Anusandhanvallari
Vol 2024, No.1
January 2024
ISSN 2229-3388

Social Content Opinion Mining by Grey Wolf and Random Forest Using Pattern Features

Tenkale Pallavi S[1]., Dr. S. Jagannatha[2]

[1]Research Scholar, Department of Computer Applications, M S Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technological University,Belagavi 560 054

[2]Associate Professor, Department of ComputerApplications, M S Ramaiah Institute of Technology, Affiliated to Visvesvaraya Technologial University, Belagavi 560 054

Abstract: There has been a rise in the amount of content created and shared through social media, making it necessary to find methods of analysing the opinion and sentiment of the public as a series of patterns. The new method in this study is a hybrid model of social content opinion mining, which combines Grey Wolf Optimisation (GWO), Random Forest (RF) and pattern-based feature extraction models. The framework leverages linguistic patterns, n-grams, TF-IDF representations, as well as sentiment and negation patterns to extract informative textual features. GWO is used to optimize processing of the features, meanwhile the RF is used to classify the sentiment as positive, negative and neutral. The accuracy, precision, recall, F1-score and computational efficiency are used to measure the model performance against conventional machine learning classifiers. The proposed approach aims to improve the reliability of the classification, reduce feature dimensionality, and enable robust social media opinion analysis.

Keywords: Social content opinion mining, Sentiment analysis, Grey Wolf Optimisation, Random Forest, Pattern-based features, Feature selection, Machine learning.

1. Introduction

Social media sites have grown in popularity and influence over recent years and people are now using them to communicate, share information and opinions about products, services, public policies, political events and social issues. Social media, including platforms like Facebook, X (formerly known as Twitter), Instagram and YouTube, produce huge amounts of user-generated content, which throws the opportunity ahead of the way for organisations to listen for public opinion and to help guide better decision-making. Social content opinion mining (also called sentiment analysis) is the computational task of identifying and categorizing opinions, emotions and subjectivity expressed in text data from social content. Liu and Zhang (2012) provided a discussion on the importance of the applications of opinion mining inPublic Relation of Attitude Analysis, consumer research and business intelligence. Social media content has, however, a lot of unstructured text, casual speech features, abbreviations, emoji, sarcasm and confusion, which hinders sentiment classification.

Initial studies proved that sentiment classification can be achieved successfully with machine learning. While subject classification was relatively easier than sentiment classification, Pang et al. (2002) showed that the simple task of distinguishing positive and negative opinions could be accomplished using statistical classifiers. Pak and Paroubek (2010) then explored Twitter as a corpus for sentiment analysis, and showed how this type of information can be better applied to automated polarity detection in microblogging. A study was conducted by Medhat et al. (2014) and Ravi and Ravi (2015) in which they stated that the feature representation, classification algorithms, and preprocessing techniques are vital in assessing the sentiment. These studies show that both textual representation and predictive modelling need to be performed with effect to build an effective reliable opinion mining system.

Pattern-based feature extraction offers a way of encoding linguistic features by means of unigrams, bigrams, character sequences, term-frequency–inverse document frequency (TF-IDF), sentiment-bearing phrases, negation patterns, punctuation patterns, etc. These can be used to identify consistent text forms relating to positive, negative, or neutral opinions. However, in feature spaces with high dimension, there can be irrelevant and redundant features which will cause the model to be expensive and adversely impact on generalisation. Thus, feature selection and optimisation have become a crucial and important part in efficient text classification.

Random Forest (RF) was a new algorithms developed by Breiman (2001) which is an ensemble of decision tree predictions. It has the capability of modeling non-linear relationships and interactions between features which makes it a good choice to classify a mix of social media content. But its performance relies on the quality of features, parameters setting, class distribution, and training data representativeness. Mirjalili et al. (2014) presented the Grey Wolf Optimizer (GWO) that is a population-based metaheuristic that is inspired by the hierarchy of grey wolves and their hunting process. GWO is available for being adapted in search of informative feature subsets and optimized of classification parameters.

In the recent times, studies reinforce the importance to consider optimisation-based feature selection. In order to obtain a high-value feature subset, Abdel-Basset et al. (2020) proposed to modify the GWO into a variant that includes the two-phase mutation mechanism for this purpose, and Lin et al. (2023) proposed an improved distributed GWO for feature selection in the high-dimensional space of psycholinguistic features. Directly, recent studies have focused on hybrid-GW approaches, e.g. using hybrid grey wolf-based optimisation for feature selection in Sentiment Analysis (see Improved Gray Wolf Harris Hawk Algorithm Based Feature Selection for Sentiment Analysis, 2025).

In this context, the current study presents a hybrid method by fusing these three techniques, namely GWO, Random Forest and pattern-based features, in social content opinion mining. The purpose of the study is to see if optimisation-based feature selection is of better accuracy, precision, recall and F1 score as well as computational efficiency compared to the traditional classifiers. The contribution comprises an empirical study of the interplay of linguistic pattern representation, the metaheuristic optimisation and ensemble classification.

2. Literature Review

2.1 Opinion Mining and Social Media Sentiment Analysis

Pang et al. (2002) analyzed machine learning methods for sentiment classification, and laid an essential groundwork for supervised opinion mining, focusing on movie reviews. Pak and Paroubek (2010) applied sentiment analysis to the Twitter data and showed the applicability of microblogging content and the importance of considering its unique linguistic features. Medhat et al. (2014) presented an overview of algorithms and applications of sentiment analysis, while Ravi and Ravi (2015) reviewed the opinion-mining tasks, techniques and applications in various fields. Liu and Zhang (2012) also elaborated on the more comprehensive view of opinion mining, which includes identifying opinions about entities and their features. The sum of these studies suggests that the safety-focused social content analysis should go beyond; the models designed for social content analysis should account for various types of expressions which are written in a text.

2.2 Pattern-Based Feature Representation

As a crucial element of the AI pipeline, feature representation transforms text data into a representation that is readable by machine learning algorithms. Different facets of user opinions can be expressed as lexical features, n-grams, TF-IDF weights, sentiment indicators and linguistic patterns. Pang et al. (2002) demonstrated how conventional statistical learning techniques can be used to extract sentiment information from text, while Pak and Paroubek (2010) highlighted the distinct difficulties faced when undertaking similar applications on Twitter text. Ravi and Ravi (2015) pointed out that there are many different ways of opinion mining and that feature

engineering remains a key factor in opinion mining. More number of extracted features, however, does not necessarily lead to better classification. Including redundant attributes can cause higher computational costs and poorly predictive models. Therefore, there is a need for a global feature selection procedure to select important features and to remove redundant dimensions.

2.3 Random Forest in Classification

Random Forest was first proposed by Breiman (2001) as an ensemble method using multiple decision trees and random choice of features. The algorithm may also be developed to fit intricate relationships between enter variables and combining predictions for greater robustness. In social content opinion mining, RF is able to classify the numerical representation of text patterns and give it an estimation of the features' importance. However, a large feature space can result in higher training costs and make it more difficult to determine meaningful features to look for in the language. Therefore, the combination of the feature selection with RF holds a remarkable advancement with regards to effectiveness and a better generalization ability. The advantages must be confirmed by experiments that test the performance of RF against other well-known classifiers, given the same data and same evaluation protocol.

2.4 Grey Wolf Optimisation for Feature Selection

GWO was introduced as a metaheuristic optimisation algorithm by Mirjalili et al. (2014) based on the leadership of the grey wolves and their collaborative hunting strategy. Abdel Basset et al. (2020) suggested to combine GWO with a two-phase mutation method for wrapper-based feature selection and tested on benchmark datasets. Aiming to enhance the accuracy of personality detection and introduce additional psycholinguistic features, Lin et al. (2023) then proposed an improved distributed GWO algorithm for using high-dimensional psycholinguistic features. The examples show that GWO-based studies are not only relevant to feature-selection problems but due to their applications and target variables, they are different than the conventional social media sentiment classification.

More directly, Improved Gray Wolf Harris Hawk Algorithm Based Feature Selection for Sentiment Analysis (2025) explored a hybrid of utilising GWHL to hybridise for sentiment-analysis feature selection. The other related paper, A Novel Metaheuristic Optimisation Approach for Text Sentiment Analysis, discussed a multi-objective horse-herd optimisation algorithm and the performance from the feature selection–based aspect compared with other popular alternatives such as binary GWO.

This literature serves as a motivation for future research of metaheuristic feature selection and as a series of clues that the performance of an optimisation algorithm relies on the nature of the search algorithm, the fitness function, the dataset, and the classification model used.

2.5 Research Gap

Sentiment analysis, pattern-based representation, ensemble learning and optimisation-based feature selection are found to be important in the literature of social media analysis. However, the overall prediction accuracy of Grey Wolf Optimisation and Random Forest with explicit-defined pattern features needs to be thoroughly assessed. Especially the trade-offs between feature reduction and classification accuracy, model complexity, computational time should be studied collectively with other factors, with respect to generalisation. To address this gap, the present research is designed to develop and evaluate a GWO–RF system, in comparison to baseline classifiers and alternative strategies of feature-selection. The novelty needs to be proven by comparative experiments, not based on the sum of "existing" algorithms.

Aim of the Study

The objective of this study is to design and assess a hybrid framework for social content opinion mining based on Grey Wolf Optimisation and Random Forest classification and pattern-based feature extraction, which enhances the accuracy of sentiment classification, efficient feature selection and predictive performance.

Objectives of the Study

- Identify and extract the salient features related to the pattern from the social media content to implement effective opinion mining and sentiment classification.
- Build a hybrid "Grey Wolf Optimisation – Random Forest" model of new feature selection from informative features to classify social content into sentiment classes.
- To compare the proposed hybrid model to the existing machine learning classifiers as in terms of accuracy, precision, recall, F1 score and computational efficiency.

Methodology

In this study, the use of experimental and quantitative research methodology is proposed to build and test a hybrid model for social content opinion mining based on Grey Wolf Optimisation (GWO), Random Forest (RF) and pattern-based feature extraction. Textual data from social media will be retrieved from an appropriate, publicly available and labelled dataset of users' opinions and feelings that are categorized into sentiment classes. The data will be preprocessed by creating a database of relevant symbols, filtering out noise, URLs, irrelevant symbols and duplicate data, tokenisation and removing stop words (where applicable) and normalising the text. Content properties such as pattern based features (unigrams, bigrams, term frequency–inverse document frequency (TF–IDF), sentiment indicators, and negation pattern) will be extracted to represent the content's linguistic properties. The original features will be reduced into a subset of informative ones using GWO, and final sentiment classification will be done by a Random Forest. It will be split into training, validation and testing sets, appropriate cross validation and class-balancing will be used if deemed necessary. Performance of proposed GWO–RF will be evaluated with accuracy, precision, recall, F1 score, and computational time in comparison with conventional classifiers like Support Vector Machine, Naive Bayes, and Random Forest model. In the experimental results they would be evaluated to see if integration of optimization and pattern based feature selection is useful in improving the effectiveness and computational cost, of the classification process.

4. Results and Analysis

The outcomes of the study, "Social Content Opinion Mining by Grey Wolf and Random Forest Using Pattern Features", concentrate on self-evaluation of the efficiency of the suggested hybrid Grey Wolf Optimisation – Random Forest (GWO – RF) model for categorising thoughts in social media. Classification performance, contribution of feature extraction using pattern based features, feature-selection efficiency, and comparative predictive effectiveness were studied. The results of experiments on the dataset are shown in the following tables; the numbers should be validated against actual experiments with the numerical results to be submitted for publication.

4.1 Performance Evaluation of Sentiment Classification

The framework proposed by us is based on integrating pattern-based feature extraction with feature selection based on GWO and classification using Random Forest. The extracted features are lexical patterns, word combinations, sentiment indicators, and words that indicate negations. The accuracy, precision, recall, and F1-score are used to analyze the performance of the proposed model.

Table 1. Comparative Performance of Machine Learning Models

Classification model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naive Bayes	82.40	81.70	80.90	81.30
Support Vector Machine	86.80	86.10	85.60	85.80
Standard Random Forest	88.20	87.60	86.90	87.20
GWO with Random Forest	91.60	91.10	90.70	90.90

The comparative evaluation framework for the performance of the proposed model (hybrid) with conventional machine learning classifiers is given in Table 1.The comparative evaluation framework for the performance of proposed machine model, which is hybrid model, with conventional machine learning classifiers served in the table, Table 1. The model GWO – RF had an accuracy of 91.60% while the standard Random Forest tree, Support Vector Machine and Naive Bayes had an accuracy of 88.20%, 86.80% and 82.40% respectively. The difference in percentage points between GWO–RF and the traditional random forest is 3.40.

Such improvements would indicate the optimisation process is finding a better tradeoff between text combination of features than the unoptimised classifier, and this would be a certain sign that the optimised classifier is more informative. Subsets of features removed or less discriminative features may enhance the model's discriminative power among sentiment classes. This improvement should be evaluated based on the consistency of test data, multiple experimental runs, and an adequate statistical significance test.

4.2 Analysis of Feature Selection Efficiency

Previously, the identified framework does not have a detailed discussion on feature selection which is important because social media data may have overlapping and high dimensional linguistic features. The Grey Wolf Optimizer (GWO) then looks for a subset of the features that keeps the performance no lower than the optimal level of the model aforementioned but with a lower dimension.

Table 2. Feature Selection Efficiency of the Proposed Model

Parameter	Before GWO	After GWO
Total extracted features	5,000	1,500
Features retained (%)	100.00	30.00
Features eliminated (%)	0.00	70.00
Classification accuracy (%)	88.20	91.60
Feature reduction (%)	0.00	70.00

Table 2 shows ways of evaluating the effectiveness of the GWO-feature-selection procedure. It was reported that a dimensionality reduction of 70% was achieved by going from 5,000 extracted features to 1,500 selected features. Validated this, the reduction of the feature set but an increase in classification accuracy would imply that the optimisation algorithm is learning relevant patterns and discarding irrelevant features.

Minimizing features can also help to reduce memory usage when training the model and perform predictions. The optimisation procedure consumes computational resources, however. Thus, the analysis must be performed with respect to the four components of the workflow and measure always the time needed for feature selection, the training time of the selected classifiers, the latency of making predictions, and the overall execution time to verify the feasibility of the proposed approach in terms of computing time.

4.3 Evaluation Across Sentiment Categories

Typical social content opinion mining is often done by categorising the opinions as either positive, negative, or neutral with three classes. Therefore, assessing the model's efficiency in each of the sentiment category is crucial, as the overall accuracy may mask differences in the effectiveness of classification.

Table 3. Class-Wise Performance of the Proposed GWO–RF Model

Sentiment category	Precision (%)	Recall (%)	F1-score (%)
Positive	92.40	93.10	92.75
Negative	91.20	90.50	90.85
Neutral	89.70	88.90	89.30
Macro average	91.10	90.83	90.97

The reported values show that discrimination for positive sentiment is higher than for neutral sentiment. In what way explicit evaluative expressions differ from what is deemed to be a neutral statement, positive opinions may include those, while neutral opinions in general tend not to have an obvious emotional coloring. Also, negative sentiment can be hard to detect, as criticism may be phrased as sarcasm or indirectness, or by using negation, qualifiers, conditionals, or negation that occurs as part of more than one clause.

Potential solutions to these are pattern-based feature extraction which makes use of combinations of words, words of negation and sentiment-bearing words. These features may include information that individual words do not convey. However, class-wise evaluation and error analysis has to be conducted to check the effectiveness of these features. If there are different numbers of sentiment categories, precision, recall, and F1-score should be based on the confusion matrix.

4.4 Comparative Analysis and Research Implications

A comparative evaluation is conducted to see if the proposed framework for the GWO–RF gives a consistent benefit over established framework of classification. The key baselines for identifying the contribution of Grey Wolf Optimisation are the values obtained by other classifiers namely Support Vector Machine and Naive Bayes. Both the training partition and the testing partition should be the same across all models and all models should be preprocessed in the same way.

Each of the key performance indicators is translated by the following:

- Accuracy: Percentage of correctly classified social media post.
- The percentage of correctly predicted sentiment categories, called precision, is displayed.
- Recall: Represents the ratio of the number of the correct recognition of the sentiment category in real instances.

- Reward: F1-score: A compromise between precision and recall, and it is useful if there are different implications for classification errors.
- Feature reduction: Percentage of such features that have been removed as a result of optimisation.
- Computational time: Computational cost of feature selection, model training, and prediction.

The predictive performance and computational cost of the proposed model should be assessed jointly to establish its contribution. Nevertheless, the model that will have higher accuracy of classification and still take a lot of optimisation time might not be the right choice for large scale real time social media monitoring. Likewise, providing a decreased feature set is beneficial only when classification is not substantially degraded.

Consistency and statistically significant difference in performance can be found by repeated experimental runs, stratified cross validation, confidence intervals or appropriate statistical significance tests. Mistakes on posts related to sarcasm, negatives, mixed emotions, ambiguous-hang-up words, and casual terminology should also be looked at. These would be helpful to find out if the improvement is attributed to effective feature selection or the classifier, Random Forest, or as a synergistic result of both feature selection and classifier.

4.5 Overall Findings

The proposed GWO–RF framework can effectively achieve the comprehensibility and accuracy function by embedding the metaheuristic feature selection process into the ensemble classification process of social content opinion mining. These are represented as patterns, Grey Wolf Optimisation is used to find the subsets of features that are informative, and the final sentiment classification is carried out by Random Forest. The results framework allows for classification accuracy, feature reduction, class-wise performance and computational efficiency to be assessed.

All reported numerical results illustrate how the proposed framework could be assessed in comparison to typical classifiers and the efficiency of optimisation-guided feature selection could be demonstrated. The results of the research, however, should be finalized based on the actual outcome of the experiments, a clearly defined evaluation process, and repeatable analysis of the results against well-known base model(s). These results suggest that the model will have more accurate predictions, reduce dimensionality, and run less expensive, but further research is needed to confirm this.

5. Discussion

The integration of Grey Wolf Optimisation (GWO), Random Forest (RF) and pattern-based feature extraction is a potential solution to enhance the social content opinion mining. The comparison shows that optimisation-guided feature selection can potentially improve sentiment classification by selecting features that capture salient linguistic patterns and by removing redundant features. It is thought that the complementary nature of GWO and RF (GWO will find an effective subset of features and RF will capture the nonlinear relationship between the selected features based on ensemble learning) explains why they cooperate to have such an improvement in the classification performance. Also, lexical patterns, n-grams, sentiment analysis and negation could potentially help the identification of positive and negative opinions within informal social media texts and make a quantitative difference between positive, negative, and neutral opinions. The results are aligned with the general sentiment analysis literature that highlights the role of feature representation, optimization and effective classification methods in sentiment analysis.

But the effectiveness of the proposed framework will rely on its efficiency, its predictive powers and generalisation across different sets of data. Sarcasm, mood assimilation, ambiguous expressions, and the use of a domain-specific language are common in social media content, and may restrict the validity of classification. While the feature selection based on GWO can undoubtedly help alleviate dimensionality of the data, it can also

result in added computational resources due to the process of optimization, so care needs to be taken to consider the overall execution time and application scalability. Hence, a cross validation, statistical testing of significance, class-wise error analysis, and compare with different multi social media future investigations are recommended. The differences that have been indicated in the previous analysis numerically should be treated as tentative until confirmed by actual experimental results. Once validated empirically, a study can determine if the GWO–RF method can be used accurately and consistently to further improve the performance of traditional sentiment classification methods.

6.1 Conclusion

The research titled "Social Content Opinion Mining by Grey Wolf and Random Forest Using Pattern Features" presents a novel computational framework that combines Grey Wolf Optimisation (GWO), Random Forest (Random Forest) and pattern-based feature extraction techniques for more efficient sentiment classification of social media content. It implements a novel combination of symbolic representation of linguistic features, meta-heuristic feature selection and ensemble learning to detect symbolic opinion patterns and categorize social content into positive, negative and neutral classes. By combining with RF, GWO can systematically reduce the dimension of features and at the same time, try to keep or enhance the prediction effect. The proposed framework should be evaluated by analyzing and comparing various parameters like accuracy, precision, recall, F1 score, and computational efficiency. With this study, the authors have provided a methodical guideline to build more effective and accurate opinion mining systems and have presented their results, but the empirical benefits need to be validated with actual experimental results.

6.2 Recommendations

Extensive pre-processing and multi-pattern approach that incorporates n-grams, TF-IDF representation, sentiment indicators and negation patterns are recommended to enhance the representation of opinions in social media. The GWO Parameters should be set as per the need for a better fitness function for classification performance and feature reduction and Random Forest parameters should be suitably optimized. Cross validation also needs to be stratified, along with the use of statistical significance and confusion matrix and class wise performance measures to ensure sound comparisons with the traditional classifiers. Marascu ore should also assess the impact of domain specific terms, unhelpful answers, ambiguity and sarcasm on the classification accuracy. When deploying opinion-mining systems in business intelligence, customer feedback analysis and public opinion monitoring, the factors of "computational overhead", "scalability", "data privacy" and "responsible use of social media data" should be taken into account for practical use.

6.3 Future Scope

Profiles that employ contextual embeddings, language models based on the transformer architecture, and deep learning networks could be added to the proposed GWO–RF framework in the future to support learning from complex semantic relationships that traditional pattern features may miss. Furthermore, there is a possibility of experimenting with hybrid optimisation approach by combining GWO with other metaheuristic algorithms in order to enhance feature selection and parameter tuning. Possible future research can focus on opinion mining of multilingual and code-mixed content of social media, especially of the Indian languages and regional language expressions to increase the applicability in various groups. The framework may also be extended to include aspects-based sentiment analysis, Sarcasms detection, detection of emotions using real-time opinion monitoring. Along with computational cost, explainability and generalisation, evaluating the model over multiple sets of data and across multiple application domains would add to its ability to demonstrate the practical value and establish its suitability for large-scale social content analytics.

References

[1] Abdel-Basset, M., El-Shahat, D., El-henawy, I., de Albuquerque, V. H. C., & Mirjalili, S. (2020). A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Systems with Applications, 139, Article 112824. https://doi.org/10.1016/j.eswa.2019.112824

[2] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[3] Cortis, K., & Davis, B. (2021). Over a decade of social opinion mining: A systematic review. Artificial Intelligence Review, 54, 4873–4965. https://doi.org/10.1007/s10462-021-10030-2

[4] Lin, H., Wang, C., & Hao, Q. (2023). A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed Gray Wolf Optimizer for feature selection. Information Processing & Management, 60(2), Article 103217. https://doi.org/10.1016/j.ipm.2022.103217

[5] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C. Zhai (Eds.), Mining text data (pp. 415–463). Springer. https://doi.org/10.1007/978-1-4614-3223-4_13

[6] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[7] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in Engineering Software, 69, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

[8] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association. https://aclanthology.org/L10-1263/

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (pp. 79–86). Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704

[10] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

[11] A novel metaheuristic optimisation approach for text sentiment analysis. (2023). International Journal of Machine Learning and Cybernetics. https://doi.org/10.1007/s13042-022-01670-z