FEATURE SELECTION USING GENTIC ALGORITHM FOR SENTIMENT ANALYSIS

A PROJECT REPORT

Submitted by

AJAY SUNDAR RAJKUMAR (20BCS006)

BHARATH S (20BCS021)

DHYANESH M (20BCS030)

HARI GOKUL B (20BCS033)

KIRAN KUMAR L (20BCS056)

In partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY

COIMBATORE-641 049

(An Autonomous Institution Affiliated to Anna University, Chennai)

May 2021

KUMARAGURU COLLEGE OF TECHNOLOGY

COIMBATORE 641 049

(An Autonomous Institution Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **“FEATURE SELECTION USING GENTIC ALGORITHM FOR SENTIMENT ANALYSIS”** is the bonafide work of **“AJAY SUNDAR RAJKUMAR (20BCS006), BHARATH S (20BCS021), DHYANESH M (20BCS030) , HARI GOKUL B (20BCS034), KIRAN KUMAR L (20BCS056)”**

who carried out the project work under my supervision.

SIGNATURE

Dr. Devaki. P, Ph.D.,

HEAD OF THE DEPARTMENT

Department of Computer Science and Engineering,

Kumaraguru College of Technology

Coimbatore – 641 049

SIGNATURE

Mr. V. Senthil Kumar, Assistant Professor

SUPERVISOR

Department of Computer Science and Engineering,

Kumaraguru College of Technology

Coimbatore – 641 049

The candidates with University register number **20BCS006, 20BCS021, 20BCS030, 20BCS033, 20BCS056** were examined in the Project Viva-Voice examination held on 31.12.2022.

Internal Examiner

External Examiner

DECLARATION

We affirm that the project work titled **FEATURE SELECTION USING GENTIC ALGORITHM FOR SENTIMENT ANALYSIS** being submitted in partial fulfilment for the award of B.E Computer Science and Engineering is the original work carried out by us. It has not formed the part of any other project work submitted for the award of any degree or diploma, either in this or any other University.

Ajay

AJAY SUNDAR RAJKUMAR (20BCS006)

Bharath

BHARATH S (20BCS021)

Dhyanesh.

DHYANESH M (20BCS030)

Hari Gokul

HARI GOKUL B (20BCS033)

KIRAN KUMAR L (20BCS056)

I certify that the declaration made above by the candidates are true.

Mr. V. Senthil Kumar,

Assistant Professor,

Department of Computer Science and Engineering,

Kumaraguru College of Technology,

Coimbatore – 641 049.

ACKNOWLEDGEMENT

We express our profound gratitude to the management of Kumaraguru College of Technology for providing as with the required infrastructure that enabled us to successfully complete the project.

We extend our gratefulness to our Principal, **Dr. Saravanan D,** for providing us the necessary facilities to pursue the project.

We would like to acknowledge **Dr. P. Devaki,** Professor and Head, Department of Computer Science and Engineering, for her support and encouragement throughout this project.

We thank our project coordinator **Dr.L.Latha,** Professor, Department of Computer Science and Engineering and guide **Mr. V. Senthil Kumar,** Assistant Professor, Department of Computer Science and Engineering, for their constant and continuous effort, guidance and valuable time.

Our sincere and hearty thanks to staff members of Department of Computer Science and Engineering of Kumaraguru College of Technology for their well wishes, timely help and support rendered to us during our project. We are greatly indebted to our family, relatives and friends, without whom life would have not been shaped to this level.

TABLE OF CONTENTS

CHAPTER NO. **TITLE** **PAGE NO.**

ABSTRACT

This paper summarizes work on an approach that combines feature selection and data classification using Genetic Algorithms. First, it describes our use of Genetic Algorithms to optimize classification by searching for an optimal feature weighting, essentially warping the feature space to individuals within groups and to separate groups from one another. This approach has proven especially useful with large data sets where standard feature selection techniques are computationally expensive. Second, it describes our implementation of the approach in a parallel processing environment, giving nearly linear speed-up in processing time. Third, it will summarize our present results in using the technique to discover the relative importance of features in large biological test sets.

1. INTRODUCTION

Now a days, The majority of individuals now communicate with one another through social media. Social networking has become a popular and a helpful tool for people to express their daily feelings in a society. Twitter is one of the social media platforms that most people utilise. According to, Twitter has 300 million monthly active users and 500 million daily tweets. Users of Twitter have the option to upload photographs and brief videos in addition to tweets with up to 280 characters.

When doing sentiment analysis, these tweets provide useful information. Numerous studies have been done on sentiment analysis. Sentiment analysis attracts scholars interested in a variety of topics, including politics, picture sentiment, natural catastrophe sentiment, and rating review prediction.

Numerous features employed or produced by the sentiment cause other issues in this area of study. Accurate classification may be improved by high dimension features. However, because many irrelevant characteristics may be employed throughout the training process, high dimension features could reduce the classification accuracy. To improve accuracy results, a good feature must be used as effectively as possible.

GA has been demonstrated to be a reliable approach for choosing the optimum features for a variety of problems. Use GA to choose the top 78% features from photos for image retrieval. The outcome demonstrates that GA can increase accuracy. By using GA, the best aspects of the network intrusion features were also chosen. GA is used to improve the classification outcome for mechanical fault classification.

1.1 CONCEPTUAL STUDY OF THE PROJECT

Feature subset selection in the context of practical problems such as diagnosis presents a multicriteria optimization problem. The criteria to be optimized include the classification's accuracy, cost, and risk. Evolutionary algorithms offer a particularly attractive approach to multicriteria optimization because they are effective in high-dimensional search spaces.

Genetic Algorithms are search algorithms inspired by Darwin's Theory of Evolution in nature.

- By simulating the process of natural selection, reproduction and mutation, the genetic algorithms can produce high-quality solutions for various problems including search and optimization.
- By the effective use of the Theory of Evolution genetic algorithms are able to surmount problems faced by traditional algorithms.

According to Darwin's theory of evolution, an evolution maintains a population of individuals that vary from each other (variation). Those who are better adapted to their environment have a greater chance of surviving, breeding, and passing their traits to the next generation (survival of the fittest).

First we split the process after preprocessing steps into two stages: training and testing. In the training stage, we create the model. In the testing stage, the model applied and evaluated using the fitness value to know the effectiveness of the algorithm. To avoid the overfitting, we split the dataset into two datasets: training data and testing data. The training and testing data was chosen randomly. The number of data training is 6443 tweets, while the number of data testing is 500 tweets.

1.2 OBJECTIVES OF THE PROJECT

The project mainly aims to do sentiment analysis of social media comments and other datasets using genetic algorithm (GA). Sentiment analysis becomes one of the topics that attract the researchers with various issues, such as, political, image sentiment, natural disaster sentiment, and rating review prediction. However, there are several challenges behind the research related to this topic. Such as the preprocessing steps and hardly classified the tweets.

1.3 SCOPE OF THE PROJECT

There are various intense forces causing customers to use evaluated data when using social media platforms and microblogging sites. Today, customers throughout the world share their points of view on all kinds of topics through these sources. The massive volume of data created by these customers makes it impossible to analyze such data manually. Therefore, an efficient and intelligent method for evaluating social media data and their divergence needs to be developed. Today, various types of equipment and techniques are available for automatically estimating the classification of sentiments. Sentiment analysis involves determining people's emotions using facial expressions. Sentiment analysis can be performed for any individual based on specific incidents. The present study describes the analysis of an image dataset using CNNs with PCA intended to detect people's sentiments (specifically, whether a person is happy or sad). This process is optimized using a genetic algorithm to get better results. Further, a comparative analysis has been conducted between the different models generated by changing the mutation factor, performing batch normalization, and applying feature reduction using PCA. These steps are carried out across five experiments using the Kaggledataset. The maximum accuracy obtained is 96.984%, which is associated with the Happy and Sad sentiments

1.4 EXISTING SYSTEM

Genetic algorithms find use in various real-world applications. In this segment, we have elaborated on some areas that utilize the genetic algorithms in machine learning.

1. Neural networks

Genetic programming in machine learning finds great applications for neural networks in machine learning. We use it for genetic optimization in neural networks or use cases like inheriting qualities of neurons, neural network pipeline optimization, finding the best fit set of parameters for a given neural network, and others.

2. Data mining and clustering

Data mining and clustering use genetic algorithms to find out the centre point of the clusters with an optimal error rate given to its great searching capability for an optimal value. It is renowned as an unsupervised learning process in machine learning, where we categorize the data based on the characteristics of the data points.

3. Image processing

Image processing tasks, such as image segmentation, are one of the major use cases of genetic optimization. However, genetic algorithms can also be used in different areas of image analysis to resolve complex optimization problems.

2. LITERATURE REVIEW

In this section we discuss the prominent literature surveys being carried out in the area of sentiment analysis and text mining. Our comparison criteria is based on the two fac- tors we discussed before; integration of sentiment analysis approaches in a unified way and a cross-disciplinary application area. We are interested to see how user's opinion and his/her social behavior can be helpful in analyzing the current geopolitical situation and uprising.

Medhat *et al* presented a comprehensive overview of the recently proposed algorithms, enhancements, and appli cations in the area of sentiment analysis. They also dis- cussed the related fields to sentiment analysis e.g., transfer learning, emotion detection, and building resources. They tried to give a full image of the sentiment analysis tech- niques and related fields with brief details. Khan *et al* proposed a rule-based domain-independent method which classifies subjective and objective sentences from reviews and blog comments. SentiWordNet is used to calculate the score and to determine the polarity. They showed that their proposed method is effective and it outperforms ML-based methods with an accuracy of 76.8% at the feedback level and 86.6% at the sentence level. Our proposed approach is aligned with these studies as we are also focusing on ML and lexicon-based methods. However, we are employing GA based optimized feature selection for training ML algorithms. Agarwal *et al* examined sentiment analysis on Twitter data. They introduced POS-specific prior polarity features and explored the use of a tree kernel to obviate the need for tedious feature engineering. Their new features and the tree kernel performed almost at the same level and both outperformed the state-of-the-art baseline techniques. Kouloumpis *et al* investigated the utility of linguistic features for detecting the sentiment of Twitter messages. They evaluated the usefulness of the exist- ing lexical resources as well as the creative language used in microblogging. Devies and Ghahramani presented a language-independent model for sentiment analysis for short text forms e.g., social networks statuses. They used Twitter datasets to model happy and sad sentiments and showed that their system

performed 10% better than Naive Bayes (NB) model. These three papers are employing sentiment analysis
on short-text data i.e., SMS, tweets etc.
Similarly, Pontiki *et al* described the aspect based sentiment analysis. They identified the aspects of given tar- get entities and the sentiment expressed for each aspect.

They used manually annotated reviews of restaurants and laptops as a dataset. Njolstad *et al* proposed, defined, and evaluated four different feature categories composed of 26 article features for sentiment analysis. They used five different ML methods to train sentiment classifier of Norweign financial internet news articles. They achieved classification precision up to 71%. When comparing ML classifiers, they found that J48 yielded the highest performance closely followed by Random Forest (RF). We have also presented a similar comparison in which we compared different classifiers and their accuracy on our system. However, we extended our evaluation by including GA optimized features in comparison.
Govindarajan proposed a hybrid classification method based on integration classification methods using arcing classifier. They analyzed the performance in terms of accuracy. They designed classifier ensemble using NB and GA. They evaluated the effectiveness of ensemble technique for sentiment analysis. Finally, they evaluated the performance under different performance metrics using movie reviews datasets. However, they do not compare the performance of different classifiers and do not provide any optimization for feature size reduction.
As we observe that most of the related work employed independent techniques for sentiment analysis while using few evaluation metrics. Furthermore, they do not provide the user with the freedom to choose different algorithms, classifiers, and optimizations according to customized needs. In contrast, our proposed framework bridges the gap between sentiment analysis and geopolitical intelligence by providing

1) a unified framework having the facility to plug different algorithms, cross-validation, and optimized feature selection
2) a two-dimensional analysis on public opinions in association with political uprisings by combining security and opinion mining.

These days, most people use social media to communicate with each other. Social media has grown as a well-known and useful tool to share their daily sentiments in a society. One of the social media used by most users is Twitter. Based on, the monthly active user of Twitter reaching a total of 300 million with 500 million daily tweets. Twitter allows the user to post tweet up to 280 characters and the ability to upload photos and short videos. These tweets become useful data to do sentiment analysis. Numerous research regards the sentiment analysis increases. Sentiment analysis becomes one of the topics that attract the researchers with various issues, such as, politic , image sentiment , natural disaster sentiment , and rating review prediction . However, there are several challenges behind the research related to this topic. Such as the preprocessing steps and hardly classified the tweets . Other problems arise in this field of study is numerous feature used or generated from the sentiment. High dimension features may increase classification accuracy. However, high dimension features could decrease the classification accuracy since many irrelevant features might be used during the training process. There is a need to use a good feature optimally to enhance accuracy result. GA has been proven as a robust algorithm to select the best features for various issues. For image retrieval, use GA to select the best features from 78 features in images. The result shows that GA able to improve accuracy. GA also used by to select the optimal features of the network intrusion features. For mechanical fault classification, uses GA to enhance the classification result by selecting the best distance based features. A general step performed in sentiment analysis is pre-processing. The common pre-processing steps are as follows: removing URL, removing duplicate lines, removing the unrelated word, twitter name, similar lines, punctuation, number, irrelevant tweets, hashtag, check the slang word and change it to the formal one. Another pre-processing step uses the Term frequency-inverse document frequency (TFIDF), stemming, stop words elimination, tokenization, and feature selection using WEKA's filter. Pre-processing step is crucial since it could help increase the accuracy. Nonetheless, many previous studies did not consider pre-processing step thoroughly. NB is a common algorithm that used by researchers nowadays to classify the sentiment of social media. Parveen and Pandey [16] implemented NB on movie dataset. Masrani and Poornalatha use modified NB to analysis the Twitter sentiment on various topics. Goel, Gautam, and Kumar use NB to classify the movie review and produce 58.4% accuracy. Based on the information described above, Naïve Bayes (NB) is used as the classification algorithms with the unigram language model in this study. Pre-processing phase in this study involving 14 steps such as remove the irrelevant tweets, change the contractions, remove URL, number, hashtags, stop words, punctuation, non-English words, non-alphabetic characters, the words smaller than two characters, features with extremely low frequency, change all tweets to lowercase, and use lemmatization. Then, in this research, the Variable Length Chromosome GA (VLCGA) is used to determine types of sentiment in a tweet based on the features examination. The topic of this research is tweets related to self-driving cars.

3. PROBLEM DEFINITION

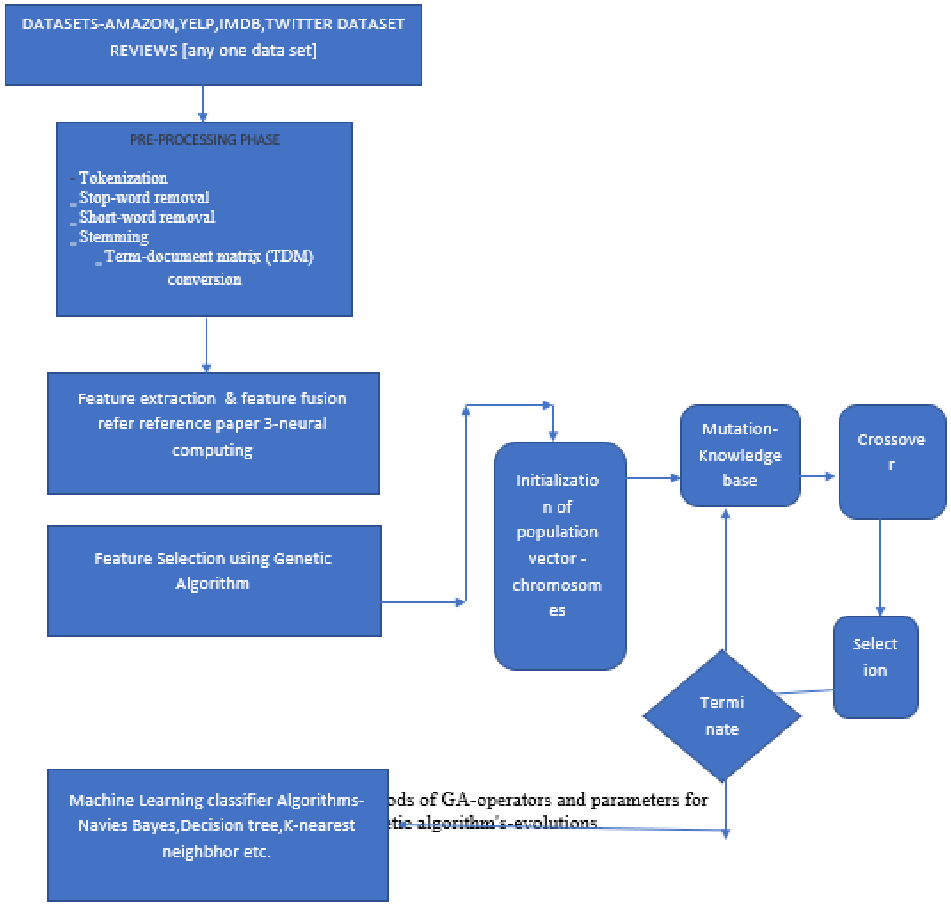
With the advent of online social networks, people became more eager to express and share their opinions and sentiment about all kinds of targets. The overwhelming amount of opinion texts soon attracted the interest of many entities (industry, e-commerce, celebrities, etc.) that were interested in analyzing the sentiment people express about what they produce or communicate. This interest has led to the surge of the sentiment analysis (SA) field. One of the most studied subfields of SA is polarity detection, which is the problem of classifying a text as positive, negative, or neutral. This classification problem is difficult to solve automatically, and many hand-adjusted resources are needed to overcome the difficulties in detecting sentiment from text. These resources include hand-adjusted textual features as well as lexicons. Deciding which resource and which combination of resources are more appropriate to a given scenario is a time-consuming trial-and-error process. Thus, in this work, we propose the use of genetic algorithm(GA) as a tool for automatically choosing, combining, and classifying sentiment from text. We propose a series of functions that allow GA to deal with preprocessing tasks, handcrafted features, and automatic weighting of lexicons for a given training set. We try to prove that the GA solution is better than SVM and superior to naïve Bayes, logistic regression.

4. PROPOSED SYSTEM

In this paper we have developed an algorithm in which we perform Sentiment Analysis using genetic algorithm. This can be done by pre-processing the data, training the data and then performing the genetic algorithm

4.1. BLOCK DIAGRAM

WORKFLOW SYSTEM


DATASETS-AMAZON,YELP,IMDB,TWITTER DATASET REVIEWS [any one data set]
PRE-PROCESSING PHASE
- Tokenization
_ Stop-word removal
_ Short-word removal
_ Stemming
_ Term-document matrix (TDM) conversion
Feature extraction & feature fusion refer reference paper 3-neural computing
Feature Selection using Genetic Algorithm
Initialization of population vector - chromosomes
Mutation-Knowledge base
Crossover
Selection
Terminate
Machine Learning classifier Algorithms-Navies Bayes,Decision tree,K-nearest neighbhor etc.
ods of GA-operators and parameters for
etic algorithm's-evolutions

Modules:

1). Pre-processing

2). Feature extraction

3). Feature selection

4). Genetic algorithm

4.2. METHODOLOGY

A. DATA CLEANING

Data cleaning is the first module in the processing pipeline of this framework. In this phase, extracted data is streamed from the files and saved in the memory for cleaning purpose. This stage consists of three sub-stages.

1) GARBAGE REMOVAL

In this step, unwanted characters (non-ASCII characters) including URLs, web addresses, and online links are removed from the text using customized regular expressions.

2) SLANG CORRECTION

This step involves correcting any slang and abbreviated word that is used in online conversations. We use predefined dictionaries and maps to translate slangs or abbreviation to their original and abbreviated form. e.g. “ttyl” to “talk to you later” and “afk” to “away from keyboard”. This is helpful for later stages because, during sentiment analysis, the abbreviated words make no sense for analysis engine. The working of this module is explained in Algorithm 1.

3) STOPWORD REMOVAL

Stopword removal removes very common words of a language e.g., ‘‘an’’, ‘‘about’’, ‘‘above’’ etc. These words usually have no impact on NLP. We use CMU’s Rainbow stopword list for finding any stopword in the data.

B. PREPROCESSING

This module includes different NLP tasks i.e., tokenization, word stemming, and part-of-speech tagging.

1) TOKENIZATION

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. In order to tokenize the text, LingPipeTokenizer from Apache Lucene package [26] is used which preserves punctuations. Initially, we used StringTokenizer but due to the inherent limitations of this tokenizer, we opted for much better LingPipeTokenizer. An important point to mention is that custom data structures are designed to hold tokens (Keyword) and sentences (list of Keywords) of each document.

2) STEMMING

Stemming is the process of reducing inflected word to its base or root word. The framework use porter-2 algorithm [27] to convert each token to its stem form and store in the Key- word object alongside the original token.

C. ANALYSIS ENGINE

This is the most vital module of the framework. It includes all the natural language based techniques for sentiment analysis. Each sentence (list of Keywords) is fed to the analysis engine and it produces the aggregated sentiment polarity score of the sentence based on different sentiment analysis techniques including lexicon-based, ML using bag-of- words as features, and hybrid approach with feature reduction using GA.

1) ML USING BAG-OF-WORDS AS FEATURES

In this approach, we mainly use different ML algorithms to classify sentiment values of given data.

Weka toolkit is used because it contains several classifiers algorithms and its richness in terms of the analysis. We start by modeling our preprocessed data for Weka classifiers. In order to model the preprocessed data (list of tokens) and generated feature vector, we employ a bag-of-words approach. This is a basic approach in which we include all the potential keywords in the feature vector. We start by reading each document and add its keywords in a feature- set. Then, we append sentiment value associated with that document as a class label and generate an ARFF file. Finally, we process this ARFF file in Weka toolkit and run prominent classifier algorithms including J48, NB, PART, Sequential Minimal Optimization (SMO), Instance-Based with k-nearest neighbours (IB-k), and JRip. Here is a description of these classifiers.

- J48: J48 is a decision tree classifier in which an attribute is selected based on information gain from the training data to build each node of the tree. The selected attributes effectively split a set of training data into subsets enriched in one class or the other. It is mostly used because of its simplicity in explanation and interpretation.

- NB: It is a classification technique based on Bayes Theorem. It works with the assumption that all the attributes on the training samples are independent. It is fast and can be used with the small amount of training data. Although it is very simple, it has outperformed many sophisticated classification methods.

2) HYBRID METHOD WITH OPTIMAL FEATURE SELECTION In this approach, we use ML algorithms to classify sentiment values of the given data. However, the problem with the previous bag-of-words approach is that it does not scale well since almost 80% of the input data gets included in the feature-set.

This problem worsens as the size of the dataset grows bigger. In order to solve this scalability problem, we have devised an efficient technique to reduce the feature-set size.

We propose an evolutionary Genetic Algorithm based approach to evaluate each document and instead of choosing all the keywords, choose a subset of keywords such that the discarded keywords do not impact the overall sentiment score of the document. In other words, we aim to reduce the feature-set size by extracting those keywords that contribute towards the sentiment score of the entire document while excluded keywords make no effect. Once the feature selection is optimized, we use this feature-set to generate ARFF file and consequently perform the analysis using ML classifiers.

Definition 2 (Chromosome (Genotype)): A set of parameters which define a proposed solution to the problem that the GA is trying to solve. A chromosome represents a candidate solution. □

Definition 3 (Population): A set of chromosomes (candidate solutions) that evolves towards a better solution over the certain generations in order to solve the problem. Different genetic operators e.g., mutation, crossover are applied to a population. □

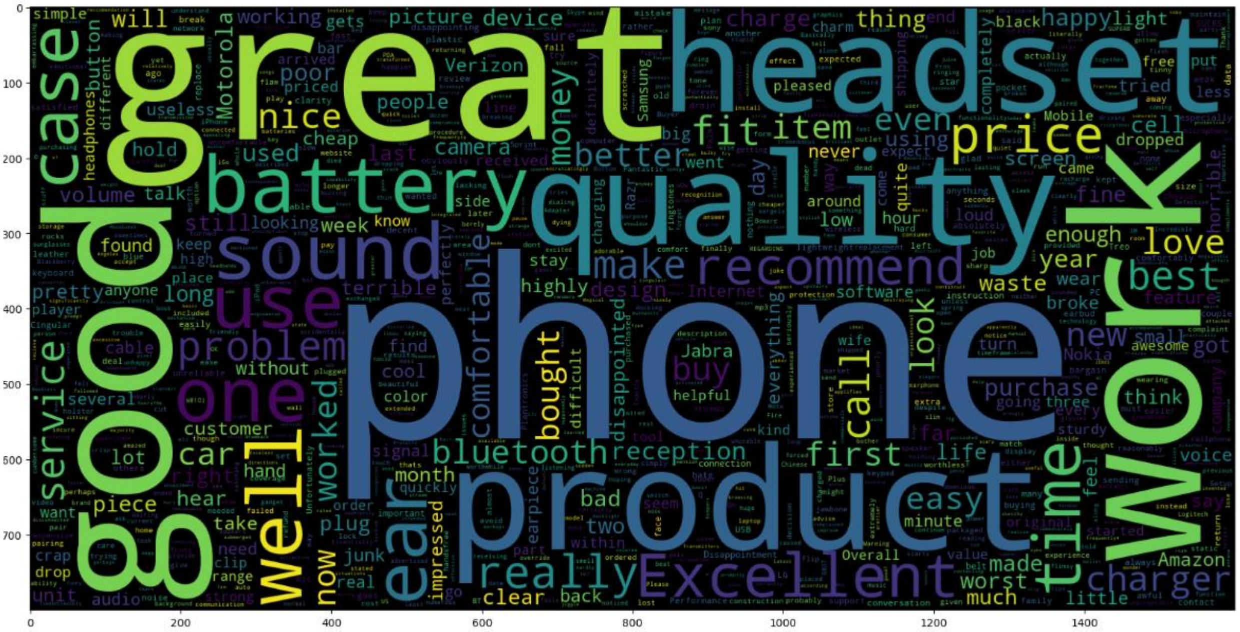
Definition 4 (Fitness Function): The core of an evolutionary algorithm. It is a particular type of objective function which is responsible for performing the evaluation and returning a ‘‘fitness value’’ that reflects how optimal the solution is. The fitness value is used to determine which candidate solution (chromosome) will be surviving in the next generation.

D. FEATURE OPTIMIZATION

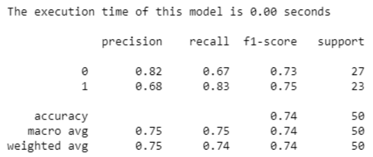
Extracting features using bag-of-words data structures results in significantly large feature vector size because all the keywords which have any associated sentiment value are included in the feature vector. This technique, however, poses significant scalability problem when using a larger dataset. To solve this problem, we need to optimize the feature vector by reducing its size while maintaining accuracy. In this section, we formulated this problem and proposed its solution by using evolutionary Genetic Algorithmic approach.

4.3. IMPLEMENTATION

Positive words cloud of words



Bernoulli Naïve-Bayes (Confusion Matrix and ROC-AUC curve)

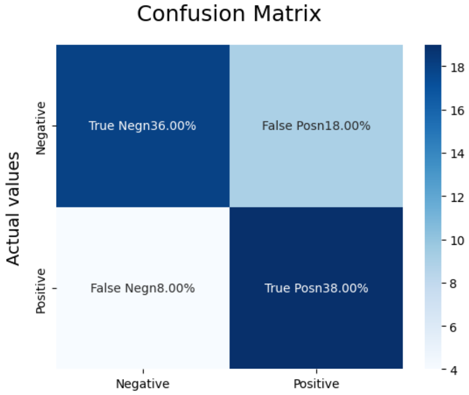


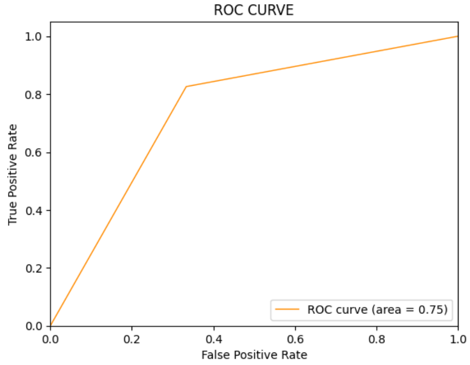
```
The execution time of this model is 0.00 seconds

              precision    recall  f1-score   support

           0       0.82      0.67      0.73        27
           1       0.68      0.83      0.75        23

    accuracy                           0.74        50
   macro avg       0.75      0.75      0.74        50
weighted avg       0.75      0.74      0.74        50
```

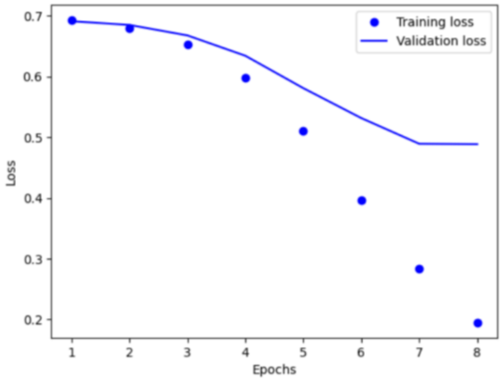


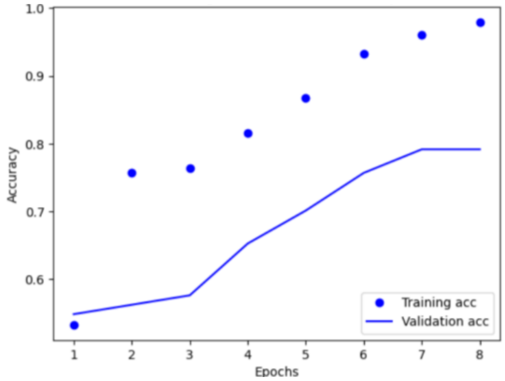


Feature selection using Naïve Bayes algorithm:

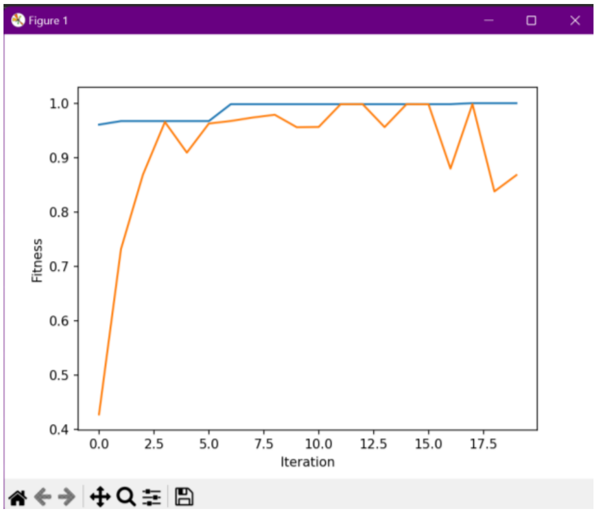
Training and validation loss:

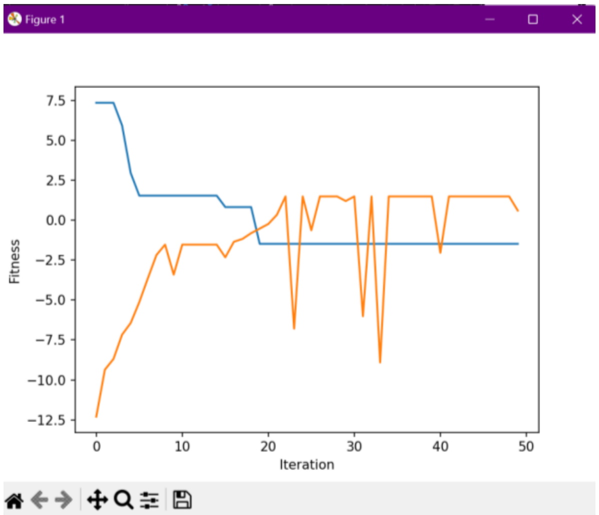
Training and validation accuracy:





Feature selection using Genetic algorithm:





5. SYSTEM REQUIREMENTS

HARDWARE REQUIREMENT

Following are the hardware requirement that is most important for the project:

a) Fluently working Laptops

b) RAM minimum 4Gb

SOFTWARE USED:

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter.

6. CONCLUSION

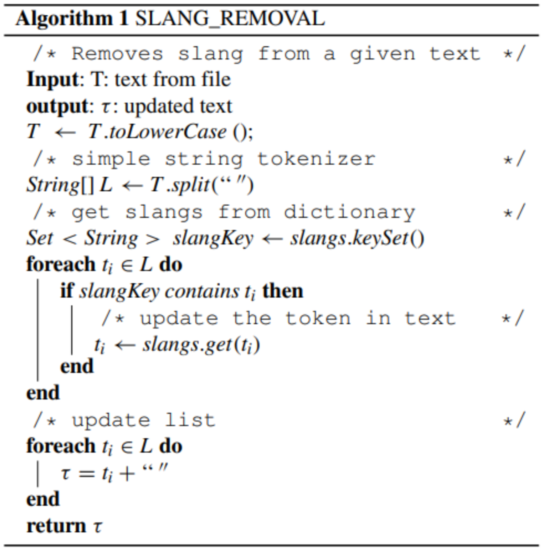
In this paper, we have presented the design, development, and evaluation of our integrated sentiment analysis frame- work in detail. We have employed a single approach to sentiment analysis which is GA with optimized feature selection. We proposed and developed an evolutionary model for feature selection using GA's evolutionary model. This novel approach resulted in 36% - 42% reduced feature size and about 5% increased efficiency as compared to a normal ML approach. We also presented a detailed evaluation of these approaches with respect to different datasets. Furthermore, our detailed analy- sis of different ML classifiers revealed that the NB classifier has the highest accuracy (about 80%) while using our GA based optimal feature selection on Twitter and reviews dataset while in case of the geopolitical dataset, IB-k outperformed all the classifiers with the accuracy of 95%.

Furthermore, we evaluated our proposed technique for scalability by using execution time comparison. We found that our system showed a linear speedup with the increased dataset size. Although, the time spent in the selection of optimal feature-set using GA took about 60% to 70% of the total execution time on reviews dataset, however, it still remained linear and produced a feature-set with 40% reduced size than the original feature-set. GA based feature set results in a speedup of modeling the classifiers up to 55%

In order to demonstrate the benefit of using our feature reduction algorithm over other feature reduction techniques, we have provided an accuracy comparison of GA based hybrid approach with PCA and LSA. The results showed that our GA based feature reduction showed up to 15.4% increased accuracy over PCA and up to 40.2% increased accuracy over LSA. This strengthens our claim that our pro- posed algorithm is fast, accurate, and scales well as the dataset grows bigger.

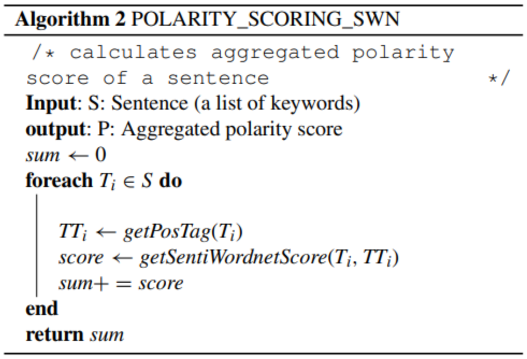
We conclude that our sentiment analysis framework has proved to be a great addition in the discipline of opinion mining. It provided the flexibility of choosing among three widely used sentiment analysis techniques according to cus- tom needs. With additional benefits of GA based optimiza- tion, it reduces feature size and improves efficiency while maintaining the scalability. In the future, we aim to extend this framework for cyber-intelligence so that it would help gener- ate recommendations for law-enforcement agencies based on user opinions.

7.APPENDIX



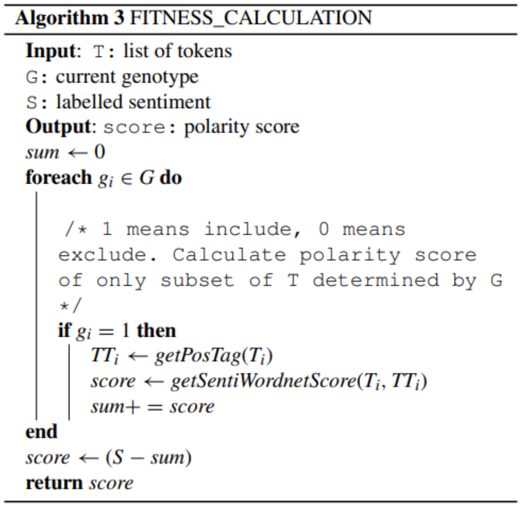
Algorithm 1 SLANG_REMOVAL

```
/* Removes slang from a given text */
Input: T: text from file
output: $\tau$: updated text
$T \leftarrow T.toLowerCase$ ();
/* simple string tokenizer */
$String[]\ L \leftarrow T.split(“ ”)$
/* get slangs from dictionary */
$Set < String >\ slangKey \leftarrow slangs.keySet()$
foreach $t_i \in L$ do
    if $slangKey\ contains\ t_i$ then
        /* update the token in text */
        $t_i \leftarrow slangs.get(t_i)$
    end
end
/* update list */
foreach $t_i \in L$ do
    $\tau = t_i + “ ”$
end
return $\tau$
```



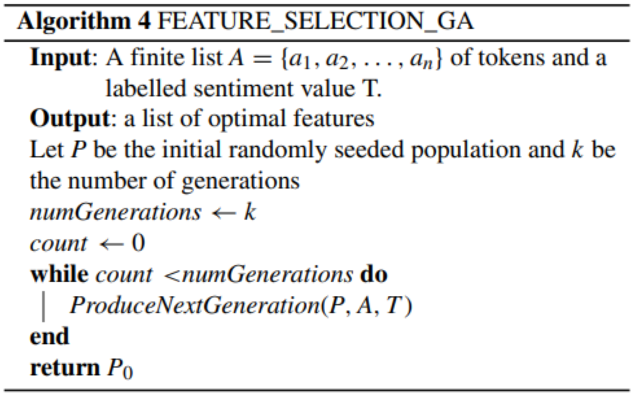
Algorithm 2 POLARITY_SCORING_SWN

```
/* calculates aggregated polarity
score of a sentence */
Input: S: Sentence (a list of keywords)
output: P: Aggregated polarity score
$sum \leftarrow 0$
foreach $T_i \in S$ do
    $TT_i \leftarrow getPosTag(T_i)$
    $score \leftarrow getSentiWordnetScore(T_i, TT_i)$
    $sum+ = score$
end
return $sum$
```



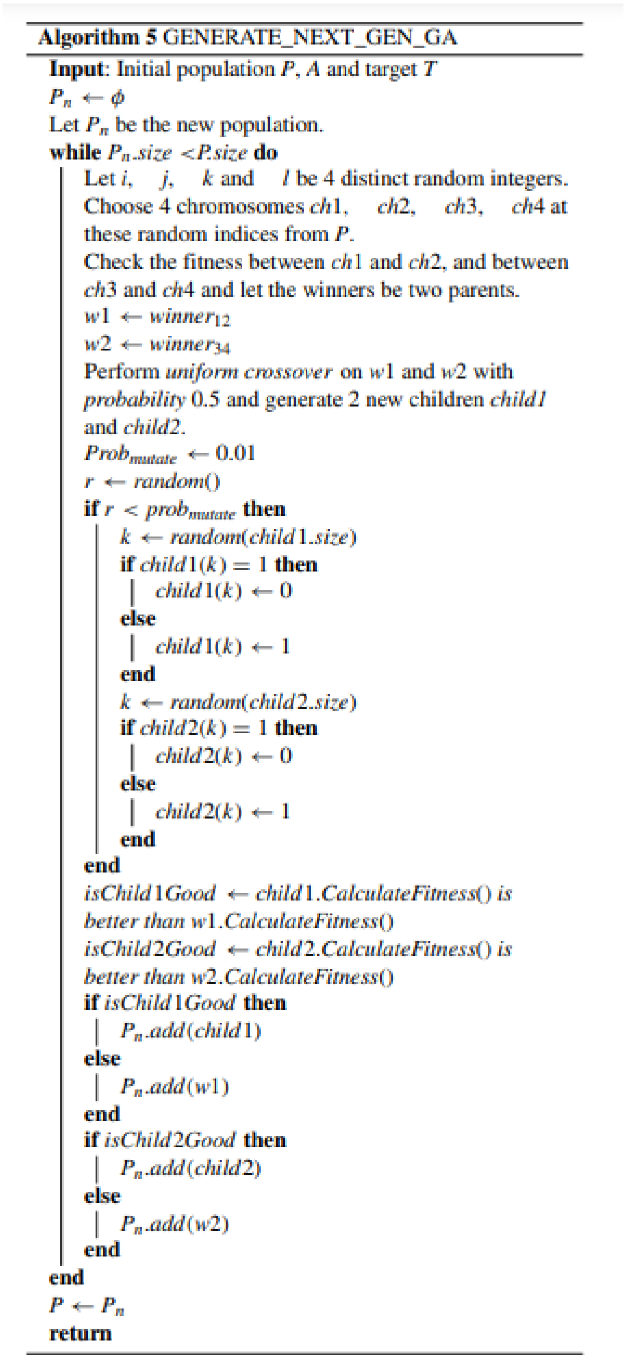
Algorithm 3 FITNESS_CALCULATION

```
Input: T: list of tokens
G: current genotype
S: labelled sentiment
Output: score: polarity score
$sum \leftarrow 0$
foreach $g_i \in G$ do
    /* 1 means include, 0 means
    exclude. Calculate polarity score
    of only subset of T determined by G
    */
    if $g_i = 1$ then
        $TT_i \leftarrow getPosTag(T_i)$
        $score \leftarrow getSentiWordnetScore(T_i, TT_i)$
        $sum+ = score$
end
$score \leftarrow (S - sum)$
return $score$
```



Algorithm 4 FEATURE_SELECTION_GA

```
Input: A finite list $A = \{a_1, a_2, \ldots, a_n\}$ of tokens and a
        labelled sentiment value T.
Output: a list of optimal features
Let $P$ be the initial randomly seeded population and $k$ be
the number of generations
$numGenerations \leftarrow k$
$count \leftarrow 0$
while $count\ < numGenerations$ do
    $ProduceNextGeneration(P, A, T)$
end
return $P_0$
```



```
Algorithm 5 GENERATE_NEXT_GEN_GA
  Input: Initial population $P$, $A$ and target $T$
  $P_n \leftarrow \phi$
  Let $P_n$ be the new population.
  while $P_n.size < P.size$ do
      Let $i$, $j$, $k$ and $l$ be 4 distinct random integers.
      Choose 4 chromosomes $ch1$, $ch2$, $ch3$, $ch4$ at these random indices from $P$.
      Check the fitness between $ch1$ and $ch2$, and between $ch3$ and $ch4$ and let the winners be two parents.
      $w1 \leftarrow winner_{12}$
      $w2 \leftarrow winner_{34}$
      Perform *uniform crossover* on $w1$ and $w2$ with *probability* 0.5 and generate 2 new children *child1* and *child2*.
      $Prob_{mutate} \leftarrow 0.01$
      $r \leftarrow random()$
      if $r < prob_{mutate}$ then
          $k \leftarrow random(child1.size)$
          if $child1(k) = 1$ then
              $child1(k) \leftarrow 0$
          else
              $child1(k) \leftarrow 1$
          end
          $k \leftarrow random(child2.size)$
          if $child2(k) = 1$ then
              $child2(k) \leftarrow 0$
          else
              $child2(k) \leftarrow 1$
          end
      end
      $isChild1Good \leftarrow child1.CalculateFitness()$ is better than $w1.CalculateFitness()$
      $isChild2Good \leftarrow child2.CalculateFitness()$ is better than $w2.CalculateFitness()$
      if $isChild1Good$ then
          $P_n.add(child1)$
      else
          $P_n.add(w1)$
      end
      if $isChild2Good$ then
          $P_n.add(child2)$
      else
          $P_n.add(w2)$
      end
  end
  $P \leftarrow P_n$
  return
```

REFERENCES

1. https://ieeexplore.ieee.org/abstract/document/8620527
2. http://www.ijstr.org/final-print/may2020/A-Literature-Survey-On-Sentiment-Analysis-Techniques-Involving-Social-Media-And-Online-Platforms.pdf
3. https://www.researchgate.net/publication/236203597_Sentiment_Analysis_A_Literature_Survey
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4923467/
5. https://towardsdatascience.com/feature-selection-with-genetic-algorithms-7dd7e02dd237#:~:text=Genetic%20algorithms%20use%20an%20approach,model%20for%20the%20target%20task.
6. https://github.com/kaushalshetty/FeatureSelectionGA/blob/master/feature_selection_ga/feature_selection_ga.py
7. https://github.com/dawidkopczyk/genetic
8. https://github.com/scoliann/GeneticAlgorithmFeatureSelection
9. https://github.com/JavierMtz5/ArtificialIntelligence/blob/main/GA_feature_selection/GA_feature_selection.ipynb