Indian Journal of Science and Technology, *Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January, 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Feature Selection using Random Forestmethod for Sentiment Analysis

Jeevanandam Jotheeswaran* **and S. Koteeswaran**

Department of CSE, Vel Tech University, Chennai - 600062, Tamil Nadu, India;
jeevanandamj@gmail.com, s.koteeswaran@gmail.com

Abstract

Background/Objectives: Online review has become important decision support system for the customers to decide on the subscription or purchse. This paper is aiming to suggest a method that improves the accuracy of the classifier. **Methods/ Statistical analysis:** Feature selection for sentiment analysis using decision forest method and Principal Component Analysis (PCA) is used for the feature reduction. The proposed method is evaluated using twitter data set. **Findings:** It is proved, that the proposed decision forest based feature extraction improves the precision of the classifiers in the range of 12.49% to 62.5% when compared to PCA and by 49.5% to 62.5% when compared to decision tree based feature selection. **Application/Improvements:** This method is applicable to product reviews, emotion detection, Knowledge transformation, and predictive analytics.

Keywords: Inverse Document Frequency (IDF), Learning Vector Quantization (LVQ), Opinion Mining, Principal Component Analysis (PCA), Sentiment analysis, Twitter

1. Introduction

One of the sorts of characteristic idiom handling is opinion mining which manages following the mind-set of the individuals with respect to a specific item or theme. This artifact gives automatic extraction of assessments, feelings and sentiments in content furthermore tracks state of mind and sentiments on the web. The method to recognize and concentrate subjective data in content archives is conclusion mining and sentiment analysis[1].

Sentiment analysis is the computational study of people's opinions, appraisals, and emotions toward objects, events and their attributes. In the past few years, this field has attracted a great deal of attention from both the academia and industry due to many challenging research problems and a range of applications. Sentiment classification classifies whether an opinionated document (such as product reviews) or sentence expresses a positive or negative opinion. Subjectivity classification determines whether a sentence is subjective or objective. Many real-life applications, however, require more analysis that is detailed because users often want to know the subject of opinions[2,3]. Sentiment analysis is about deciding the subjectivity, polarity (positive or negative) and polarity quality (weakly positive, moderately positive, categorically positive, and so on.) of a bit of content

The fundamental job of opinion mining is polarity classification, which happens when a bit of content expressing an assessment on a solitary issue is named one of two contradicting sentiments. Presently sentiment analysis depends on vector extraction to express to the most striking and vital content features. This vector can be utilized to order the most important features. Twocommonly utilized features are term recurrence and neighborhood.

Sentiments don't happen just at the archive level, nor are they restricted to a single valenceor target. Henceforth, later work embraced a portion level assessment analysis that utilized diagram based procedures to recognize sentimental from unsentimental areas[4]. Not at all like standard linguistic NLP assignments, for example, outline and auto order, emotion mining for the most part

** Author for correspondence*

spotlights on semantic deductions and emotional data connected with common dialect, and doesn't oblige a profound comprehension of content.

The fundamental issue in sentiment analysis is to distinguish how sentiments are communicated in writings and whether the expressions demonstrate positive (good) or negative (unfavorable) assessments toward the subject[5]. Consequently, sentiment analysis includes recognizable proof of Sentiment expressions, Polarity and quality of the expressions, and their relationship tothe subject.

Inverse document frequency (IDF) is commonly used in Information Retrieval. IDF is defined as -log $2dg_w$ / D, where D is the number of documents in the collection and df_w is the document frequency, the number of documents that contain w. obviously, there is a strong relationship between document frequency, df_w and word frequency, f_w. Computing IDF is critical for many downstream applications in information retrieval. It is easy to compute document frequency over textual documents, but spoken documents are challenging[6, 7].

In this paper, the focus is on feature selection for sentiment analysis using decision forest method. The proposed method is evaluated using twitter data set, and Principal Component Analysis (PCA) is used for the feature reduction.

2. Related Work

K. Church et al.[8] proposed a novel generative theme demonstrate, the Joint Aspect/Sentiment (JAS) model, to mutually separate angles and angle subordinate sentiment dictionaries from online client audits. Trial results showed that the adequacy of the JAS demonstrate in learning perspective subordinate sentiment vocabularies and the functional estimations of the separated dictionaries when connected to these down to earth errands.

J. Jeevanandam et al.[9,10] displayed a novel central part analysis (PCA)-upgraded cosine spiral premise capacity neural system classifier. The new wavelet-bedlam neural system strategy yielded high EEG classification exactness (96.6%) and is truly vigorous to changes in preparing information with a low standard deviation of 1.4%.[11] proposed an online calculation for portion PCA. Test results showed the adequacy of the proposed methodology, both on manufactured information set and on pictures of written by hand digits, with examination to traditional bit PCA and iterative part PCA. The adaptable assessment mining model in view of sentiment tree was made and hence coarse-grained, medium-size and fine-grained (three unique granularities) conclusion mining are all acknowledged in one bound together adaptable model.

J. Chen et al.[12] propose a technique to handle sentiment analysis for Cantonese feeling mining To take care of the issue of needs deals with how to direct Cantonese Opinion Mining. K Khan et al.[13] exhibited a study which secured strategies and techniques that guarantee to empower us to get sentiment arranged data from content. This examination exertion managed strategies and difficulties identified with sentiment analysis and feeling mining. The emphasis was predominantly on machine learning strategies on the premise of their utilization and significance for feeling mining.

Y. Q. Xia et al.[14] proposed the bound together collocation system (UCF) and depicted a novel brought together collocation-driven (UCD) feeling mining strategy. The UCF fused quality sentiment collocations and additionally their linguistic features to accomplish sensible speculation capacity. L. Liu et al.[15] proposed a novel technique to manage the feature level feeling mining issues. The feature grouping relies on upon three viewpoints: the comparing assessment words, the similitude of the features and the structures of the features. The test results showed that the proposed system performed well.

H. Binali et al.[16] assessed existing work and exhibited a feeling mining system and uncovered new ranges of examination in conclusion mining. People, organizations and government can now effectively know the general feeling winning on an item, organization or open strategy. At the center of this field is semantic introduction of subjective terms in records or audits which tries to build up their context oriented intention through feeling mining. Q. Qi et al.[17] embraced the Conditional Random Fields (CRFs) model to perform the assessment mining errands. The proposed system was contrasted with the lexicalized Hidden Markov Model (L-HMMs) based opinion mining strategy in the examination, which demonstrated its altogether better exactness from a few angles.

P. Han et al.[18,19] concentrated on growing fine-grained item feature extractions with insignificant tailor fabricate dialect models and naming. A limit standardized sentence-level word model is proposed for conclusion feature mining. The sentiment feature extraction was then settled by means of network factorization system[24]. Assessment on feature entropies, sentence-entropies and human assessment showed the predominance of our methodology.

3. Methodology

The proposed methods are evaluated using the twitter dataset. The twitter corpus is a large database consisting of relevant and comprehensive information about products - past, present and future. In particular, this paper worked on movie reviews on twitter corpus. It began as a set of shell scripts and data files. The information presented for a title at twitter consists of five main sections: The Title, Production Status, Cast, Crew, and Miscellaneous.

Twitter offers a rating / ranking scale that allows users to rate films by choosing one of ten categories in the range 1–10, with each user able to submit one rating. The points of reference given to users of these categories are the descriptions "1 (awful)" and "10 (excellent)"; and these are the only descriptions of categories. Due to the minimum category being scored one, the mid-point of the range of scores is 5.5, rather than 5.0 as might intuitively be expected given a maximum score of ten. This rating system has since been implemented for television programming on an episode-by-episode basis1.

The steps in the methodology to classify opinions are given below:

Start

Step 1: Extract movie data from Twitter

Step 2: Feature Extraction using JAS method

Step 3: Feature Selection using PCA

Step 4: Classification algorithm using Random Forest method

End.

The features from the movie documents are extracted using Inverse Document Frequency (IDF). IDF represents scaling. When a term *a* occurs frequently in documents, its importance is scaled down due to lowered discriminative power. The term document frequency is computed as follows: for a document set *x* and a set of terms a. A document is modelled as a vector *v* in a dimensional space R^a. When term frequency denoted by *freq (x,a)*, expresses number of occurrences of term a in document x. Term-frequency matrix *TF (x,a)* measures term association *a* regarding a given document *x*. *TF (x,a)* is assigned zero when document has no term and *TF (x,a)* = 1when term *a* occurs in document *x* or uses relative term frequency; term frequency as against total occurrences of document terms.

Estimate the rarity of a term in the whole document collection. (If a term occurs in all the documents of the collection, its IDF is zero.)

$$idf_i = \log \frac{|D|}{\left|\{j : t_i \in d_j\}\right|}$$

with $|D|$: cardinality of D, or the total number of documents in the corpus $\{j : t_i \in d_j\}$: number of documents where the term ti appears (viz. the document frequency) (that is ni,j ≠ 0). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1 + \{j : t_i \in d_j\}$

The features separated are chosen utilizing Principal Component Analysis (PCA), and the proposed choice woodland based featureselection[20,21,22]. PCA is a system which utilizes a direct change to shape a streamlined information set holding the qualities of the first information set. Expect that unique grid contains d measurements and n perceptions and it is obliged to decrease the dimensionality into a k dimensional subspace then its change can be given by

$$Y = E^T X$$

Here $E_{d\times k}$ is the projection matrix, which contains k eigen vectors corresponding to k highest eigen values, and where $X_{d\times n}$is mean centered data matrix.

The target of PCA is to locate a direct change for every class utilizing the preparation designs for that class in the feature space. This gives class-subordinate premise vectors. The primary premise vector is toward most extreme change of the given information. The remaining premise vectors are commonly orthogonal and, all together, augment the remaining changes subject to the orthogonal condition. The key tomahawks are those orthonormal tomahawks onto which the remaining differences under projection are greatest. These orthonormal tomahawks are given by the overwhelming eigenvectors (i.e. those with the biggest related eigenvalues) of the covariance network. In this classifier, every class is described by class-subordinate premise vectors and the quantity of premise vectors utilized for portrayal must be not exactly the dimensionality d of the feature space .

The disservices of the PCA are its worldwide linearity opinion and a missing fundamental measurable model that would take into consideration delicate choices about the participation of a certain item utilizing probabilities.

Dimensionality Reduction (DR) alludes to calculations and systems which make new properties as mixes of the first ascribes keeping in mind the end goal to diminish the dimensionality of an information set. The most imperative DR method is PCA which delivers new properties as direct mixes of the first variables. Interestingly, the objective of a variable analysis is to express the first properties as straight mixes of a little number of shrouded or inactive qualities.

3.1 Proposed Feature Selection Based on Decision Trees

Decision trees are mainstream systems for inductive deduction. They are powerful to boisterous information and learn disjunctive expressions. A choice tree is a k-exhibit tree in which every inside hub indicates a test on a few qualities from data list of capabilities speaking to information. Every branch from a hub relates to conceivable feature values determined at that hub. Furthermore, every test results in branches, speaking to fluctuated test results. The choice tree prompting fundamental calculation is a ravenous calculation developing choice trees in a top-down recursive partition and-vanquish way.

The calculation starts with tuples in the preparation set, selecting best quality yielding greatest data for classification. It produces a test hub for this and after that a top down choice trees affectation partitions current tuples set by test trait values. Classifier era stops when all subset tuples fit in with the same class or on the off chance that it is not qualified to continue with extra partition to further subsets, i.e. on the off chance that more quality tests yield data for classification alone underneath a prespecified edge. In this paper, it is proposed to construct the edge quantify situated in light of data addition and Manhattan progressive group.

In the proposed featureselection, a Decision tree impelling chooses significant features. Choice tree actuation is the learning of choice tree classifiers building tree structure where every inside hub (no leaf hub) signifies quality test. Every branch speaks to test result and every outside hub (leaf hub) signifies class forecast. At each hub, the calculation chooses best segment information credit to individual classes. The best credit to apportioning is chosen by characteristic selection with Information pick up. Trait with most astounding data increase parts the characteristic. Data addition of the trait is found by

$$\inf o(D) = -\sum_{i=1}^{m} p_i \log_2(p)$$

Where p_i is the probability that arbitrary vector in D belongs to class c_i. A log function to base 2 is used, as information is encoded in bits. Info (D) is just average information amount required to identify vector D class label. The information gain is used to rank the features and the ranked features are treated as features in hierarchical clusters[23]. The proposed Manhattan distance for n number of clusters is given as follows:

$$MDist = \sum_{i=1}^{n}(a_i - b_i)$$

A cubic polynomial comparison is determined utilizing the Manhattan values and the limit rule is resolved from the incline of the polynomial mathematical statement. The features are thought to be superfluous for arranging if the slant is zero or negative and pertinent when the incline is sure.

An arbitrary decision backwoods is a group of arbitrarily prepared decision trees. The timberland model is portrayed by various segments. Case in point, we have to pick a group of part capacities (likewise alluded to as \ weak learners" for consistency with the writing). Also, we must choose the kind of leaf indicator. The haphazardness demonstrate additionally has incredible impact on the workings of the woodland. This area examines every segment each one in turn.

Pseudo code for the random forest algorithm:

```
To generate $c$ classifiers:
for $i$ = 1 to $c$ do
    Randomly sample the training data $D$with replacement to produce $D_i$
    Create a root node, $N_i$containing $i$ $D$
    Call BuildTree($N_i$)
end for
BuildTree(N):
if$N$ contains instances of only one class then
    return
else
    Randomly select x% of the possible splitting features in $N$
    Select the feature $F$ with the highest information gain to split on
    Create f child nodes of $N$ ,$N_1$,..., $N_f$, where $F$ has $f$ possible values ( $F_1$, … ,$F_f$)
    for $i$ = 1 to $f$ do
        Set the contents of $N_i$to $D_i$, where $D_i$is all instances in $N$ that match $F_i$
        Call BuildTree($N_i$)
    end for
end if
```

The features chose are characterized utilizing CART, Naïve Bayes and Learning Vector Quantization (LVQ) system. LVQ arranges through direct and unsupervised learning[23]. The model is partitioned into two layers. The main layer is focused layer, in which every neuron speaks to a subclass, and the second is yield layer, in which every neuron speaks to a class. A class may be made out of a few subclasses. The second layer joins a few subclasses into a class through W2 network[24]. So LVQ system may make complex limits through joining a few subclasses into a class. Hence, LVQ is suited to characterize extra messages which have a few subclasses.

LVQ is an algorithm that learns appropriate prototype positions used classification and is defined by P prototypes set $\{(m_j, c_j), j = 1\ldots P\}$, where m_j is a K-dimensional vector in feature space, and c_j its class label. The prototypes number is larger than classes number. Thus, each class is represented by more than one prototype. Given an unlabeled data point x_u, its class label y_u is determined as class c_q of nearest prototype m_q

$$y_u = c_q, q = \arg\min_j d\left(x_u, m_j\right)$$

Where *d* is Euclidean distance. Other distance measures are used depending on the problem.

4. Results and Discussions

The proposed feature selection feature choice is assessed utilizing Twitter dataset. Features are separated from the film information utilizing IDF and feature chose utilizing PCA and the proposed featureselection strategies. The chose features are arranged utilizing CART, Naïve Bayes and LVQ with proposed decision tree and decision backwoods based feature extraction. Figure 1 demonstrates the classification exactness got from LVQ and contrasted, Naïve Bayes classifier, Classification, and Regression Tree (CART). Figure 3 gives the Root Mean Squared Error (RMSE)

It can be seen from Figure 1, the classification accuracy obtained through LVQ with proposed Decision forest based feature extraction achieves the best classification accuracy of 81.25%. It is seen that the proposed Decision forest based feature extraction improves the efficiency of the classifiers in the range of 8.33% to 30% when compared to PCA and by 4.3% to 13.69% when compared to decision tree based feature selection. Figure 2 shows the Average Precision achieved.

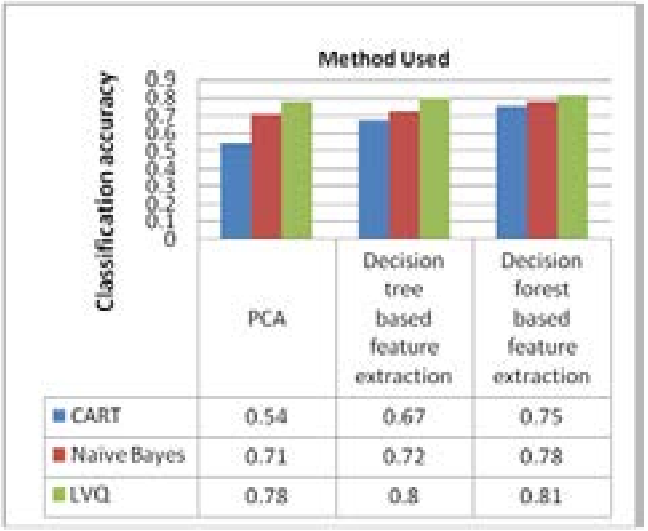


Figure 1. Classification accuracy.

It can be seen from Figure 2, that the precision of the classifiers improves significantly with the proposed decision forest based feature extraction. LVQ with proposed decision forest based feature extraction achieves the highest precision of 0.8125.

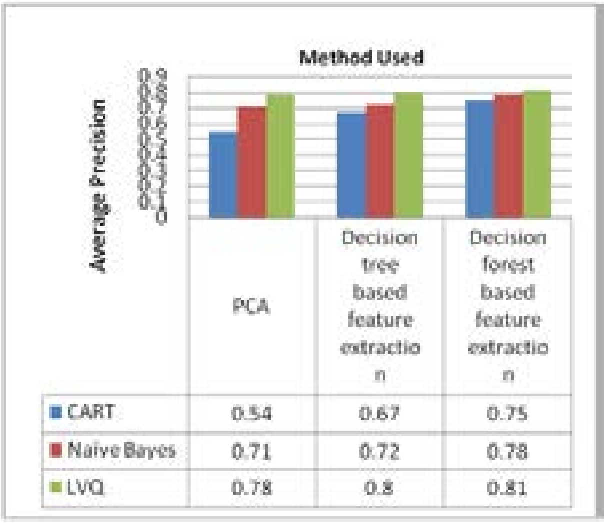
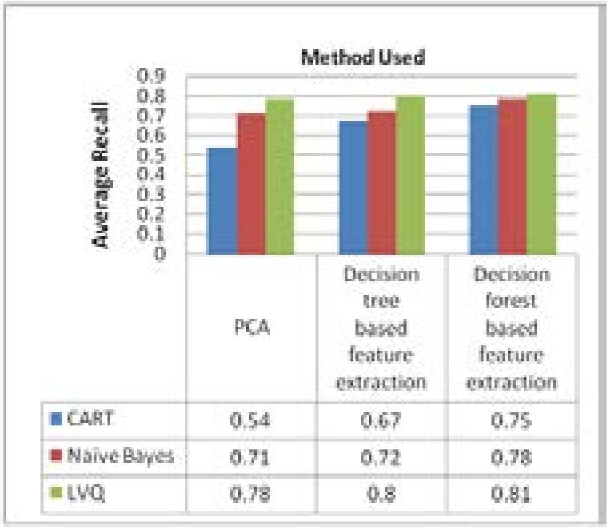


Figure 2. Average precision.

It is seen that the proposed decision forest based feature extraction improves the precision of the classifiers in the range of 12.49% to 62.5% when compared to PCA and by 49.5% to 62.5% when compared to decision tree based feature selection. Figure 3 shows the average recall achieved.



	PCA	Decision tree based feature extraction	Decision forest based feature extraction
🟦 CART	0.54	0.67	0.75
🟥 Naïve Bayes	0.71	0.72	0.78
🟩 LVQ	0.78	0.8	0.81

Figure 3. Recall.

It can be seen from Figure 3 that the recall of the classifiers improves with the proposed decision forest based feature extraction. LVQ with proposed decision forest based feature extraction achieves the highest precision of 0.8145. It is seen that the proposed decision forest based feature extraction improves the precision of the classifiers in the range of 5.38% to 38.03% when compared to PCA and by 1.88% to 11.03% when compared to decision tree based feature selection.

5. Conclusion

In this paper, a feature selection for Opinion mining using decision forest is proposed. Multilayer neural network is used for classification. LVQ type learning models constitute popular learning algorithms due to their simple learning rule, their intuitive formulation of a classifier by means of prototypical locations in the data space, and their efficient applicability to any given number of classes. Movie review features obtained from Twitter was extracted using inverse document frequency and the importance of the word found. Principal component analysis was used for feature selection based on the importance of the work with respect to the entire document. It can be concluded from the experimental results that the LVQ classifier performs better than the CART and Naïve Bayes classifiers. And the proposed decision forest based feature selection improves the efficiency of the classifiers. Further investigations to improve the performance of LVQ are to be studied.

6. References

1. Liu B. Sentiment analysis: A multi-faceted problem. IEEE Intelligent Systems. 2010; 25(3):76–80.
2. Jeevanandam J, Kumaraswamy YS. Feature reduction using principal component analysis for opinion mining. International Journal of Computer Science and Telecommunications. 2012; 3(5): 118–21.
3. Osimo D, Mureddu F. Research challenge on opinion mining and sentiment analysis. The CROSSROAD Roadmap on ICT for Governance and Policy Modeling; 2010. p. 1–9.
4. Jeevanandam J, Kumaraswamy YS. Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure. Journal of Theoretical and Applied Information Technology. 2013; 58(1):72–80.
5. Jeevanandam J, Koteeswaran S. Sentiment analysis: A survey of current research and techniques. International Journal of Innovative Research in Computer and Communication Engineering. 2015; 3(5).
6. Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture, 2003; ACM. p. 70–7.
7. Xu X, Cheng X, Tan S, Liu Y, Shen H. Aspect-level opinion mining of online customer reviews. China Communications. IEEE Explorer. 2013; 10(3):25–41.
8. Church K, Gale W. Inverse document frequency (idf): A measure of deviations from poison, Natural language processing using very large corpora; 1999; Netherlands: Springer. p. 283–95.
9. Jeevanandam J, Koteeswaran S. Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis. ARPN Journal of Engineering and Applied Sciences. 2015; 10(14):5883–94.
10. Jeevanandam J, Koteeswaran S. A weighted semantic feature expansion using hyponymy tree for feature integration in sentiment analysis. ICGCIoT'15; 2015.
11. Honeine P. Online kernel principal component analysis: a reduced-order model. Pattern Analysis and Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012; 34(9):1814–26.
12. Chen J, Liu Y, Zhang G, Cai Y, Wang T, Min H. Sentiment analysis for cantonese opinion mining, emerging intelligent data and web technologies (EIDWT), 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies; Xi'an. 2013. p. 496–500.
13. Khan K, Baharudin BB, Khan A. Mining opinion from text documents: A survey, Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on Digital Ecosystems and Technologies, Istanbul; 2009. p. 217–22.
14. Xia Y Q, Xu RF, Wong KF, Zheng F. The unified collocation framework for opinion mining, 2007, International Conference on Machine Learning and Cybernetics, Hong Kong. 2007; 2:844–50.
15. Liu L, Lv Z, Wang H. Opinion mining based on feature-lev-

el, 2012 5th International Congress on Image and Signal Processing (CISP), Chongqing; 2012. p. 1596–600.

16. Binali H, Potdar V, Wu C. A state of the art opinion mining and its application domains. ICIT 2009. IEEE International Conference Industrial Technology, Gippsland, VIC; 2009. p. 1–6.
17. Qi L, Chen L. Comparison of model-based learning methods for feature-level opinion mining. Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2011; 1:265–73.
18. Han P, Du J, Chen L. Web opinion mining based on sentiment phrase classification vector. 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content; 2010. p. 308–12.
19. Deegalla S, Bostrom H. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification, 2006. ICMLA'06. 5th International Conference on Machine Learning and Applications, Orlando, FL; 2015. p. 245–50.
20. Koteeswaran S, Visu P, Kannan E, Enhancing JS–MR Based Data Visualisation using YARN. Indian Journal of Science and Technology. 2015; 8(11):1–5.
21. VictoSudha George G, Cyril Raj V. Accurate and Stable Feature Selection Powered by Iterative Backward Selection and Cumulative Ranking Score of Features. Indian Journal of Science and Technology. 2015 June; 8(11):1–6.
22. Devi KS, Ravi R. A new feature selection algorithm for efficient spam filtering using adaboost and hashing techniques. Indian Journal of Science and Technology. 2015; 8(13):1–8.
23. SasiRegha R, Uma Rani R. A Novel Clustering Based Feature Selection for Classifying Student Performance. Indian Journal of Science and Technology. 2015 Apr; 8(S7):135–40.
24. Ruba K V, Venkatesan D, "Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts", Indian Journal of Science and Technology. 2015 July; 8(13):1–9.