Robust Feature Selection for Text Categorization using Cross Validation

Arpita[1], *Dr. Pardeep Kumar*[2], *Dr. Kanwal Garg*[3]
Ph.D. Scholar[1], *Associate Professor*[2], *Assistant Professor*[3]
Department of Computer Science and Applications
Kurukshetra University, Kurukshetra
arpitagrover05@kuk.ac.in[1], *pmittal@kuk.ac.in*[2], *gargkanwal@kuk.ac.in*[3]

Abstract

High accessibility to computational facilities encourage generation of large volume electronic data. Accumulation of this huge amount of data persuade researchers to critically analyze the data so as to extract maximum possible benefit for wiser decisiveness. Sentiment analysis is one such area where social media sites are the source of assembling data for analyzing opinion of people towards any subjective aspect. Further, accurate interpretations from such massive content require a mechanism for feature reduction. Thereby, it was observed that feature extraction, dimensionality reduction and feature selection were three major phases of producing reduced set of attributes. But, all three had some limitations in tackling enormous set of features. Therefore, a hybrid combination of extraction, reduction and selection is proposed in this paper. Feature extraction converts lump of text to meaningful lexicons or tokens removing redundancy but faced the curse of dimensionality for which dimensionality reduction is applied next. Besides, the dimensionality reduction also prevents problem of overfitting which could arise due to genetic algorithm for feature selection. While genetic algorithm, being final step of feature selection, helped election of most promising features for classification.

Keywords: Feature Extraction, Dimensionality Reduction, Learning Algorithm, Genetic Algorithm.

1 Introduction

With a fast growing Internet, feedbacks regarding products, policies, services and people, are being generated in large volume via different online portals. High accessibility to such computational facilities provide valuable electronic information. This huge volume data is known as Big data.

One of the biggest source of big data in today's era are the social media sites. Table 1 specifies the progress and rundown in utilization of some most popular social media sites. It can be clearly deduced from Table 1 that Twitter and Facebook shares significant amount of data. This accumulation of data creates a huge unstructured big data which is analyzed for making decisions. Big data from such places is considered as a great source of real time estimation due to its high frequency of creation and low cost integration[22]. From past few years to assess the feelings of social media users towards a subject, a common method called sentiment analysis or opinion mining is increasingly been used.

Table 1: Average Information Shared Per Second [10]

S.No.	Social Platform	Rise of Data Per Second	Rank
1.	Skype	310,832	3
2.	Facebook	39,852,495	1
3.	Twitter	554,301	2
4.	Instagram	84,352	5
5.	Tumblr	241,217	4

Big data era leads to progressive intensification of data samples in conjunction with features to applications like sentiment analysis. This expeditious expansion of data make proficient data management very challenging [17]. Therefore, the primary obligation for acquiring desired knowledge out of this huge data would be to deal with its bulkiness first.

The remaining paper is organized as follows: Section 2 includes discussion of various author's work in concerned arena. Further, entire methodology opted in this research for selection of optimum features is postulated in section 3. Then, the proposed algorithm for feature selection is mentioned in section 4. Thereafter, Section **??** postulates generated results. Finally, section 6 provides conclusion of entire work.

2 Related Work

Studies centered around the issue of feature selection have been discussed here.
Mwangi et al. [19] in their review emphasized on the importance of feature reduction before analyzation of data for useful patterns. Thereafter, Goyal and Parveen[7]improvised the method of feature selection for twitter data. Ultimate agenda of their research was to build a simplified model to distinguish spam tweets from useful tweets. Further, Agarwal and Mittal [1] drew an empirical analysis between various methods of feature selection. Whereas, Kamkarhaghighi and Makrehchi [14] in their study came up with a novel approach for feature extraction. The given methodology was named as CTWE or content tree based word embedding. The method dealt well with ambiguity risk of words. Thereby, Zhenf et. al. [26] had explored effect of feature selection on sentiment analysis of Chinese review data. N-char-grams and N-POS-grams had been chosen at first as potential opinionated features, followed by an enhanced document frequency method to select subset of features. Subsequently, Srinivas Mekala [23] investigated impact of dimensionality reduction for applying clustering on documents that were in textual form. While, Gwelo et al. [9] made an attempt to utilize functionality of PCA, approach of dimensionality reduction to overcome the problem of multicollinearity. Moreover, Ji and Shi [13] presented various procedures for selection based on Bayesian model. All those procedures were founded on priors of slab along with spike. The ultimate goal of presented procedures was to elect considerable variables. Followed to that, Jagdhuber et al. [12] attempted to extend two methods of feature selection including genetic algorithm and forward

selection of type greedy. Extension to these two approaches helped controlling total cost.

Research Gap

On the basis of literature, it was identified that process of feature reduction with respect to sentiment analysis still needs attention of research. The major issues that were identified are discussed below:

- Prevailing algorithms of feature selection wipes of irrelevant features but redundant features are somewhat ignored. Only a few algorithms focus on both, but a pairwise correlation has to be performed amongst features rather than correlation of multi-feature. Otherwise stated, features with strong power of discriminatory together are ignored with traditional feature reduction algorithms but weakens in case of individual features. Moreover, high dimensionality of data results in increase of computational expenses. Besides, many different criteria of evaluation for segregation of features is adopted by existing algorithms. There is still need of identifying best criteria for feature selection.

- Hybrid algorithms of feature selection [3] [6] contingent on optimization method at its stage of wrapper evaluation were also proposed. However, methods like hill climbing clogs up to solution of local optima while looking for most viable features.

- Genetic algorithm methods [11] were also proposed for feature selection. Genetic algorithm though solved the problem of local optima but at the same time suffered with over-fitting problem.

Aim of Research

The aim of this research is to structure, implement and assess an algorithm for reduction of feature subset so as to deal with scalability issue resulting efficient memory usage in data with high dimensionality. Goal is to reduce the burden of computation for classification model in accordance with sentiment analysis.

- **Identification of suitable evaluation criteria**
  Various algorithms of feature selection vary on the basis of evaluation criteria they apply. The primary goal of this research is to determine a criteria of evaluation that minimizes the feature size to least possible value. The suggested feature selection approach is intended to remove irrelevant as well as redundant features.

- **To solve problem of over-fitting**
  Another objective underlying this thesis is to prevent the condition of getting stuck to local optima while searching for most prominent features along with the problem of over-fitting suffered by wrapper methods of feature selection.

- **Extracting appropriate features from high dimensional data for efficient memory usage and low computational burden**
  Social media sites generate high dimensional data containing abundance of opinionated text that is beneficial in determining people's opinion regarding different aspects. Working on entire dataset results in wastage of resources like memory as well as cause computational burden over classifier. This ultimately results in problem of scalability. Therefore, the research aims in selection of appropriate features to provide as input to classifier for efficient memory usage and low computational burden.

3 Research Methodology

The process of producing just features is broken down to three major phases for this research. Attributable to capability of genetic algorithm for dodging the local optima, it is implemented on level of feature selection. However, the problem of over-fitting had to be resolved in this case. Therefore, principle component analysis i.e. PCA was implemented at level of dimensionality reduction. Furthermore, redundancy in feature set is taken care with TF-IDF on very first step of feature extraction.

Conclusively, as demonstrated in Figure 1 the foremost milestone involves extraction of features from preprocessed text. Spotted attributes are then passed for dimensionality reduction. Subsequently, the dimensionally reduced features undergo feature selection procedure. Ultimately, competent features were generated reducing the storage requirement.

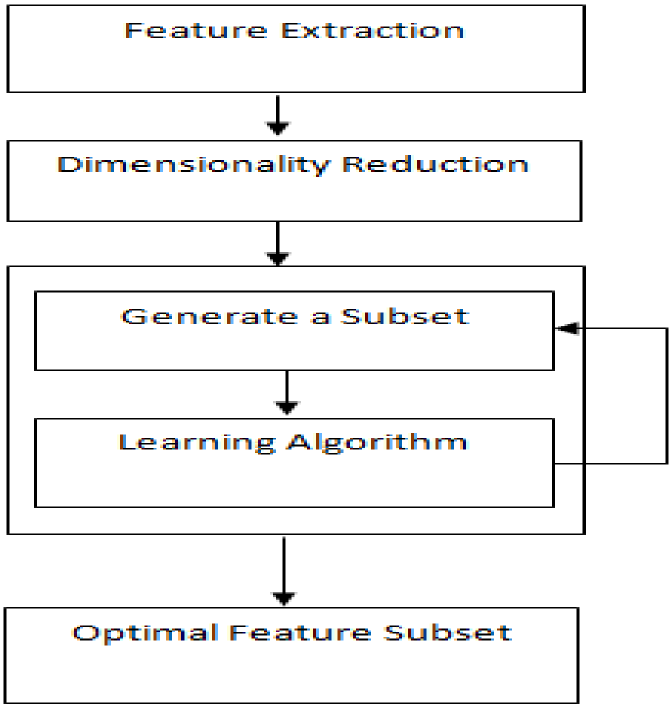

Figure 1: Process of Feature Selection

Feature Extraction

Large volume data has a characteristic that it encompass a wide range of variables demanding lot many resources for processing. Feature extraction is a process that turn these variables into tokens called features. It reduces the magnitude of data under operation in such a way that it still describes original dataset entirely. This way feature extraction proves to be a valuable approach when enormity of feature subset has to be reduced keeping its important features intact. When it comes on reducing attributes holding entire meaning of document to its true form, removal of redundant data is the first possible way of dealing with it. Therefore, a prominent role of feature extraction is to tackle repetitions in data [2]. For this research, feature extraction is used to extract unique words from a sentence and weight them with respect to its frequency of usage.

Dimensionality Reduction

Feature extraction helped pulling out unique features from a huge volume of preprocessed data. The major phenomena for withdrawal of features from such voluminous dataset was to drop all those attributes which were repeating itself. Additionally, each feature was weighed by process of feature extraction according to its occurrence within the document as well as among the documents. Now, we have a dataset that still endures many correlated variables. This signifies that, even now our data to be analyzed holds a lot of bulkiness. Also, it is clearly demonstrated in [15] [24] [5] [16] that as many attributes a dataset contain that much difficulty a machine learning algorithm faces at the time of classification. Besides the performance parameters too come out to be very poor. For this purpose, we propose to incorporate step of dimensionality reduction [21] followed to feature extraction.

Information theory, statistics and machine learning defines dimensionality reduction as an operation of lowering the abundance of arbitrary variables in consideration by acquiring a group of principle variables [25]. Implementation of dimensionality reduction reduces requirement of storage space to a great extent. Additionally, time for training is greatly minimized. Further, the issue of multicollinearity is handled efficiently by abolishing correlated specifics [9]. Also, visual analysis of voluminous data becomes troublesome. Henceforth, by reducing the dimensions of data, dimensionality reduction techniques make it easy to observe patterns by plotting visually.

Feature Selection

Various attributes are consolidated in different forms to generate a dimensionally reduced data. This way the data was shrinked to a great extent but still had some irrelevant information according to our requirement. Therefore, third filtration is applied over features before passing them for classification hence providing a 3-Stage Model for feature selection. Feature selection [20] is simply the process of identifying appropriate variables and pick them for classification. It can be said that, feature selection is mere the inclusion and exclusion of attributes without actually transforming them to come up with something new. Contrary to dimensionality reduction where attributes were combined to provide new dimensionally reduced attributes.

Selection of variables serve 3-fold benefit. [8]. First is to improvise classifier's predictability. Second, it helps in building a cost effective classifier with faster predictions. Number three, it reduces complexity making it easy to understand the process with which data is generated.

4 Algorithm Proposed

The proposed algorithm for feature selection is formulated in this section. Entire process is established in three major steps. Foremost step is the extraction of features. Subsequently, the extracted features are set forth for dimensionality reduction. Final step is the selection of optimal features from dimensionally reduced set of attributes using a wrapper method of genetic algorithm. Algorithm 1 is drawn up to express the proposed algorithm of feature selection. Apart from this, Figure 1 depicts flow of control for selection of features by algorithm proposed.

Algorithm 1 Feature_Select(Preprocessed_data)

```
Input: Preprocessed_data
Output: Selected_Features
tfidfconverter ← TfidfVectorizer()
Extract_features ← tfidfconverter.fit.transform(preprocessed_data).toarray()
X_Scalar ← StandardScalar.fit_transform(Extracted_features)
pca ← PCA(explained_variance_ratio)
X_scalar_pca ← pca.fit(X_scalar).transform(scalar)
X ← X_scalar_pca
y ← preprocessed_data(target)
n_features ← X.Shape[1]
n_gen ← number of generations
Population ← Initialize()
for i 1 to n_gen
    population ← generate(Population)
end for
Selected_Features ← best_chromosomes[-1]
```

Mathematical Formulation

Cross validation mean square error is the basis of fitness function used to assess performance of estimator for distinct sets of variables chosen in each iteration of genetic algorithm so as to select best set of features. The mathematical formulation for same is given below.

Let X be the set of data points, Y be corresponding actual outcomes and $\hat{Y}$ be the predicted outcomes:

$X = \{x_1, x_2, ..., x_n\}$
$Y = \{y_1, y_2, ..., y_n\}$
$\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$

Divide X to k equal subsets of λ

$k_1 = \{x_1, x_2, ..., x_\lambda\}$
$k_2 = \{x_{\lambda+1}, x_{\lambda+2}, ..., x_{2\lambda}\}$
.
.
.
.
$k_n = \{x_{\lambda(k-1)}, x_{\lambda(k-1)+1}, ..., x_{\lambda n}\}$

Validation Set	Training Set	Error
k=1	k=2, k=3,....,K	$MSE_1(\lambda)\ \sum_{i\in k_1 part} y_i - \hat{y}_i$
k=2	k=1, k=3,. ..,K	$MSE_2(\lambda)\ \sum_{i\in k_2 part} y_i - \hat{y}_i$
.		
.		
.		
.		
K	k=1, k=2,....,K-1	$MSE_n(\lambda)\ \sum_{i\in K_{th} part} y_i - \hat{y}_i$

$$CV - MSE = \frac{1}{k}\sum_{k=1}^{K} MSE_k(\lambda)$$

Functionalities

This section puts forward all the functionalities of genetic algorithm that are called in Algorithm 1. Algorithm 2 defines initialization of population for genetic algorithm. Computation of fitness function is manifested in Algorithm 2. While, Algorithm 4 presents procedure of selecting suitable parents for generating fit offsprings. Thereafter, Algorithm 5 and algorithm 6 convey the course of crossover and mutation over individuals elected by selection method. Ultimately, Algorithm 7 is presented that combines up all the functionalities of genetic algorithm for generation of optimum features.

Algorithm 2 Initialize()

```
    Input: Number ofchromosomes in population
    Output: population
Size ← Number of chromosomes in population
for i 1 to Size
    individual ← ones(n_features, d_type = bool)
    mask ← random(len(individual))
    population.append(individual)
end for
return population
```

Algorithm 3 Fitness(population)

```
    Input: population
    Output: Sorted list of scores and population
estimator ← learning algorithm for optimal feature selection
for individual in population
    -1.0 * mean(cross_val_score(estimator, X[:,individual], y, cv = 5, scoring = "neg_mean_square_error"))
end for
sort ← argsort(score)
return list(score[sort]), list(population(sort,:))
```

Algorithm 4 Selection(population_sorted)

```
    Input: population_sorted
    Output: Next_population
n_best ← number of best individuals to select
n_rand ←random individuals to select
for i 1 to n_best
    Next_population.append(population_sorted[i])
end for
for i 1 to n_rand
    Next_population.append(random.choice(population_sorted))
end for
random.shuffle(Next_population)
return Next_population
```

Algorithm 5 Crossover(population)

```
    Input: population
    Output: Next_population
n_children ← number of children created during crossover
for i to $\lfloor len(population/2) \rfloor$
    for j 1 to n_children
        Parent1 ← population[i]
        Parent2 ← population[len(population-1-i)]
        child ← Parent1
        Mask ← random(len(child)) ¿ 0.5cm
        child[Mask] ← Parent2[Mask]
        Next_population.append(Child)
    end for
end for
return Next_population
```

Algorithm 6 Mutation(population)

```
    Input: population
    Output: Next_population
mutation_rate ← probability of individual mutation
for i to len(population)
    individual ← population[i]
    if random.random() < individual.mutation_rate
        mask ← random(len(individual))
        cm individual[mask] ← False
    end if
    Next_population.append(individual)
end for
return Next_population
```

Algorithm 7 Generate(population)

```
    Input: fitness(population)
    Output: population
sorted_scores, sorted_population ← fitness(population)
population ← Selection(sorted_population)
population ← Crossover(population)
population ← Mutation(population)
best_chromosomes.append(sorted_population)
return population
```

5 Result and Interpretations

All results demonstrated in this section were generated on jupyter notebook by execution of code in python environment over three datasets. First being Mixed dataset of movie reviews and news, second being Airline dataset and the final one was dataset of Amazon product reviews.

First step is extraction of unique features followed by inspection of internal data consistency before application of dimensionality reduction technique. For this purpose, Cronbach Alpha test [18] [4] is carried out. It generates a numerical output between range of 0 and 1. If result approaches to 1 then data can be depicted as highly consistent making it reliable. Whereas, if its outcome is near to 0, that indicates unreliable nature of data. For former case, data is assumed to be suitable for carrying out a statistical approach over it. The results of test depicted in Table 2 ensures the adequacy of data for application of PCA.

Table 2: Cronbach Alpha Test Results

	Cronbach Alpha Test Result
Mixed Dataset	0.9822
Airline Dataset	0.9943
Amazon Dataset	0.9981

Subsequent to dimensionality reduction, genetic algorithm with fitness function formulated on basis of 10-fold cross validation is applied for feature selection. Figure 2 to Figure 19 demonstrates convergence of genetic algorithm for all three datasets without getting stuck to local optima with respect to cross validation mean square error. Finally, Table 3 shows reduction in feature subspace after application of proposed model.

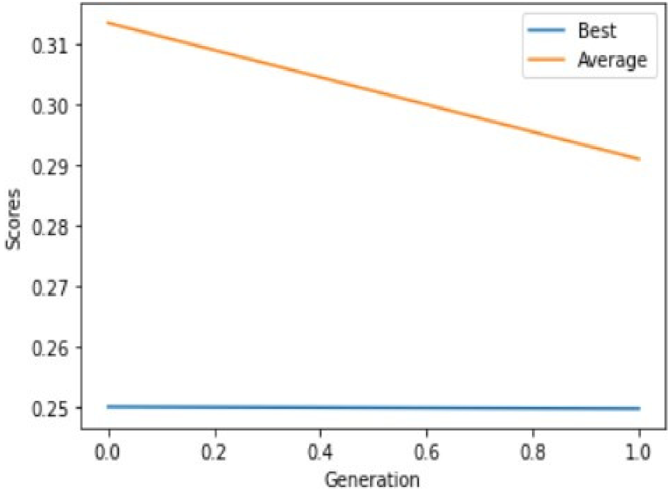

Figure 2: Output of Genetic Algorithm for Two Iterations on mixed dataset

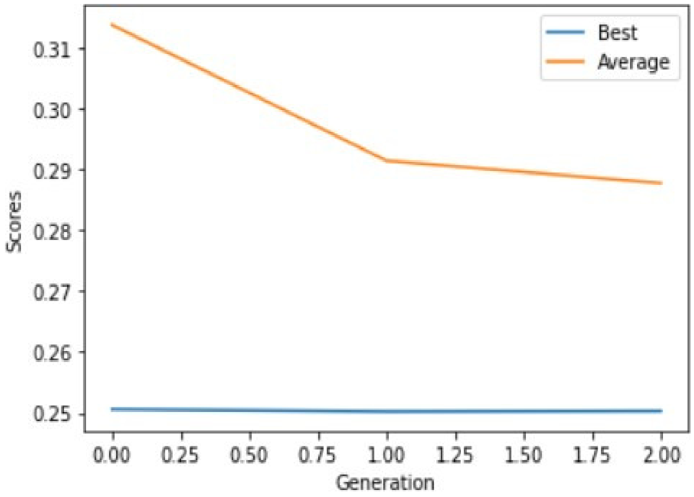

Figure 3: Output of Genetic Algorithm for Three Iterations on mixed dataset

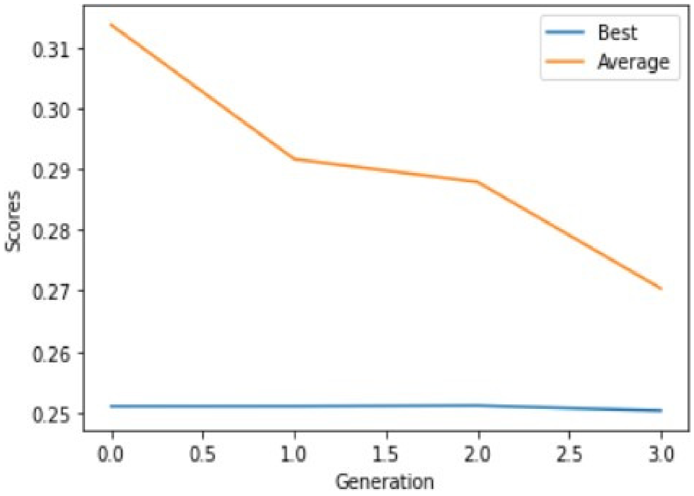

Figure 4: Output of Genetic Algorithm for Four Iterations on mixed dataset

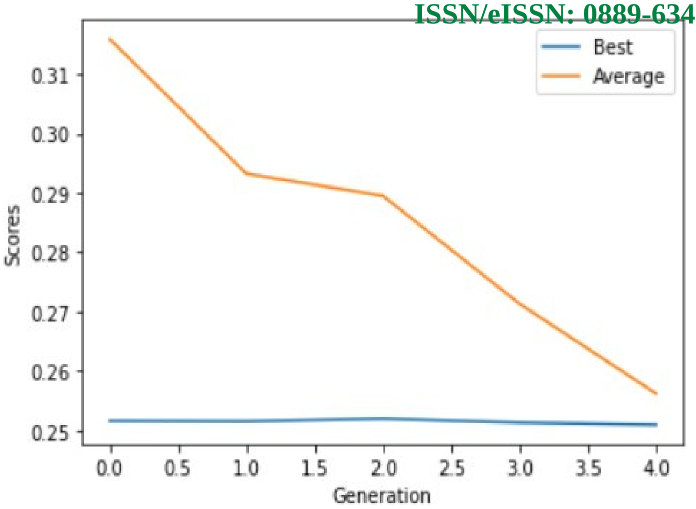

Figure 5: Output of Genetic Algorithm for Five Iterations on mixed dataset

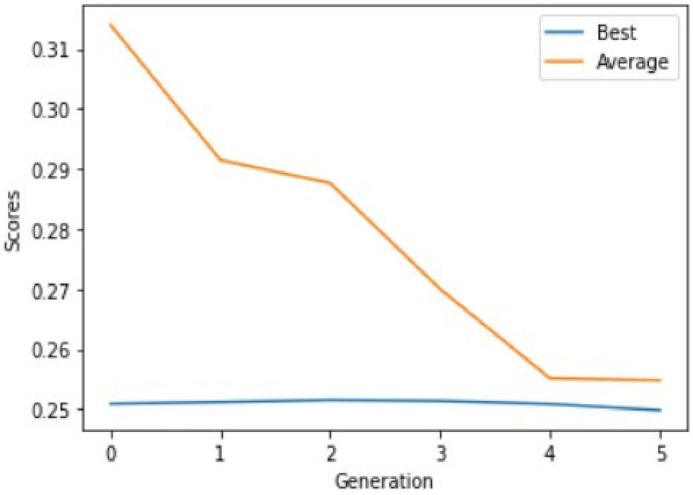

Figure 6: Output of Genetic Algorithm for Six Iterations on mixed dataset

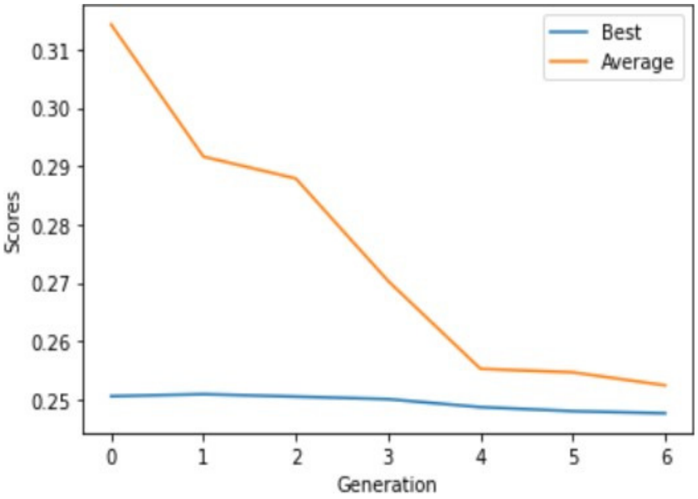

Figure 7: Output of Genetic Algorithm for Seven Iterations on mixed dataset

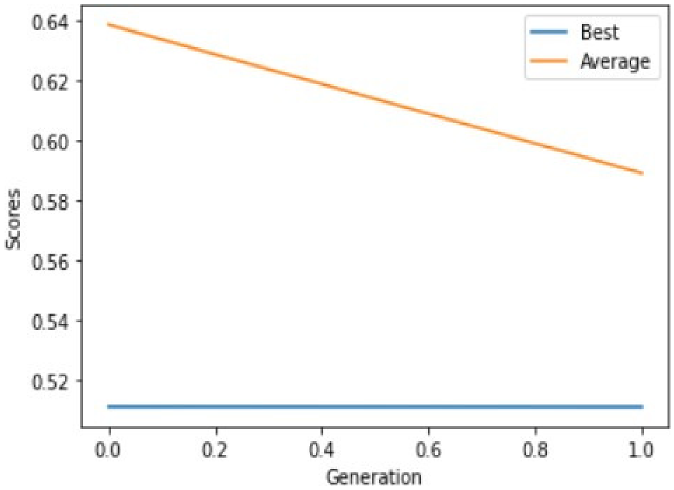

Figure 8: Output of Genetic Algorithm for Two Iterations on airline dataset

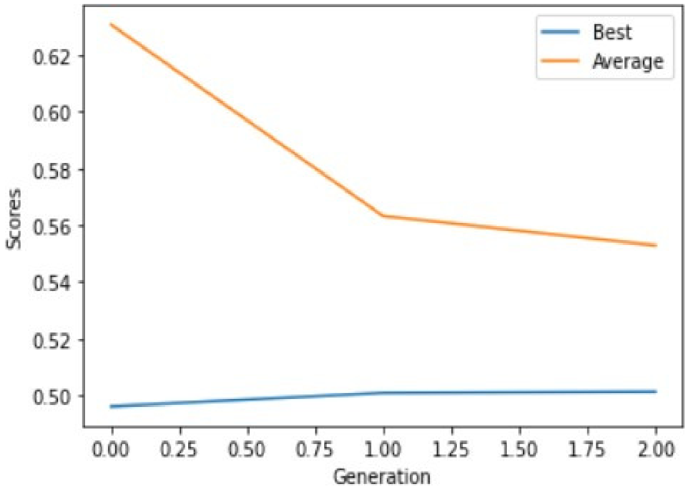

Figure 9: Output of Genetic Algorithm for Three Iterations on airline dataset

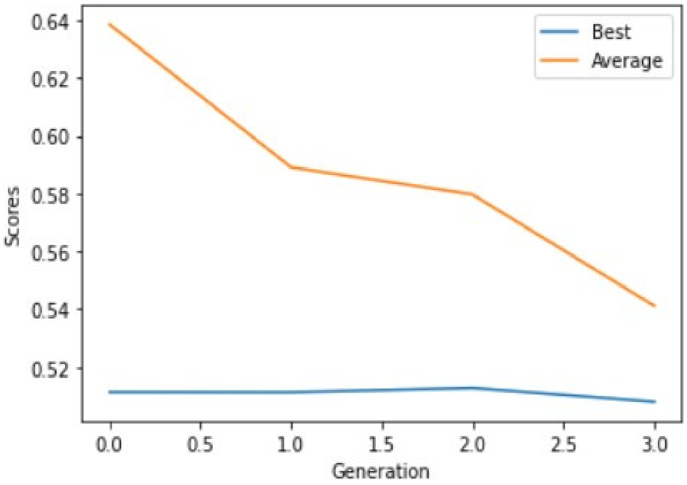

Figure 10: Output of Genetic Algorithm for Four Iterations on airline dataset

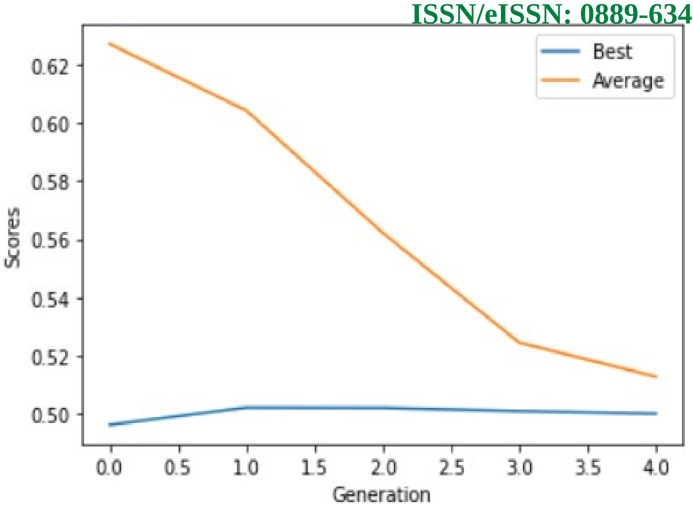

Figure 11: Output of Genetic Algorithm for Five Iterations on airline dataset

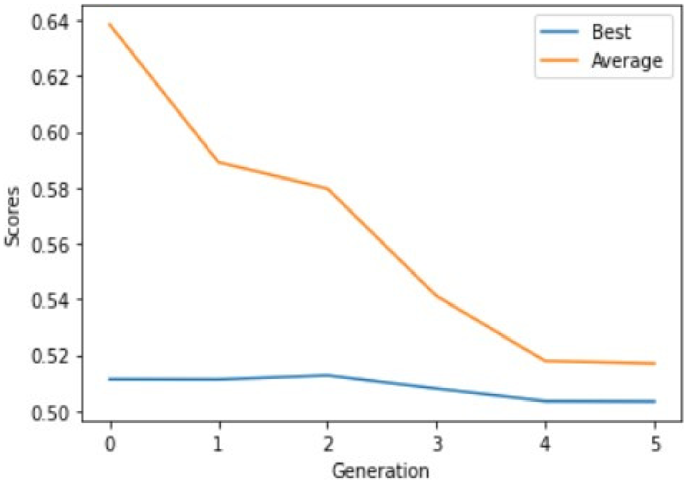

Figure 12: Output of Genetic Algorithm for Six Iterations on airline dataset

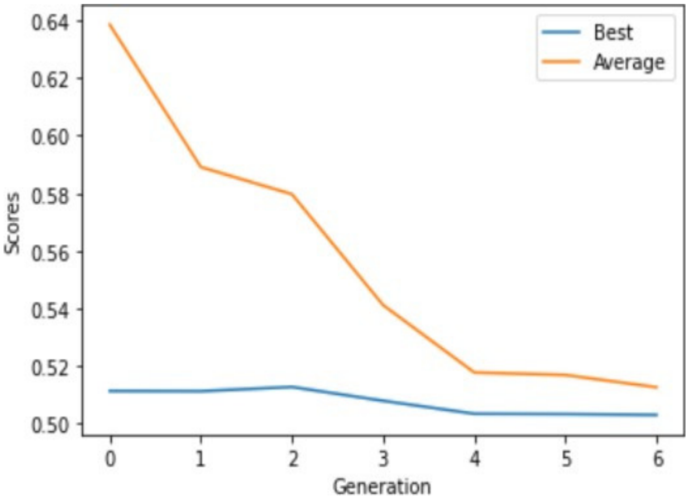

Figure 13: Output of Genetic Algorithm for Seven Iterations on airline dataset

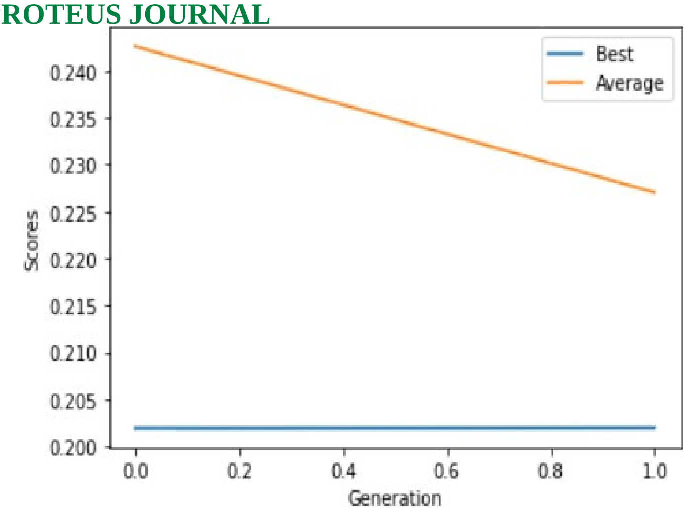

Figure 14: Output of Genetic Algorithm for Two Iterations on amazon dataset

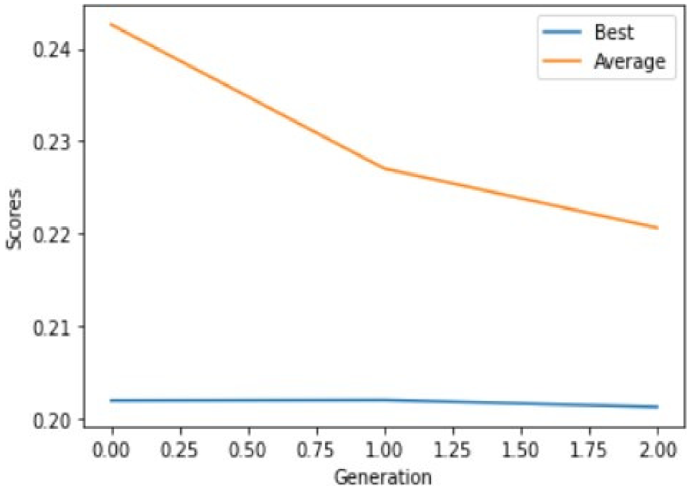

Figure 15: Output of Genetic Algorithm for Three Iterations on amazon dataset

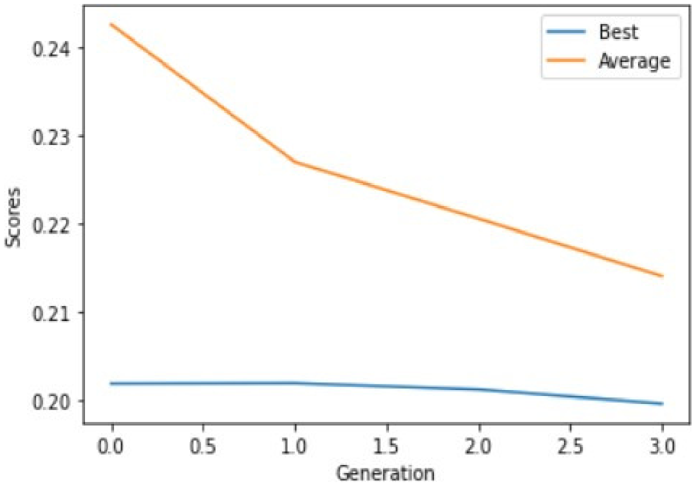

Figure 16: Output of Genetic Algorithm for Four Iterations on amazon dataset

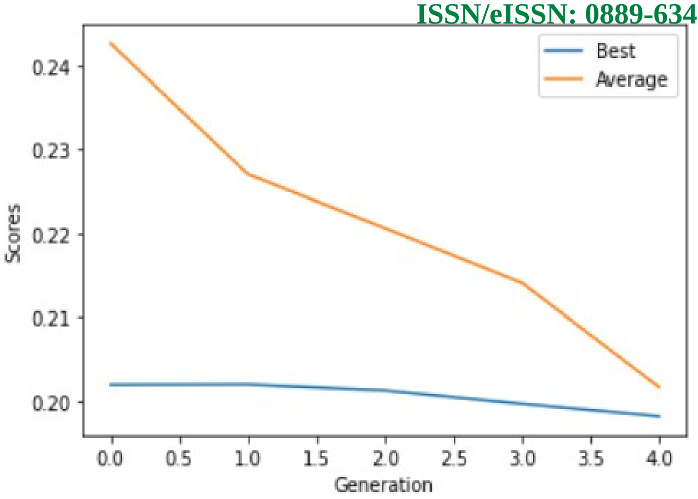

Figure 17: Output of Genetic Algorithm for Five Iterations on amazon dataset

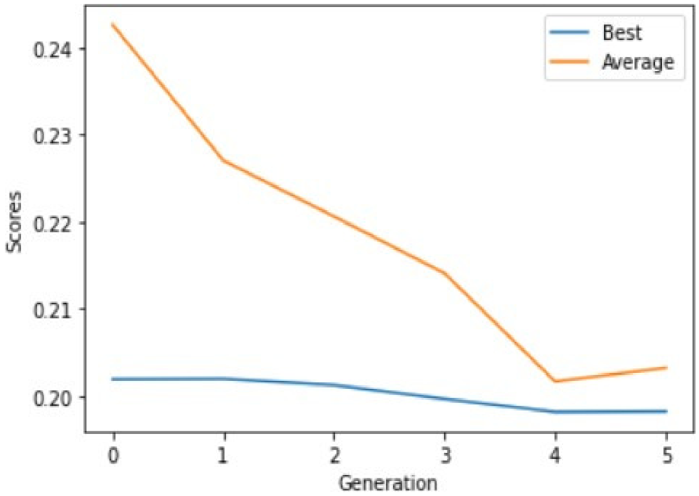

Figure 18: Output of Genetic Algorithm for Six Iterations on amazon dataset

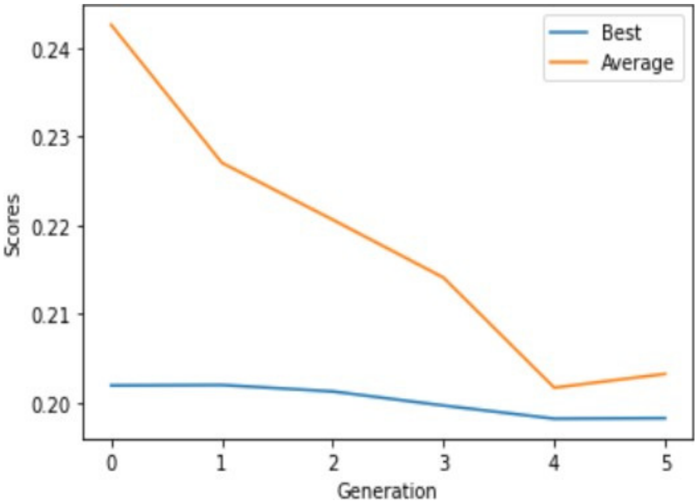

Figure 19: Output of Genetic Algorithm for Seven Iterations on amazon dataset

Table 3: Reduction in Feature Size

	Percentage of Reduction
Mixed Dataset	52.07%
Airline Dataset	45.63%
Amazon Dataset	50.3%

6 Conclusion

This paper presents a three-tier model for selection of optimal features to generate accurate results for mining of opinion from lump of raw data. The model blends advantageous characteristics of feature extraction, dimensionality reduction and feature selection to produce supreme set of features. Feature extraction at one hand converts raw pile of text to tokens while dimensionality reduction deals with the curse of dimensionality and overfitting. Genetic algorithm in the end, elects most prominent feature to predict the polarity of entire text document. Further, in future we will work on implementation of proposed algorithm in connection with sentiment analysis over real time data.

References

[1] Agarwal, B. and Mittal, N. (2016). Prominent feature extraction for review analysis: an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(3):485–498.

[2] Alhaj, Y. A., Al-qaness, M. A., Dahou, A., Elaziz, M. A., Zhao, D., and Xiang, J. (2020). Effects of light stemming on feature extraction and selection for arabic documents classification. In *Recent Advances in NLP: The Case of Arabic Language*, pages 59–79. Springer.

[3] Bermejo, P., Gámez, J. A., and Puerta, J. M. (2011). A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets. *Pattern Recognition Letters*, 32(5):701–711.

[4] Ertando, A., Muflihaini, M., et al. (2020). Validity and reliability test in the questionnaire of javanese local wisdom knowledge aspect through the myth of beringin (ficus sp.). In *Journal of Physics: Conference Series*, volume 1440, page 012067. IOP Publishing.

[5] Fernández, A., del Río, S., Chawla, N. V., and Herrera, F. (2017). An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2):105–120.

[6] Ghosh, M., Kundu, T., Ghosh, D., and Sarkar, R. (2019). Feature selection for facial emotion recognition using late hill-climbing based memetic algorithm. *Multimedia Tools and Applications*, 78(18):25753–25779.

[7] Goyal, S. and Parveen, S. (2015). Improved feature selection for better classification in twitter. *International Journal of Computer Applications*, 122(1).

[8] Guyon, I. and Elisseeff, A. (2003). An introduction to vari- able and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[9] Gwelo, A. S. et al. (2019). Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics*, 4(1):79–91.

[10] InternetLive (2020). Social media data. `https://www.internetlivestats.com/one-second/`. [Online; accessed 20-November-2020].

[11] Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., and Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access*, 7:14637–14652.

[12] Jagdhuber, R., Lang, M., Stenzl, A., Neuhaus, J., and Rahnenführer, J. (2020). Cost-constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms. *BMC bioinformatics*, 21(1):1–21.

[13] Ji, Y. and Shi, H. (2019). Bayesian model selection in order-restricted two-way anova mixed models. *Statistics in Biopharmaceutical Research*, pages 1–11.

[14] Kamkarhaghighi, M. and Makrehchi, M. (2017). Content tree word embedding for document representation. *Expert Systems with Applications*, 90:241–249.

[15] Kwon, O. and Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, 40(5):1847–1857.

[16] Lheureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797.

[17] Li, J. and Liu, H. (2017). Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 32(2):9–15.

[18] Malhotra, N. K. and Birks, D. F. (2007). *Marketing research: An applied approach*. Pearson education.

[19] Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244.

[20] Nafis, N. S. M. and Awang, S. (2020). The impact of preprocessing and feature selection on text classification. In *Advances in Electronics Engineering*, pages 269–280. Springer.

[21] Raunak, V., Gupta, V., and Metze, F. (2019). Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243.

[22] Sohangir, S., Wang, D., Pomeranets, A., and Khoshgoftaar, T. M. (2018). Big data: deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):3.

[23] Srinivas Mekala, D. B. (2019). Kernel pca based dimensionality reduction techniques for preprocessing of telugu text documents for cluster analysis. *International Journal of Engineering Research and Technology*.

[24] Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719.

[25] Xu, X., Liang, T., Zhu, J., Zheng, D., and Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 328:5–15.

[26] Zheng, L., Wang, H., and Gao, S. (2018). Sentimental feature selection for sentiment analysis of chinese online reviews. *International journal of machine learning and cybernetics*, 9(1):75–84.