

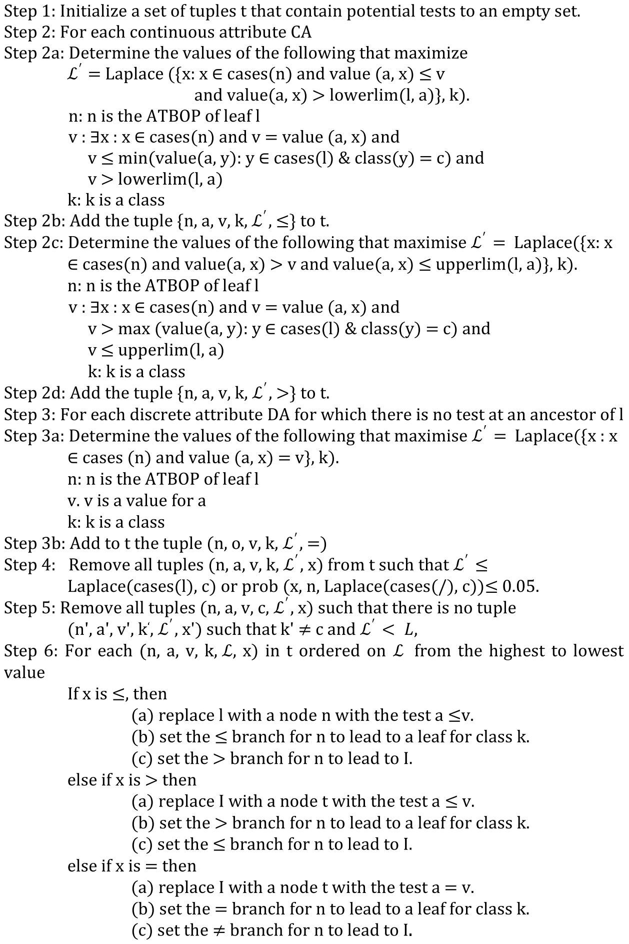
Classifier	Accuracy (%)	
	Without ensemble	With ensemble
RT	82.49	87.07
C4.5	85.19	87.76
Grafted C4.5	84.30	87.83
REPT	85.67	86.08
NBT	86.05	
BFT	85.77	86.29
FT	86.25	80.27
LMT	87.21	87.79
CART	86.73	87.31

The graph evidently demonstrates the positive influence of feature selection and the ensemble on emotion classification by the various classifiers considered for experimentation. The next sub-section presents the other measures for every emotion type obtained by the best performing classifier (Fig. 12).

4.3 Performance of the proposed methodology

This section presents the accuracy, precision, recall and f-measure attained by the bagged Grafted C4.5 on every emotion class namely joy, surprise, fear, sadness, and disgust, mixed and neutral. It also presents the average values for these measures. Table 4 portrays these measures.

Table 4 depicts the satisfactory performance of the proposed framework in predicting the emotions from news



```
Step 1: Initialize a set of tuples t that contain potential tests to an empty set.
Step 2: For each continuous attribute CA
Step 2a: Determine the values of the following that maximize
        $\mathcal{L}' = \text{Laplace}(\{x: x \in \text{cases}(n) \text{ and value}(a, x) \le v$
                    $\text{and value}(a, x) > \text{lowerlim}(l, a)\}, k).$
        n: n is the ATBOP of leaf l
        $v : \exists x : x \in \text{cases}(n)$ and $v = \text{value}(a, x)$ and
           $v \le \min(\text{value}(a, y): y \in \text{cases}(l)\ \&\ \text{class}(y) = c)$ and
           $v > \text{lowerlim}(l, a)$
        k: k is a class
Step 2b: Add the tuple $\{n, a, v, k, \mathcal{L}', \le\}$ to t.
Step 2c: Determine the values of the following that maximise $\mathcal{L}' = \text{Laplace}(\{x: x$
        $\in \text{cases}(n) \text{ and value}(a, x) > v \text{ and value}(a, x) \le \text{upperlim}(l, a)\}, k).$
        n: n is the ATBOP of leaf l
        $v : \exists x : x \in \text{cases}(n)$ and $v = \text{value}(a, x)$ and
           $v > \max(\text{value}(a, y): y \in \text{cases}(l)\ \&\ \text{class}(y) = c)$ and
           $v \le \text{upperlim}(l, a)$
           k: k is a class
Step 2d: Add the tuple $\{n, a, v, k, \mathcal{L}', >\}$ to t.
Step 3: For each discrete attribute DA for which there is no test at an ancestor of l
Step 3a: Determine the values of the following that maximise $\mathcal{L}' = \text{Laplace}(\{x : x$
        $\in \text{cases}(n) \text{ and value}(a, x) = v\}, k).$
        n: n is the ATBOP of leaf l
        v. v is a value for a
        k: k is a class
Step 3b: Add to t the tuple $(n, o, v, k, \mathcal{L}', =)$
Step 4:  Remove all tuples $(n, a, v, k, \mathcal{L}', x)$ from t such that $\mathcal{L}' \le$
        $\text{Laplace}(\text{cases}(l), c)$ or $\text{prob}(x, n, \text{Laplace}(\text{cases}(/), c)) \le 0.05.$
Step 5: Remove all tuples $(n, a, v, c, \mathcal{L}', x)$ such that there is no tuple
        $(n', a', v', k', \mathcal{L}', x')$ such that $k' \ne c$ and $\mathcal{L}' < L,$
Step 6: For each $(n, a, v, k, \mathcal{L}, x)$ in t ordered on $\mathcal{L}$ from the highest to lowest value
        If x is $\le$, then
                (a) replace l with a node n with the test $a \le v$.
                (b) set the $\le$ branch for n to lead to a leaf for class k.
                (c) set the $>$ branch for n to lead to I.
        else if x is $>$ then
                (a) replace I with a node t with the test $a \le v$.
                (b) set the $>$ branch for n to lead to a leaf for class k.
                (c) set the $\le$ branch for n to lead to I.
        else if x is $=$ then
                (a) replace I with a node t with the test $a = v$.
                (b) set the $=$ branch for n to lead to a leaf for class k.
                (c) set the $\ne$ branch for n to lead to I.
```

Fig. 11 Grafting procedure

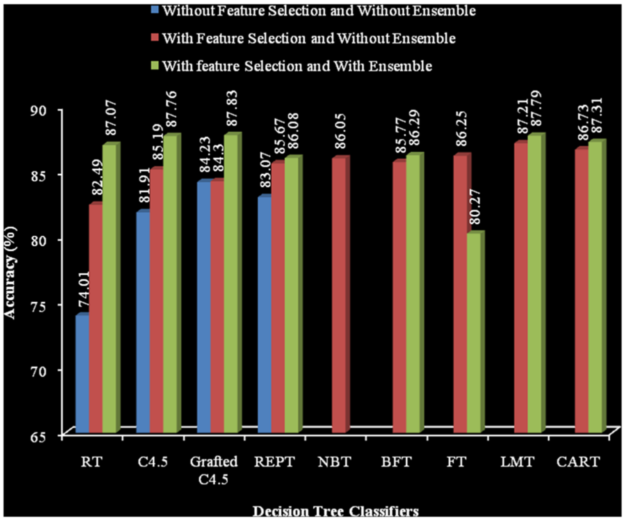


Fig. 12 Graphical illustration of performance evaluation

Table 4 Performance metrics attained by the best performing classifier

Emotion class	Accuracy	Precision	Recall	F-Measure
Joy	0.912	0.967	0.938	0.912
Surprise	0.563	0.669	0.612	0.563
Sadness	1	0.032	0.063	1
Disgust	0	0	1	0
Fear	0.853	0.744	0.795	0.853
Mixed	0.523	0.569	0.573	0.523
Neutral	0.484	0.337	0.397	0.484

articles. The values encourage its usage in real time thereby serving the social community.

5 Conclusion

Emotion prediction from text content can help in many applications. In this work, emotion prediction is attempted on news articles. News articles contain text that does reveal emotions evidently. The proposed framework comprises of two phases namely the data preparation phase and the data mining phase. The data preparation phase includes dataset collection, and text analysis processes to derive the stem words that convey emotions. These words are provided as input to the data mining phase in which a pipeline of procedures are executed to evolve efficient rules for emotion prediction. Investigations have been made to analyze the impact of feature selection, various decision tree classifiers and their ensembles. The maximum possible accuracy of 87.83% is accomplished through the bagged grafted C4.5 decision tree classifier. The results portray the potential of the proposed work and justify its utilization in the real time. The future directions in this work may include developing an improved stemming algorithm that can correctly derive the base forms, developing enhanced hybrid classification algorithm that can predict the emotions with still higher accuracy.

The solution can be a system or scheme which aids an external auditor to audit user's outsourced data in the cloud deprived of the wisdom on the data content. In this paper, we propose a protocol to achieve data hosting with sensitive information hiding in cloud storage; that yields fully the doles of ECDSA of Hung-Zih Liao (2006) viz., increased security, no need to update the host with secrets in the field, double secret key and less computation. The solution of the secure privacy preserved data sharing infrastructure necessitates also several features like Scalable, Decentralized (to evade a single point of failure), Robust, Highly accessible, Control mechanism for data ownership and sharing,a good structure for galvanizing data sharing costs, Directories and Token Vaults, Access based on privileges, facility to run refined data analytics and machine learning.

References

Adankon MM, Cheriet M (2009) Support vector machine. Encyclopedia of biometrics. Springer, USA, pp 1303–1308

Bandyopadhyay S, Saha S (2012) Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer Science & Business Media, Berlin

Bhowmick PK (2009) Reader perspective emotion analysis in text through ensemble based multi-label classification framework. Comput Inf Sci 2(4):64

Bhowmick PK, Basu A, Mitra P (2010) Classifying emotion in news sentences: when machine classification meets human classification. Int J Comput Sci Eng 2(1):98–108

Burget R, Karasek J, Smekal Z (2011) Recognition of emotions in Czech newspaper headlines. Radioengineering 20(1):39–47

Crawford SL (1989) Extensions to the CART algorithm. Int J Man Mach Stud 31(2):197–217

Dietterich TG (2002) Ensemble learning. Handb Brain Theory Neural Netw 2:110–125

Fürnkranz J, Widmer G (1994) Incremental reduced error pruning. In: Machine learning proceedings, pp 70–77

Gama J (2004) Functional trees. Mach Learn 55(3):219–250

Gray EK, Watson D, Payne R, Cooper C (2001) Emotion, mood, and temperament: similarities, differences, and a synthesis. In: Payne R, Cooper C (eds) Emotions at work: Theory, research and applications for management, Wiley, pp 21–43

Hall MA (2000) Correlation-based feature selection of discrete and numeric class machine learning. Working Paper. ISSN : 1170-487X

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam

Hui JLO, Hoon GK, Zainon WMNW (2017) Effects of word class and text position in sentiment-based news classification. Procedia Comput Sci 124:77–85

Jia Y, Chen Z, Yu S (2009) Reader emotion classification of news headlines. In: International conference on natural language processing and knowledge engineering, 2009. NLP-KE 2009, IEEE, pp. 1–6

Kanagaraj R, Rajkumar N, Srinivasan K (2020) Multiclass normalized clustering and classification model for electricity consumption data analysis in machine learning techniques. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-01960-w

Kim Y, Street WN, Menczer F (2003) Feature selection in data mining. Data Mining Oppor Chall 3(9):80–105

Kirange DK (2013) Emotion classification of news headlines using SVM. Asian J Comput Sci Inf Technol 2(5):104–106

Kohavi R (1996) Scaling up the accuracy of Naïve-Bayes classifiers: a decision-tree hybrid. KDD 96:202–207

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160:3–24

Lee H, Choi YS, Lee S, Park IP (2012) Towards unobtrusive emotion recognition for affective social communication. In: Consumer communications and networking conference (CCNC), IEEE, pp. 260–264

Li M, Wang D, Lu Q, Long Y (2016a) Event based emotion classification for news articles. PACLIC 30:153

Li X, Rao Y, Chen Y, Liu X, Huang H (2016) Social emotion classification via reader perspective weighted model. In AAAI, pp 4230–4231

Lin KHY, Yang C, Chen HH (2007) What emotions do news articles trigger in their readers? In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, ACM, pp. 733–734

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval, vol 1. Cambridge University Press, Cambridge, p 496

Mukwazvure A, Supreethi KP (2015) A hybrid approach to sentiment analysis of news comments. In: 4th International conference on reliability, infocom technologies and optimization (ICRITO) (Trends and Future Directions), IEEE, pp. 1–6

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Quinlan JR (2014) C4.5: programs for machine learning. Elsevier, Amsterdam

Sadhana SA, Sai Ramesh L, Sabena S, Ganapathy S, Kannan A (2017) Mining target opinions from online reviews using semi-supervised word alignment model. In: 2017 Second international conference on recent trends and challenges in computational models (ICRTCCM), IEEE, pp. 196–200

Shi H (2007) Best-first decision tree learning, doctoral dissertation, The University of Waikato

Strapparava C, Mihalcea R (2007) Semeval-2007 task 14: affective text. In: Proceedings of the 4th international workshop on semantic evaluations, Association for Computational Linguistics, pp 70–74

Strapparava C and Valitutti A (2004) “WordNet Domains; Wordnet-Affect”. https://wndomains.fbk.eu/wnaffect.html. Accessed 2004

Sulthana AR, Jaithunbi AK, Ramesh LS (2018) Sentiment analysis in twitter data using data analytic techniques for predictive modelling. J Phys Conf Ser 1000(1):012130

Vijayarani S, Ilamathi MJ, Nithya M (2015) Preprocessing techniques for text mining-an overview. Int J Comput Sci Commun Netw 5(1):7–16

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.