Report - 21/11/2022

**SC with GA**

Process Involved

horizontal line

# Introduction

Nowadays, people are highly interested in buying online products from any e-commerce site. The quality of the products is mainly based on customer reviews in the text that has been posted on e-commerce platforms. While there are an abundant number of reviews posted by customers, it is impossible for customers to read all the reviews to determine the likelihood, support, and quality of the product from the customers’ perspective. Currently, sentiment analysis using machine learning techniques is the fastest-growing field to accurately predict whether the posted text reviews symbolize positive or negative perspectives. Based on this identified review, the rating of the products is evaluated from one to five stars. However, scalability, overfitting, and underfitting are critical concern that arises during the training of machine learning data. In order to reduce the above concerns significant features are selected to train the machine learning model.

Hence, various approaches are used to select the features from the training data, and genetic algorithms select optimal features due to their searching capability. Furthermore, the proposed work excellency is compared with the existing sentiment analysis model in order to find the proposed work efficiency using the metrics namely precision, recall, F-measure, and accuracy.

## A) Data Preprocessing

The dataset used is the Amazon dataset, which has 1000 entries each with 2 attributes, namely the command and the score (1000x2). The commands are the review commands received for different products on amazon. And the score is a binary value, with 1 corresponding to the review being positive and 0 to the review being negative for the product.

### Tokenization

Tokenization involves the splitting of large paragraphs and sentences into individual words and symbols. As the dataset we have is just a collection of sentences, here we are performing the word tokenization where the sentences are broken down into words.

For this, we will use the *word\_tokenize* method from the *NLTK* library.

### Stopwords removal

In this step, very common words that have no impact on the sentence like ‘a’, ‘an’, ‘the’, etc.. are removed with the help of NLTK’s stopwords method.

### Lemmatization

Lemmatization is similar to the *Stemming* process, where all the word tokens from the previous step are disintegrated into their respective root word. The only difference between the Stemming and the Lemmatization technique is that the root word reduced by lemmatization guarantees the word to be meaningful, unlike Stemming where sometimes the root word produced might make no sense.

For this, we will use the WordNetLemmatizer class from the *NLTK* library.

### Bag of Words

Bag of Words is a method of feature extraction on text data, that involves transforming the tokenized sentences into a numerical vector of the frequency of every unique word from the whole dataset.

*CountVectorize* from *sklearn.feature\_extration.text* is used in this step

## B) Feature Selection

The Feature selection approach proposed to be is based on an evolutionary algorithm - *The genetic Algorithm.* We can use those selected features from the GA, instead of using all the features from the preprocessed data. So that we can reduce the size of the feature set by selecting only those features that correlate more towards the sentiment score of the sentence.

### Genetic Representation

In the Genetic algorithm, each document or vector is called a *Chromosome* or *genotype.* Each chromosome will be of the same shape as the converted bag of words vector.

### Generating Population

The population of a generation is the collection of multiple chromosomes generated at random in the first generation. Where after generations, the population gets manipulated with respect to their fitness values (*weights*).

### Fitness Function

Fitness function is responsible for setting weight for each chromosome based on the accuracy is gets from the Naive Bayes model itself. Fitness function determines if a chromosome will produce off-springs and survive to the next generation.

### Pair Selection

The pair of chromosomes or in other words, the two parents selected to produce their off-springs are selected based on this pair selection process. We will rank the chromosomes in a population based on the weights and pick the top 100 chromosomes. And in random choose any two parents and make them produce off-springs using *crossover method*.

### Single-Point Crossover

Based on the crossover *probability*, we can choose either to perform crossover on the pair of parents selected. At random an *index* is generated for the chromosome. And at the randomly generated index, the single-point crossover is performed.