**GA-based Feature Selection Method for Sentiment Classification using Machine Learning Classifiers**

**Abstract:**

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.

**Datasets:**

1. Amazon Dataset - 1000 entries

2. YELP Dataset -

3. IMDb movie review Dataset - 1496 entries

4. Twitter Dataset -





















**The values and methods of GA operators and parameters for the genetic algorithm's evolutions.**

**86GA Parameters:**

| No of features | 1553 features (len of the chromosome) (amazon) |
| --- | --- |
| Size of population | 1000 sentences in amazon dataset |
|  |  |
| Chromosome representations | bitwise representation |
| Population selection | filtering the top 100 best-fitted chromosomes |
| Parent selection | random selection |
| Crossover function | choosing a random crossover-point and exchanging the genes of the parent's |
| Fitness function | based on the accuracy from the naive Bayes model |
| Type of crossover | single point crossover |
| Type of Mutation | Bit Flip mutation |
|  |  |
| Crossover probability | (0.1 to 0.9) |
| Mutation probability | (0.1 to 0.9 |
|  |  |
| Maximum No. generations | For now for testing, is set to 10000 |
| No of runs | min. 30 runs  (As GA is a stochastic algorithm, to get a reliable measure, we run multiple times and average the results) |

