Twitter Data Feature Selection Using Enhanced Genetic Algorithm

[1] K. Brindha, [2] Dr. E. Ramadevi

[1]Research Scholar, NGM College, Pollachi,Tamilnadu, India

[2]Associate Professor, Dept of Computer Science,NGM College, Pollachi,Tamilnadu, India

Abstract

Feature selection is a basic critical task in sentiment analysis, especially while analyzing Twitter data for stock market sentiment. This paper proposes an enhanced genetic algorithm (GA) for feature selection utilizing Finance Yahoo stocks data and openly accessible Twitter data. The objective is to distinguish the most relevant features that can successfully anticipate stock market sentiment. The proposed GA integrates methods to enhance the investigation and double-dealing capacities, empowering it to look through a bigger feature space and work on the nature of chosen features. The algorithm starts by introducing a populace of random binary chromosomes, with every chromosome addressing a feature subset. Wellness assessment is performed utilizing sentiment analysis strategies to survey the prescient force of each feature subset. Trial assessment utilizing Finance Yahoo stocks and Twitter data shows that the enhanced GA beats customary GA and PSO strategies concerning exactness and forecast performance. The proposed approach gives important experiences to sentiment analysis and feature selection with regards to stock market sentiment utilizing Twitter data.

Keywords: *Sentiment analysis, Twitter, Feature selection, Genetic algorithm, Chi Square;*

I. Introduction

The widespread adoption of social networking platforms like Facebook, Twitter, Google, and similar platforms has reshaped the internet, turning it from a static repository into a dynamic arena where information is in constant flux. Within this dynamic online landscape, sentiment analysis plays a pivotal role in assessing public sentiment and enhancing user interfaces (UI). Amidst the myriad of available platforms, Twitter stands out as one of the largest social networks, boasting a user base of 326 million active users worldwide and generating a staggering 500 million tweets on a daily basis. The Twitter Dataset is widely acknowledged as a valuable resource for sentiment analysis due to its simplicity and easy accessibility. By employing a variety of machine learning techniques, sentiment analysis categorizes the emotions conveyed in tweet messages as positive, negative, or neutral.

Machine learning is a method that empowers computers to gain knowledge and make predictions within a given context. Among the numerous machine learning algorithms available, Naïve Bayes is renowned for its effectiveness and extensive use, particularly in classification tasks. The Chi-square algorithm, crafted for feature selection, can be smoothly incorporated with Naïve Bayes. This integration elevates the accuracy and efficiency of classification, surpassing the constraints of prior approaches and delivering an optimal solution for Twitter sentiment analysis.

Twitter, a widely recognized micro-blogging platform, serves as a digital arena where users communicate through concise status messages referred to as tweets. These tweets provide users with a means to convey their personal emotions, share viewpoints, and often incorporate elements such as hashtags, emoticons, or regular words to engage in discussions about various subjects, including events, movies, products, or celebrities. Sentiment Analysis (SA) frequently revolves around the task of categorizing these messages into polarities, typically positive or negative sentiments. Within the domain of sentiment classification, substantial Natural Language Processing (NLP) research has been dedicated to user-generated content. Traditionally, much of this research has revolved around the analysis of substantial bodies of text, such as product and film reviews, which encapsulate the generalized opinions of their authors. Sentiment analysis on Twitter involves the examination of emotions, attitudes, and sentiments conveyed in tweets. Individuals express their feelings using a variety of mediums, such as text, images, and emojis. Many methodologies tend to view emojis and abbreviations in tweets as undesirable noise that should be eliminated. However, it's crucial for sentiment analysis to encompass the assessment of these emojis and abbreviations. In essence, sentiment analysis can be approached in two primary ways: lexicon-based and machine learning-based. The lexicon-based approach can be further subdivided into two categories: corpus-based and dictionary-based. Within the corpus-based approach, two methods stand out: semantic and measurement. Linear classifiers like Support Vector Machine (SVM) and neural networks are part of this category, while probabilistic classifiers include Naïve Bayes, Bayesian Network, and Maximum Entropy.

The rationale behind focusing on feature selection techniques in conjunction with tweet sentiment analysis is twofold. Due to the unique characteristics of tweets, feature engineering methods applied to Twitter data have the potential to generate a substantial number of features. However, each individual tweet contains only a few features from the entire feature set, as tweets are limited to 140 characters. Feature selection techniques play a crucial role in cherry-picking a subset of features, significantly smaller than the total feature pool. This reduction in feature dimensionality not only decreases the computational time required for training and classifying tweets but also enhances classifier performance. This improvement is achieved by eliminating redundant or irrelevant features and mitigating the risk of overfitting.

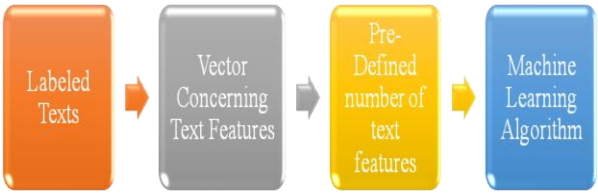
Labeled Texts
Vector Concerning Text Features
Pre-Defined number of text features
Machine Learning Algorithm

Figure 1.Feature Selection Process

Figure 1 depicts the feature selection process. To begin with, the text document is given a name and subsequently converted into a vector comprising a multitude of features derived from various words. This can result in a substantial number of features. Consequently, it becomes imperative to streamline the selection process, preserving only the features with the utmost relevance. Features possessing higher values encompass more substantial information compared to their lower-value counterparts. Following the download of the word document, it can then progress to the data training phase.

The core concept behind various feature selection methods remains consistent. Each algorithm assesses the value of each feature, and important features are chosen based on either their initial value or a predefined threshold for the feature's score. In fact, there are several commonly employed feature selection techniques.

Ii. Existing Methodologies

2.1 Genetic Algorithm

"A Genetic Algorithm-based Feature Selection for Twitter Sentiment Analysis" by I. Tuba et al. (2018). The authors suggest employing a genetic algorithm-driven feature selection method for sentiment analysis on Twitter data. They represent features in a binary format and employ a fitness function to evaluate the quality of feature subsets. The research demonstrates an enhanced performance in sentiment classification compared to alternative feature selection techniques.

2.2 Hybrid Feature Selection

M. Alam (2019) this research presents a hybrid feature selection framework for sentiment analysis on Twitter data. It consolidates a genetic algorithm with a channel based way to deal with select relevant features. The study assesses the performance of the framework involving different assessment measurements and exhibits its viability in further developing sentiment classification accuracy.

2.3 Feature Selection for Twitter Sentiment Analysis

M. Z. Ali et al. (2020) the authors direct a similar study of various feature selection techniques for sentiment analysis on Twitter data. They assess the performance of information gain, chi-square, and term frequency-inverse document frequency (TF-IDF) methods. The study gives experiences into the viability of every technique in further developing sentiment classification accuracy.

2.4 Particle Swarm Optimization

A. S. Zibran et al. (2015) this study presents a feature selection method in light of particle swarm optimization (PSO) for sentiment analysis on Twitter data. The proposed method improves feature selection by limiting an expense capability. The creators exhibit the adequacy of the PSO-based approach in further developing sentiment classification accuracy.

2.5Bag of Words and selection Features Pearson's Correlation

F. R. Saputra Rangkuti et.al proposed Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection. Microblogging has gained immense popularity among internet users, serving as a valuable source of information, especially for film reviews and opinions. In this study, sentiment analysis of film reviews is conducted using ensemble features and the Bag of Words technique, coupled with Pearson's Relationship for feature selection to reduce dimensionality and identify optimal feature combinations. Multiple Naïve Bayes models, such as Bernoulli, Gaussian, and Multinomial Naïve Bayes, are employed for classification. The results indicate that non-standard words in tweet assessments yield an accuracy of 82%, precision of 86%, recall of 79.62%, and an f-measure of 82.69% with 20% feature selection. Furthermore, manual standardization of words boosts accuracy by 8%, resulting in 90% accuracy, 92% precision, 88.46% recall, and an f-measure of 90.19% with 85% feature selection. These findings suggest that word standardization improves classification performance, and Pearson's feature selection optimizes feature combinations while reducing dimensionality.

Iii .Proposed Methodology

3.1 Feature Selection

Feature selection is a crucial process that addresses feature ambiguity by identifying definitive and relevant features. Machine learning algorithms can be severely impacted by excessive features, making it challenging to discern ideal feature sets from noisy ones. Heterogeneous datasets, such as Twitter data in social media, often contain complex feature relationships, redundancies, and irrelevant sets. Consequently, these features can hinder classification accuracy. The goal of feature selection is to optimize learning-based models by selecting a subset of essential features, improving classification accuracy while reducing dimensionality and computational complexity.

Feature selection has gained increasing significance due to its pivotal role in curtailing classification expenses, particularly in terms of time and computational resources. One viable strategy for reducing the dimensionality of the feature space is the utilization of Genetic Algorithms (GAs). However, when applied to text feature selection, GAs encounters a noteworthy challenge - premature convergence. This issue arises from a dearth of diversity in subsequent generations of the algorithm. To tackle this challenge, enhancements have been implemented in the GA's crossover operator. These improvements encompass two key aspects: a) the introduction of a variable slice-direction for determining the size of genes to be exchanged during offspring creation, and b) the incorporation of feature frequency scores to inform the selection of genes for exchange.

3.2 Genetic Algorithm (GA)

Genetic Algorithms (GAs) are a machine learning model inspired by the emulation of natural evolutionary processes. Within a computer-based context, a population of individuals is generated, and these individuals are represented as chromosomes in the form of character strings. These chromosomes serve as potential solutions to the optimization problems under consideration. In GAs, individuals are typically depicted as n-digit binary vectors, mapping to an n-dimensional Boolean space within the search domain. GAs employs specific fitness-based probabilistic selection mechanisms to choose individuals from the current population for the creation of the next generation. The chosen individuals then undergo genetic operations, such as mutation, which introduces slight random alterations to a single string, and crossover, which combines two parent individuals to yield two offspring (Govindarajan 2013). This process of fitness-driven selection and the application of genetic operations are iterated multiple times until a satisfactory solution is attained. The basic operation of the GA is framed as follows:

```
Procedure:
begin
        $t \le 0$
initialize $P(t)$
while (not ter min ation condition)
        $t \le t + 1$
 select $P(t)$ from $P\ (t-1)$
crossover $P(t)$
mutate $P(t)$
evaluate $P(t)$
end
end
```

The basic stages in genetic algorithm are given in figure 2. Improving the efficiency and effectiveness of feature selection for Twitter data involves integrating enhancements into the Genetic Algorithm (GA) framework. These enhancements encompass advanced selection, crossover, and mutation techniques. Additionally, fine-tuning through experimentation and boundary adjustments is often essential to ensure the GA operates at its peak performance.

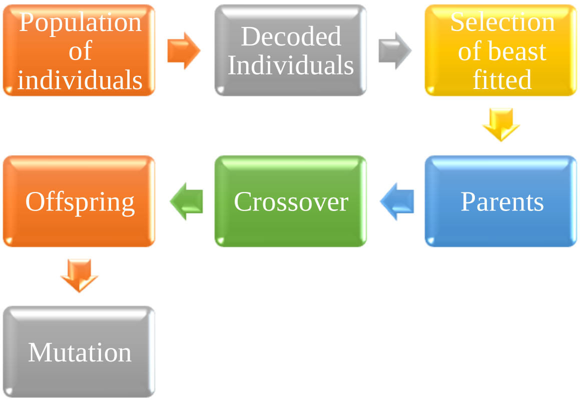

Figure 2.Basic GA

The augmented genetic algorithm in this situation functions as a feature subset enhancer and collaborates with a machine learning (ML) method as an accuracy estimator for fitness evaluation. This synergy entails integrating GA-based search techniques with ML accuracy evaluations to iteratively sample and refine the feature subset. As a result, this iterative process results in the feature set's reduction and optimisation, preparing it for use in a classification method later on. The point by point procedures are summarized as follows:

1. Chromosome Encoding. To address the unique nature of the feature selection problem, which may be effectively described using 1s and 0s, the researchers modified the binary encoding of chromosomes. A chromosome is effectively a string of 0s and 1s in this encoding technique, where a '1' denotes the inclusion of a feature and a '0' denotes its exclusion from the set of features as a whole. This string's length reflects the overall number of features offered.

2. Population Initialization. Every chromosome is randomly assigned a number between 0 and 1, with lengths equal to the absolute characteristics.

3. Fitness function. The fitness of each chromosome in the current population is assessed based on two main criteria: the number of features that have been chosen and the classifier accuracy of the particular machine learning algorithm when applied to the dynamic feature combination encoded in the chromosome. This is accomplished by using a fitness function that has been taken directly from a particular source [19]. The fitness is defined as:

$$fitness = \left(B * acc(S)\right) + (1 - B) * (1/S) \qquad (1)$$

In this context, "S" stands for the dynamic or chosen features, and "B" is the balance factor that controls the trade-off between the size of the feature subset and the accuracy of the classification. Accuracy is prioritised over the size of the feature subset when "B" has a greater value. The accuracy score generated from assessing the machine learning algorithm using the chosen feature subset is denoted by "Acc (S)" in the meantime.

4. Selection. Using the rank-based selection method, two parent chromosomes are chosen, guaranteeing that those with the best fitness aren't passed over as the population changes. The best-performing chromosomes have a greater chance of pairing up and producing improved children in succeeding generations thanks to this method. The population is initially arranged according to their fitness ratings in descending order. The remaining high-fitness chromosomes are then retained by a selection process that determines which ones will experience crossover and mutation based on predetermined probabilities.

5. Crossover. The crossover operation in this improvement procedure proceeds as follows: First, a modification was made by designating a variable slice point for genes that would be transferred during the production of offspring. Second, the cumulative feature frequency score constrains the crossover operator and directs the choice of the genes to be switched. The same slice locations were consistently applied to both parent chromosomes using this dynamic determination method. The feature subsets represented by these specific slice points were documented to calculate the total frequency score, following the same procedure for other feature subsets. Subsequently, the feature subsets from both parents were compared. The child1 was chosen as the one with the higher score, while child2 incorporated the remaining feature subset. This variable-slicing multi-point crossover process is depicted in Figure 3 for clarity.

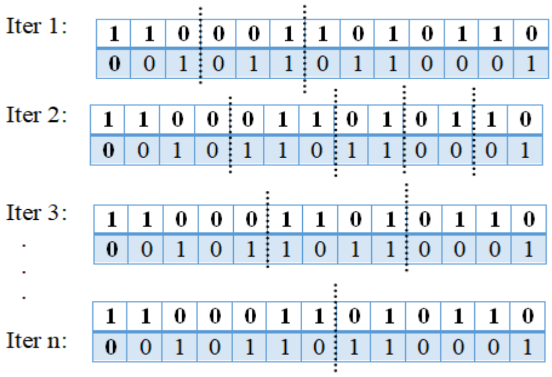
Iter 1:
1 1 0 0 0 1 1 0 1 0 1 1 0
0 0 1 0 1 1 0 1 1 0 0 0 1
Iter 2:
1 1 0 0 0 1 1 0 1 0 1 1 0
0 0 1 0 1 1 0 1 1 0 0 0 1
Iter 3:
1 1 0 0 0 1 1 0 1 0 1 1 0
0 0 1 0 1 1 0 1 1 0 0 0 1
.
.
.
Iter n:
1 1 0 0 0 1 1 0 1 0 1 1 0
0 0 1 0 1 1 0 1 1 0 0 0 1

Figure 3.Example of variable slice-point for genes swapping in the offspring generation loop

6. Mutation. By switching the cycle, the offspring go through the mutation process using the mutation probability (Pm) value.

7. Population Update. When the population's total number of chromosomes reaches its maximum, this population is then designated as the new generation and will be carried forward into the future. Algorithm 1 furnishes a breakdown of the procedures executed within the improved Genetic algorithm.

Proposed Enhanced Genetic Algorithm:

```
Step 1: Initialize the population by creating random chromosomes consisting of 0s and 1s. Every chromosome has a length equivalent to the complete number of features.
Step 2: Evaluate the fitness value for each chromosome in the population.
Step 3: Using the fitness values of the chromosomes, sort the population in descending order.
Step 4: While until the desired number of generations is reached do
        Select the best-performing chromosomes as parents based on their fitness values, using a probability-based selection method.
Step 5: While until the desired population size is reached do
        Create offspring through crossover and mutation operations:
Step 6: Choose two parent chromosomes from the selected parents.
Step 7: Perform crossover based on a crossover probability (PC).
Step 8: Determine a random slice point in both parent chromosomes, using a randomly generated number between 2 and (chromosome length / 15).
Step 9: Determine the length of the feature subset by isolating the chromosome length by the random slice point.
Step 10: Set a feature subset counter to 1.
Step 11: Make offspring chromosomes:
Step 12: While the feature subset counter is less than or equal to the random slice point:
Step 13: For both parents (P1 and P2), determine the summative frequency score of the chosen subset of bits, and store them as $P1_{subset}$ and $P2_{subset}$ , respectively.
Step 14: If the score of $P1_{subset}$ is greater than
```

```
or equal to the score of $P2_{subset}$ :
        Perform a random bit flip for $P1_{subset}$ .
        If the bit is flipped to 0, append $P1_{subset}$ to Child1, and append $P2_{subset}$ to Child2.
        Else:
            Append $P2_{subset}$ to Child1, and append $P1_{subset}$ to Child2.
Else:
Append $P2_{subset}$ to Child1, and append $P1_{subset}$ to Child2.
Step 15: Addition the feature subset counter.
Step 16: Perform mutation on the offspring chromosomes based on a mutation probability (Pm).
Step 17: Append the offspring chromosomes to the parents set.
Step 18: Update the population with the new offspring.
Step 19: Using the fitness values, rank-sort the population of chromosomes in descending order.
Step 20: Give back the best feature subset that was determined using the highest fitness value.
```

Iv.Experimental Results

Figure 4 illustrates a comparison chart of Precision, showcasing the performance of the existing Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and the proposed Enhanced Genetic Algorithm (EGA). The x-axis represents different datasets, while the y-axis indicates the Precision ratio. Notably, the proposed EGA consistently outperforms the existing algorithms, with Precision values ranging from 90.78 to 99.76, compared to the existing algorithm's range of 73.94 to 83.12 and 74.72 to 85.37. Figures 5, 6, and 7 similarly present comparisons of Recall, F-Measure, and Accuracy, revealing the superior results achieved by the proposed method across various metrics.

4.1 Precision

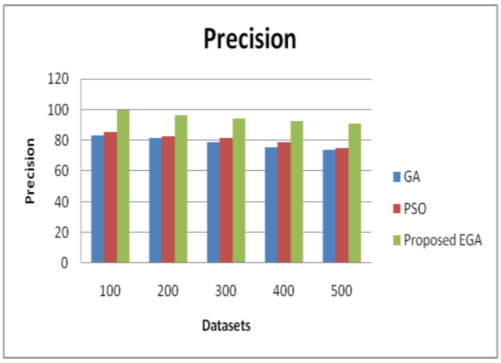

Figure 4.Comparison chart of Precision

4.2 Recall

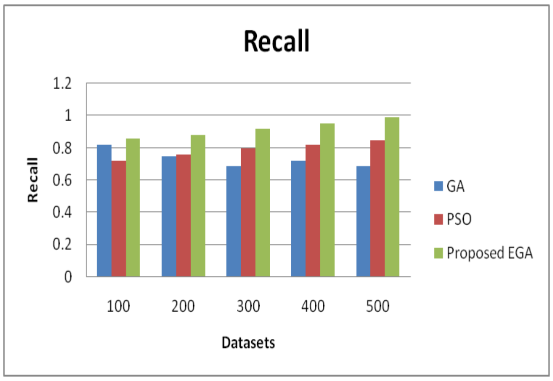

Figure 5.Comparison chart of Recall

4.3 F -Measure

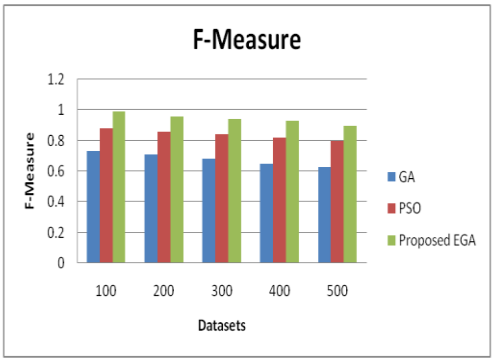

Figure 6.Comparison chart of F –Measure

4.4 Accuracy

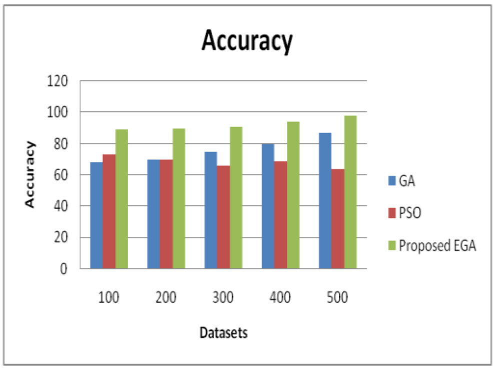

Figure 7.Comparison chart of Accuracy

V. Conclusion

In this paper, we proposed an enhanced genetic algorithm (GA) for feature selection to anticipate stock market sentiment using Finance Yahoo stocks data and Twitter data. The enhanced GA incorporated techniques to further develop investigation and exploitation capabilities, enabling a more comprehensive search of the feature space and enhancing the nature of selected features. The enhanced GA further developed investigation and abuse capabilities, prompting better feature selection. Experimental results showed that our methodology outperformed traditional GA and PSO methods in accuracy and forecast performance. The selected feature subsets caught important patterns and sentiments connected with stock market movements. This combination of data sources is important for anticipating sentiment, giving insights to traders, investors, and market analysts in settling on informed decisions.

References

[1] Tuba, S. Mirjali, S. Mirjalili, and M. A. Razzaq. "A Genetic Algorithm-based Feature Selection for Twitter Sentiment Analysis." In Proceedings of the International Conference on Engineering & MIS (ICEMIS), 2018.

[2] M. Alam and M. M. Islam. "A Hybrid Feature Selection Framework for Twitter Sentiment Analysis." In Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.

[3] M. Z. Ali, K. Muhammad, A. A. Khan, and S. Khan. "Feature Selection for Twitter Sentiment Analysis: A Comparative Study." In Proceedings of the International Conference on Computational Intelligence (ICCI), 2020.

[4] S. Zibran, M. H. Kabir, and R. Ahmed. "A Feature Selection Technique for Twitter Sentiment Analysis Using Particle Swarm Optimization." In Proceedings of the International Conference on Informatics, Electronics & Vision (ICIEV), 2015.

[5] F. R. Saputra Rangkuti, M. A. Fauzi, Y. A. Sari and E. D. L. Sari, "Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection," 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018, pp. 88-91, doi: 10.1109/SIET.2018.8693211.

[6] R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 21-25, doi: 10.1109/CCOMS.2019.8821650.

[7] M. Bibi et al., "Class Association and Attribute Relevancy Based Imputation Algorithm to Reduce Twitter Data for Optimal Sentiment Analysis," in IEEE Access, vol. 7, pp. 136535-136544, 2019, doi: 10.1109/ACCESS.2019.2942112.

[8] E. S. Usop, R. R. Isnanto and R. Kusumaningrum, "Part of speech features for sentiment classification based on Latent Dirichlet Allocation," 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2017, pp. 31-34, doi: 10.1109/ICITACEE.2017.8257670.

[9] A. Yang, J. Zhang, L. Pan and Y. Xiang, "Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination," 2015 International Symposium on Security and Privacy in Social Networks and Big Data (SocialSec), 2015, pp. 52-57, doi: 10.1109/SocialSec2015.9.

[10] N. K. Suchetha, A. Nikhil and P. Hrudya, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862033.