INTERACTION OF ENSEMBLE FEATURE TECHNIQUES WITH INCREMENTAL LEARNING USING STOCHASTIC GRADIENT DESCENT OPTIMIZATION ON NEWS CATEGORIZATION

[1]Abdullah, Khadijha-Kuburat Adebisi, [2]Sodimu, Segun Michael, [3]SolankeOlakunle O, [4]Efuwape Biodun Tajudeen[1]

uwaizabdullah9@gmail.com, [2]princesegzy01@gmail.com, [3]olasolanke@gmail.com.
[4]efuwapebt@yahoo.com
Department of Mathematical Sciences
Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria.
Corresponding Email: uwaizabdullah9@gmail.com,Phone:+2348060046592

ABSTRACT

The problem of categorization lies in the number of features that exist in the text documents. Single feature method can pose irrelevant, redundant, noise with high dimensionality, hence, increases computational cost when treated as an independent feature. Then, become relevant when combine with other feature techniques and creates interaction between features. In this paper, different feature techniques are combined to enhance news categorization. Feature selection methods filter the features but scale to high-dimensionality that results to irrelevant features and lack of understandability. Reduction techniques extract relevant features to obtain low dimensional representation as it combines to minimise the error and maximise the variance. In this study, *t*he concept of the feature reduction does not solve the presence of missing values and noise which has many features associated with nonlinear models for large data set. Therefore, Stochastic Gradient Descent (SGD) optimization with $L1$ regularization which solves the effect of missing values and noisy gradient. These results show that SGD is least affected by dataset sparsity and it shows that SGD algorithm provides potent predictions when handling sparse data. Finally, performance evaluations are done.

Keywords: Stochastic gradient descent, Interaction, Regularization, Incremental, PCA

Accepted Date: 20 Nov., 2020

Introduction

Feature space faced the problem of high dimensionality in natural language processing (NLP) due its representation. Though, the features contained in texts are very complex, this result to high memory space, time complexity and poor classification performance (Verbeek, 2000).Many researchers have exploited different feature methods in NLP tasks with difficulty in interpretation and harmonisation of the classification performances. According to Werner *et al.,*(2014), different subsets of features were chosen to find an optimal feature set but cannot be used for the generalizable classification problem. This led to non-uniformityin the selection of features for classification which are used due to good categorization performance (George, 2003). Biricik *et al,* (2012) presented that in order to reduce computational cost of the algorithm, fewer features can be imbibed. It is also known that relevant information is chosen to perform the desired task using reduced representation if features are carefully extracted. Apparently, some algorithm ignore less relevant features in the classification process and work only on the relevant ones such decision tree algorithms (Quinlan, 1993). Meanwhile, Multi-layer perceptron exclude the irrelevant features automatically with strong regularization (Duch *et al,* 2001). This results to improper interpretation and sub-optimal design and/or defection of other features. To provide robust categorization for text documents, different

feature techniques are required(Guyon *et al,* 2008).Many related work used in the literatures involve document frequency (DF), Information Gain (IG), Chi Square, concept indexing (CI), Latent Semantic Index (LSI), Principal component analysis (PCA) etc. It was shown according to Abbasi *et al.,* (2011)that information gain was better than Document Frequency (DF) with less frequency than a certain threshold, but with little impact on filtration (Myra *et al.,* 2016) Most existing text categorization occurs in a single class but it is better for feature to occur in more than one class, this forces information gain to be among features that occurred in multiple classes. Similarly, experiments show that Chi-Square ($\chi 2$) improves the performance of text categorisation (Meesad *et al.,* 2011) with relatively low time complexity (Uysal & Gunal, 2012)

Existing text feature techniques are categorised into feature selection and reduction, this involves filtering, fusion, mapping and clustering. These categories are used in the evaluation of news categorization in this study except for fusion method. Fusion category such as K-Nearest Neighbour (K-NN) gives each feature a weight and lead to multiple weighted methods for the algorithms. Thus, information Gain (IG), word frequency and Mutual Information (MI) e.t.c. involve filtering of features and are suitable for large scale text extraction. According to Abbasi *et al,*(2011), information gain has been shown to be better as feature selection and improves performance of text categorization with good results (Meesad *et al.,* 2011)and takes only one contribution of class feature at a time but select low quality feature subsets. Meanwhile, scale to very high-dimensional datasets and do not consider much interaction with classifier, hence, degrades categorization methods. Feature extractions were shown to be Nondeterministic Polynomial (NP)-hard (Candes *et al*, 2011) which can be equivalent to convex program with high probability. The linear combination of the features was returned with no information attached to the class variable. Principal Component Analysis (PCA) and Chi-Square ($\chi 2$) are use as feature extraction methods for mapping and clustering respectively. Chi-Square ($\chi 2$) clusters features to have low time complexity after ranking (Gomez *et al,* 2012)thus, reduces computational cost and improves visualisation in text processing (Josse *et al.,* 2011) (Gomez *et al,* 2012). In order to preserve important information in high dimensionality and feature redundancy, PCA used singular value decomposition (SVD) to preserves most of the information in the datasets (Josse et *al.,* 2011).

Consequently, training datasets may still contain some errors such as corrupted data as well as failure to process missing elements. Replacing the missing elements with an extreme value cannot be apply when a significant portion of the measurement matrix is unknown, therefore, lead to noise. There have been several attempts to improve and solve the problem of sparsity and other forms of errors in text representation with time complexity, efficiency and computational cost. (Candes *et al*; 2011) proposed Robust PCA (RPCA) as convex de-convolution method for extracting low dimensional subspace structure in the presence of errors. Apparently, uniformly sparsely distributed with element-wise corruptions requires an expensive time to perform decomposition to find the Principal Components (PCs).(Feng *et al.,*2012) employed robust PCA methods and implemented all the samples in batch manner but require access to every sample on iteration of the optimization. Though, the method required that subspaces changes gradually by memorising all samples. Whereas, only covariance is needed when considering standard PCA, therefore, there are some algorithms that are faster with memory though their practicability is not in used for large datasets (Xu *et al.,* 2012).

In this study, SVD is used to compute PCA to detect and extract small signals from noisy dataset. The SVD technique obtains the eigenvalues and eigenvectors for matrices to determine the size of the matrix needed for dimensionality reduction(Li *et al.,* 2017).Though, the projection of computing SVD is easy for PCA but it is time consuming with large storage requirements when the matrix size is big. There are lots of methods to solve this issue, for instance, (Zou & Tibshirani 2006) formulated PCA as a regression problem using Sparce PCA (SPCA), few of the input variables were combined by imposing the Lasso (elastic net) constraint in order to allow interpretation of the principal components. RPCA does not support low-dimensional approximation or outliers but very

vast in real application such as subspace recovery (Liu *et al.,* 2013)but cannot be used for big data as well as it creates noise in an incremental setting. Halko *et al.,* (2011) proposed that the matrix of the SVD can be accurately reconstructed from the arithmetic of the small matrix with low reconstruction error, but Halko's approach is known as an algorithm of the randomised SVD. Li *et al.,*(2017) developed algorithm971 by modifying the preconditioning step so that the calculation time is improved. The goal of PCA is to learn a linear transformation by finding a lower dimensional space U to transform the dataset ($X=\{x_1 \cdots x_n\}$) from a high dimensional space $\Re^d$ to low dimension space $\Re^k$ n and i^{th} samples is extracted for the most important dataset. PCA works well if interest is not on the interpretation of the features but on random projection of data matrices to make a large matrix smaller (Halko *et al.,* 2011), (Li *et al.,* 2017). These approaches are employed on incremental implementation where only a subset of the data matrix is loaded into the memory and used to update the classifier. Finally, PCA algorithms are used where data are incrementally observed such as subspace tracking Balzano (such as subspace tracking Balzano *et al.,* 2018) The stochastic approximation algorithms have been shown to be computationally effective on different classification problems with good empirical performance (Nemirovski et *al.,* 2009). Using gradient descent-based method as an objective function where the initial vectors are updated to the reverse direction utilises all the data to calculate the gradient. This calculation of the full stochastic gradient is decomposable to the sum of the gradient of individual data points yet not scalable. This method is known as (Oja & Karhunen, 1985) method or the generalized Hebbian algorithm (Oja & Karhunen, 1985)With the success of stochastic gradient descent (SGD) algorithms in the large datasets, (Arora *et al.,*2012) formulated PCA objective as a stochastic optimization problem that reduces computational cost (Alexander & Tapani , 2010) and solve the problem of missing data in PCA. Our work is build on Halko *et al.,*(2011), and Li *et al.,*(2017)that randomly project data matrices to formulate the PCA objective as a stochastic optimisation problem computed with Singular Value Decomposition (SVD). Categorization with stream datasets can adapt to changes (Myra *et al.,* 2016) either implicitly (Sebastian *et al.,* 2015)or explicitly (Vasileios, 2017). In this work, different feature methods interact with PCA objective function using incremental stochastic gradient descent (SGD) to optimize the objective function on $L1$-regularization penalty in order to reduce the presence of errors and sparsity in samples at instance time (t). The problem finds the maximal variance k-dimensional subspace in the distribution

(I) The stream datasets are generated incrementally and independently by underlying probability distribution D (i) to detect when changes occur in the datasets Errors, sparsity representation and noise are dealt with using Stochastic Gradient Descent optimization on L1-

(i) regularization penalty.

(ii) Comparative studies of results are obtained in terms of performance from the features techniques using attributes such as accuracy, time complexity and memory requirement.

There maining work is organised as follows: Part 2 presents the research concepts and proposed method with SGD as optimization classifier on news categorization while 3 presents the experimental analysis results and discussion. Finally, conclusion and recommendation are presented.

Material and Methods

Statistical NLP has emerged as the primary option for modeling natural language tasks but difficult to find effective features for text representation. This led to the motivation of representations of feature of words in low-dimensional space. In this section, three different feature techniques interact with incremental stochastic gradient descent as optimisation objective function. The sub-section describes the proposed methods.

Also, preprocessing of text document considered how content can be represented in form of vector in D-dimensional vector. The document x^j is in numerical feature vector $v_j := [wt_{1,}, \ldots wt_{D,j}]$.

$$wt_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \qquad (1)$$

where

$f_{f,j}$ = term frequency and

df_f = document collection.

The Feature Extraction Techniques

In this study, three different features algorithms are adopted for effective and efficient classification with optimization where individual feature techniques are reduce to their respective *k*-best as describe below:

Information Gain (IG): This measures the entropy concepts that filter strategies in order to evaluate features separately. Let distribution D with N examples $E_i = (x_i, y_i)$ such that $I = 1 \cdots n$ is associated with a feature vector $v_{i1} \cdots v_{in}$, D attributes and C is the class. Then, information gain of I ($G(I)$) is described using entropy as in equation 2 respectively in form of single and multi-class labels;

$$G(I) = -\sum_{i=1}^{n} \Pr(t_i) \log \Pr(t_i) \quad (2a)$$

$$G(I) = -\sum_{i=1}^{n} \Pr(t_i) \log \Pr(t_i) + \Pr(I) \sum_{i=1}^{n} \Pr(t_i | I) \log \Pr(t_i | I) \quad (2b)$$

i. ***Chi-Square*** (χ^2):This involves the degree of relationship of features between the expected count and observed counts deviate from one another as in equation 3.

$$\chi^2 = \sum_{i=0}^{N} \frac{(observed - \exp ected)^2}{\exp ected}$$

$$\chi^2(N, i, k) = s_{e_{i \in \{0,1\}}} s_{e_{k \in \{0,1\}}} \frac{(N_{e_i e_k} - E_{e_i e_k})^2}{E_{e_i e_k}} \quad (3)$$

Where

e_t and e_k = contingency table binary variables

$e_i = 1$ (if $e_i = 0$ (document not exist in i)

$e_k = 1$ (if $e_k = 0$ (document not exist in class k)

N = observed frequency in class kcontaining termsi

I. The Singular Value Decomposition of PCA

The PCA space consists of a number of principal components (pc= *k*), in order to compute the PCs, SVD is used to compute PCA analytically by finding covariance matrix of eigenvalues and eigenvectors which form the basis for *principal components* (PCs) as described in equation (4a&4b):.

For $Wt = \{wt_1 \cdots wt_n\} \in \Re$

Where Mean= $\mu = \frac{1}{n} \sum_{i=1}^{n} wt_i$ (4a)

Covariance $(C) = \Sigma(wt_i - \mu)]$ (4b)

The SVD diagonalise the matrix $Wt = \{wt_1 \cdots wt_n\} \in \Re^{D \times N}$ as input data with N points, where each point has $D = \{d_1, d_2 \cdots d_n\}$ dimensions. Wt is a $D \times N$ matrix, with the point wt_i .In solving the SVD of matrix Wt with rank r, SVD factorised into product of three (3) matrices of equation (5) with left singular vectors $U = \{u_1, u_2 \cdots u_m\}$ for $(M \times N)$ and right singular vector $A = \{a_1, a_2 \cdots a_n\}$ for $(N \times N)$, then, U, A are orthogonal matrices with E the eigenvalues such that $\delta = \sqrt{\lambda_i}$ therefore, $\Sigma = diag(\delta_1 \cdots \delta_r)$, $r = \min(D \times N)$ with $\delta_1 \geq \delta_2 \geq \cdots \delta_r$. The equation for the matrix Wt can rewritten as $W = U_r \Sigma_r A_r^T$.

$$Wt = U \Sigma A^T \quad (5)$$

The k largest singular value from matrix $(k < r)$ is obtained and a new matrix is form as $Wt_k = U_k \Sigma_k A_k^T$ .Thus, left eigenvalue decomposition problem is solve as the covariance matrix as the product of $Wt\ Wt^T$. The equation is obtained by substituting W by its SVD, this implies:

Then if $A^T A = 1$, $WtWt^T = U \Sigma^2 U^T$ (6)

Where

Σ^2 = diagonal matrix of eigenvalues of $WtWt^T$, similarly, right singular vector A represent the Eigenvector decomposition of $Wt^T Wt = A \Sigma A^T$.

Formulating PCA as objective function $W = \{v_1 . v_2 \cdots v_k\}$, this is done as a low dimensional space of PCA is constructed by combining selected Pcs in k with most eigen values so that maximum amount of variance direction in the data are kept. Solving PCA via SVD is equivalent to solve the Eigen-problem for the covariance.

$$Wt_k = U diag(\delta_1 \cdots \delta_k, 0, \cdots 0) A^T \quad (7)$$

$$Z = Wt^T D = \sum_{i=1}^{n} Wt^T (wt_i - \mu) \quad (8)$$

Where $Z \in \Re^k$
represent data after projection onto the PCA space while the other eigenvectors or PCs are neglected but to prevent this, the PCs vector can be optimised such that the projection yields the minimum error as well as selecting large number of PCs and increases the total variance of W Also, in the presence of missing values and noise, the gradient descent update for W is projected on the data that describe the maximum variation. With incremental stochastic gradient descent (OGD) only a subset of the data matrix is loaded into the memory and used to update the iteration

Algorithm: Calculating PCAwith SVD.

1. *Given a Matrix* $Wt = \{wt_1 \cdots wt_n\} \in \Re^{D \times N}$, *where* n *is number of samples,* x *with* i^{th} *sample* (x_i).
2. *Generate mean as in Equation [4a]*
3. *Generate Covariance as in Equation [4b]*
4. *Construct a Matrix* $Wt = D^T$, $Wt(N \times M)$
5. *Calculate SVD for matrix Wtas in Equation [5].*
6. $r = \sqrt{v_1} \cdots v_k$ *with PCs as columns A.*
7. *Select Eigenvectors with highestEigenvalues* $Wt = \{v_1 . v_2 \cdots v_k\}$*for PCA space.*
8. *PCA of matrix Wt has low dimensional space* $Z = \sum_{i=1}^{n} Wt^T (x_i - \mu)$

Incremental Stochastic Gradient Descent Optimization (SGD)

In this section, the major concern is large datasets which cannot be loaded entirely into the computer memory. Using incremental SGD, a small randomly selected subset of the training samples optimizes the feature sets by minimizing the least square errors (z_i). Thus, maximize the classification prediction with L_i regularization penalty of the parameters to achieve sparse representation with coefficient larger than threshold. The problem is formulated as follows: Given dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ are vector and $i = \{1,2,\cdots N\}$ with features set. A set $D \subseteq \Re$ is a convex set if for any $\vec{x}, \vec{y} \in X$, with features $x_n \in \Re^d$ and label $y_n \in \Re$ and loss function of a linear model $R(w)$ is the regularization term which prevents the model from overfitting with parameters $w \in \Re^d$ where $R(w) = C\sum_i |w_i|$, C controls the degree of regularization tuned by cross validation as express in equation 9 (a &b).

$$\min_w L(w) = \sum_{i=1}^{n} \ell((x_n), y_n; w) - R(w) \quad (9a)$$

$$\min_w L(w) = \sum_{i=1}^{n} \ell((x_n), y_n; w) - C\sum_i |w_i|$$

Let $x_n = \{w, z_i\}$

$$\min_w L(w) = \sum_{i=1}^{n} \ell(w, z_i), y_n) - C\sum_i |w_i| \quad (9b)$$

Incremental SGD is done in a sequential prediction using online datasets of weight (w) time (t) which is optimised with explicitly constraints $w \in W$ and the prediction is base on previous weight. As the weights of the feature is updated at training samples i then $w \in W\, \Re^d\, x_i$ unlabelled text is added, $f(w)$, is observed output and it incurs some loss function $L(w_i z_i)$.

Then $$f(w) \Rightarrow w_0 + \sum_{i=1}^{t} L(w_i z_i) \quad (10)$$

At each round the learner chooses decisions from a convex feasible set with a learning rate $\eta > 0$, Therefore, updating the weight (w) at time (t): $f(w_{t+1}) = w_t - \eta \nabla L(w_{it}, z_i)$

With iteration counter k then:

$$f(w_{t+1}^{k+1}) = w_t^k - \eta_t \nabla L(w_{it}, z_i) - C\sum_{i=1}^{n} |w_i| \quad (11)$$

Hence, let learning rate $\eta_t = \frac{1}{\lambda t}$ for convergence analysis of stochastic approximations which will satisfy the conditions

$\sum_{t=0}^{n} \eta_t < \infty$ and $\sum_{t=0}^{n} \eta = \infty$ as t increases

Using $\eta_t = \frac{1}{\lambda t} \Rightarrow$

$$f(w_{t+1}^{k+1}) = w_t^k - \frac{1}{\lambda t} \nabla L(w_{it}, z_i) - C\sum_{i=1}^{n} |w_i| \leq 1 \quad (12)$$

Experimental Results and Discussion

We carried out step by step processes to evaluate the impact of feature techniques on news categorization. The dataset crawled (8,454 News) from different categories and about 35000 features vector (FV)are generated. Only 6000 dataset and top 20000 FV were used in order to prevent error in memory usage. The dataset are used in sequence of 1000, 2000, 3000, 3000, 4000, 5000 and 6000 on each feature technique. In Table 1, the features are reduced to k = 3000 so that accuracy and time can be studied.

Table 1: Accuracy and Time for each Feature Techniques with TFIDF on Incremental Dataset

Training	TF-IDF (20000)		TF-IDF (20000) + IG (3000)		TFIDF (20000) +CHI2(3,000)		TF-IDF (20000) + PCA (3,000)	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
1000	0.871	0.384	0.834	0.053	0.922	0.100	0.854	0.048
2000	0.805	0.361	0.882	0.051	0.902	0.095	0.918	0.046
3000	0.937	0.367	0.858	0.052	0.923	0.093	0.822	0.053
4000	0.900	0.354	0.907	0.051	0.925	0.093	0.926	0.056
5000	0.917	0.360	0.929	0.051	0.884	0.093	0.930	0.045
6000	0.935	0.249	0.929	0.034	0.931	0.063	0.935	0.033

From the Table 1, it shows that combining Information Gain (IG) with TFIDF reduces the time spent for each sequence but the accuracy increases as the sequence increases except when the sequence is 3000. The same occurs with Chi2 with increase in the accuracy better than IG but with higher time than IG. However, Principal Component Analysis (PCA) has reduced time than IG and Chi2 and accuracy increases as the size of the dataset increases as depicted in Figure 1(a) and 2(a).

In Table 2, the dataset are subject to the same feature vector but at one instance of 6000. The following results are achieved and it shows that the accuracy are better for each of the feature techniques but the time was too high compared to the time generated with incremental datasets most especially with TFIDF as shown in Figure 1(b) and (2b).

Table 2: Dataset at One Shot with Accuracy and Time for each TFIDF with Feature Technique

Training	TF-IDF (20000)		TF-IDF (20000) + IG (3000)		TFIDF (20000) + CHI2,(3,000)		TF-IDF (20000) + PCA (3,000)	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
6000	0.954	13.200	0.945	2.088	0.953	2.984	0.951	2.395

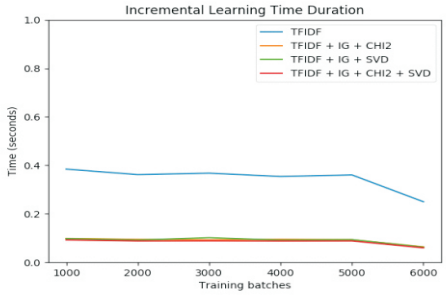
In order to make the features techniques interact with one another, these features are combined with TFIDF in the proportion that each feature would be reduced to k = 15000, 10000 and 5000 for IG, Chi2 and PCA respectively in the incremental sequence of 1000, 2000, 3000, 4000, 5000 and 6000 as depicted in Table 3 and Figure 3(a) and 4(a)

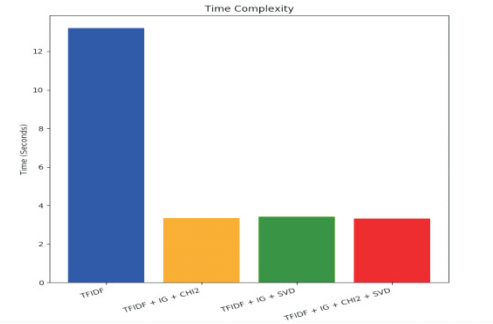
Table 3: Accuracy and Time for Feature Techniques with TFIDF on Incremental Dataset

Training	TFIDF(20k) + IG(15k)+CHI2(10k)		TFIDF(20k) + IG(15k) + PCA(5k)		TFIDF(20k) + IG(15k) + CHI2(10k) + PCA(5k)	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
1000	0.856	0.098	0.674	0.097	0.808	0.092
2000	0.884	0.094	0.907	0.091	0.906	0.088
3000	0.872	0.092	0.920	0.101	0.869	0.088

(a) (b)

Figure 3(a) and (b): Accuracy onInteraction of Features in Incremental and One Shot





(a) (b)

Figure 4(a) and (b): Time Complexity on Interaction of Features in Incremental and One Shot

Conclusion and Recommendation

The purpose of this study is to provide optimization based context for ensemble feature techniques which can be viewed as providing optimal solutions. Finding these solutions involve using Information Gain and Chi*2*as feature selection as well as using covariance of the Singular Value Decomposition (SVD) to find Principal Components which strives to reduce redundancy between features. It was observed that using the top most minimum terms (k-best) related to corpus generated is relevant with the class of the dataset. Experimentation is conducted to verify the effectiveness of various feature techniques which enhance the categorization performance. The proposed method performs well for ensemble feature techniques interacting with one another, hence the accuracy and performance enhanced by adopting the proposed methodology with appropriate time consumption. These results show that SGD is least affected by dataset sparsity and it shows that SGD algorithm provides potent predictions when handling sparse data.

For future work, we intend to improve the performance of categorization with the machine learning methods or deep learning techniques with word embedding as representation technique to tackle the challenge of categorization prediction on online or streams big datasets.

Acknowledgement

Data used in the paper were collected from punch.com. We would like to thank the developer of NLTK toolkit (Steven Bird, Edward Loper and EwanKlein) and machine leaning library developed by Google, and some libraries such as Keras and SKLearn.

References

Abbasi A., France S., Zhang Z., & Chen H. (2011). Selecting attributes for sentiment classification using feature relation networks. Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering, 447–462.

Alexander I. and Tapani R. (2010). Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research 11 1957-2000.

Arora R., Cotter A., Livescu K. and Srebro N. (2012). Stochastic optimization for PCA and PLS. In *Allerton Conference*,861-868.

Balzano L., Chi Y. and Lu Y.M. (2018). Streaming PCA and subspace tracking: The missing data case. *Proceedings of the Institute of Electrical and Electronics Engineers Transactions 106(8): 293-1310.*

Biricik G., Diri B., Sonmez A.C. (2012). Abstract feature extraction for text classification. Turk J Electric Engineering & Computer

Science, 20(1): 102-1015.

Candes E. J., Li X., Ma Y. and Wright J. (2011). Robust principal component analysis, Journal of the Association of Computing Machinery (JACM), Article 11, 58(3):1-36.

Duch W., Adamczak R., Grabczewski K. (2001). *A new methodology of extraction, optimization and application of crisp and fuzzy logical rules.* Institute of Electrical and Electronics Engineers Transactions Transactions on Neural Networks, 12: 277-306.

Feng J., Xu H., and Yan S. (2012). Robust PCA in high-dimension. A deterministic approach. In International Conference on Machine Learning,1-8.

George F. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3:1289–1305.*

Gomez J. C., Boiy E., Moens M. F. (2012). Highly discriminative statistical features for email classification. *Knowledge and Information System, 31(1): 23-53.*

Guyon I., Gunn S., Nikravesh M. and Zadeh L. A. (2008). *Feature extraction: Foundations and Applications*, Springer 207: 29-64.

Halko, N., Martinsson, P.G., Shkolnisky, Y., M, T. (2011). An algorithm for the principal component analysis of large data sets. Society for Industrial and Applied Mathematics (SIAM) *Journal on Scientific Computing 33(5): 2580-2594* .

Zou H. T. and Tibshirani R.(2006). Sparse principal component analysis. *Journal of Computational. Graph Statistic. 15(2): 265–286.*

Josse J., Pagfies J. and Husson F. (2011).Multiple imputations in principal component analysis. *Advances in Data Analysis and Classification, 5(3):231-246.*

Li, H., C L.G., Szlam, A., Stanton, K.P., Kluger, Y., Tygert, M. (2017).Algorithm 971: An implementation of a randomized algorithm for principal component analysis. Association of Computing Machinery Transactions on Mathematical Software 43(3):1-36.

Liu G., Lin Z., Yan S., Sun J., Yu Y. and Ma Y. (2013). Robust recovery of subspace structures by low-rank representation. Institute of Electrical and Electronics Engineers Transactions *Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184.

Meesad P., Boonrawd P. and Nuipian V. (*2011).* A chi-square-test for word importance differentiation in text classification *International Conference on Information and Electronics Engineering IPCSIT. 6: 110-114.*

Myra S., Eirini N., and Max Z.. (2016). Opinion stream mining. Springer US, Boston, MA, https://doi.org/10.1007/978-1-4899-7502-7_905-11–10.

Nemirovski A A., Juditsky A., Lan G. and Shapiro A. (2009). Robust stochastic approximation approach to stochastic programming. Society for Industrial and Applied Mathematics (SIAM) *Journal on Optimization, 19(4):1574-1609.*

Oja, E., Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix author links open overlay panel. *Journal of Mathematical Analysis and Applications 106(1):69-84.*

Quinlan J. R. (1993). *C4.5: Programs for machine learning,* San Mateo, *Morgan Kaufman Publishers, Incorporation. 16(2): 235-240.*

Sebastian W., Max Z., Eirini N., and Myra S. (2015). Ageing-based multinomial naïve-bayes classifiers over opinionated data streams. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 401–416.

Uysal A. K., Gunal S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based System. 36(6): 226–235.

Vasileios I., Annina O. and Eirini N. (2017). Sentiment classification over opinionated data streams through informed model adaptation. *International Conference on Theory and Practice of Digital Libraries. Springer, 369–381.*

Verbeek J. J. (2000). Supervised feature extraction for text categorization. Tenth Belgian-Dutch Conference on Machine Learning (Benelearn '00) Dec. 2000, Tilburg,

Netherlands.1-7.

Werner P., Al-Hamadi A., Niese R., Walter S., Gruss S., and Traue H. C. (2014). Automatic pain recognition from video and biomedical signals. In 2014 22nd *International Conference on Pattern Recognition* (Manchester), 4582–4587.

Xu H., Caramanis C., and Sanghavi S. (2012). Robust PCA via outlier pursuit. Information Theory, Institute of Electrical and Electronics Engineers Transactions on, 58(5):3047–3064.