Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 6, No. 2, 2024

Subjective or Objective: A Study of Classifying News Content based on Machine Learning Algorithms

-- An Example of Sports Articles

Chong Zhu [1, *], Yiman Yu [2]

[1] School of Business Administration, Anhui University of Finance and Economics, Bengbu Anhui, 233030, China
[2] School of Journalism and Communication, Henan University, Kaifeng Henan, 475001, China

[*] **Corresponding author**: Chong Zhu (Email: Chongzhu218@163.com)

Abstract: In today's world, self-media as well as streaming media platforms have gradually become the main channels for news dissemination. With the advent of the artificial intelligence era, AI-generated technology will profoundly affect the news industry. Many press releases generated by AI are sufficiently "fake" as well as cascading, so the subjectivity and objectivity of news media are gradually blurred. In order to better standardize and regulate the articles published on self-media platforms, this study is based on 1,000 sports articles and machine learning algorithms to categorize and describe the content of the articles, so as to determine whether the articles are objective or not. At the end of the study, we found that the KNN algorithm performed best in classification.

Keywords: Machine Learning; Classification Algorithms; Sports News Articles; Subjective and Objective.

1. Introduction

The rise of self-media platforms in recent years has led to an unprecedented abundance and growth in the content and quantity of news, with platforms such as Jitterbug, Shutterbug, and Weibo gradually becoming the main way for people to learn about the world. Of course, this explosive development has had a huge impact on the field of news dissemination, among which is the much-anticipated sports news. However, the objectivity of the content of these news has been the center of debate, and the bias and subjective interpretation of the self media may affect the accuracy and objectivity of the information conveyed to the audience. In this context, analyzing the objectivity of self-published media pieces becomes crucial, not only to maintain the integrity of the news, but also to ensure that people receive accurate and unbiased information.

To address this issue, this study aims to take sports news with objective and subjective colors as the object of study, and use the frequency of the more common linguistic conventions (e.g., the frequency of personal pronouns, the frequency of possessive pronouns, the frequency of adverbs, the frequency of comparative adverbs, and the frequency of high-level adverbs intonation and other vocabulary) of sports news as the independent variable. The dataset is from the UCI Machine Learning Repository and can be used for model training and evaluation. The richness and extensiveness of this dataset provide a reliable basis for our study, enabling us to obtain reliable results during the training and validation of our classification algorithms. At the same time, we hope that this research contributes to the development of the field of news dissemination in the era of artificial intelligence, and provides a strong support for the authenticity and effectiveness of news through artificial intelligence technology.

One of the primary objectives of this study was to investigate the efficacy of various classification algorithms, such as K-nearest neighbor (KNN) algorithms and support vector machines (SVM), in discerning between objective and subjective sports articles. By employing these algorithms, our aim is to develop a classification model that accurately categorizes the objectivity and subjectivity of articles based on their content and language usage. This will facilitate news editors and reviewers in promptly identifying potential biases in news manuscripts, thereby enhancing the objectivity and accuracy of overall news reports. Additionally, it will foster a more conducive public opinion environment within we-media and streaming media platforms.

In addition, this study holds significant value and significance for the future regulation and monitoring of we-media and streaming media, particularly in the realm of artificial intelligence (AI) driven content regulation. With the increasing role of artificial intelligence and generative large models in writing, monitoring, and filtering multimodal content, comprehending the subtleties between objectivity and subjectivity in sports reporting can offer guidance for developing more advanced AI systems capable of distinguishing reliable, unbiased reporting from potentially misleading or inflammatory content. This will contribute to establishing a fairer and more transparent environment for news dissemination while ensuring public access to truthful and objective information.

In summary, this study aims to contribute in some way to the ongoing debate about media objectivity and journalistic integrity by using machine learning techniques to analyze the subjectivity and objectivity of sports articles. Through classification modeling as well as evaluation, it provides insights into whether machine learning algorithms are effective for the field and which machine learning algorithms perform better, in an attempt to provide insights that can guide both journalistic practices and AI-based content regulation strategies in the digital age. This research is expected to drive the news industry in the direction of greater objectivity, transparency, and accountability, providing the public with a more credible source of information.

2. Data Sources

1000 sports articles were labeled using Amazon Mechanical Turk as objective or subjective. The raw texts, extracted features, and the URLs from which the articles were retrieved are provided. Table 1 shows the interpreted meaning of the variables [1].

Table 1. Explanatory implications of training variables

Variable Name	Variable Interpretation
Label	objective vs. subjective
Total-Words-Count	total number of words in the article
Semantic-objscore	Frequency of words with an objective SENTIWORDNET score
Semantic-subjscore	Frequency of words with a subjective SENTIWORDNET score
CC	Frequency of coordinating conjunctions
CD	Frequency of numerals and cardinals
DT	Frequency of determiners
EX	Frequency of existential there
FW	Frequency of foreign words
INs	Frequency of subordinating preposition or conjunction
JJ	Frequency of ordinal adjectives or numerals
JJR	Frequency of comparative adjectives
JJS	Frequency of superlative adjectives
LS	Frequency of list item markers
MD	Frequency of modal auxiliaries
NN	Frequency of singular common nouns
NNP	Frequency of singular proper nouns
NNPS	Frequency of plural proper nouns
NNS	Frequency of plural common nouns
PDT	Frequency of pre-determiners
POS	Frequency of genitive markers
PRP	Frequency of personal pronouns
PRP$	Frequency of possessive pronouns
RB	Frequency of adverbs
RBR	Frequency of comparative adverbs
RBS	Frequency of superlative adverbs
RP	Frequency of particles
SYM	Frequency of symbols
TOs	Frequency of 'to' as preposition or infinitive marker
UH	Frequency of interjections
VB	Frequency of base form verbs
VBD	Frequency of past tense verbs
VBG	Frequency of present participle or gerund verbs
VBN	Frequency of past participle verbs
VBP	Frequency of present tense verbs with plural 3rd person subjects
VBZ	Frequency of present tense verbs with singular 3rd person subjects
WDT	Frequency of WH-determiners
WP	Frequency of WH-pronouns
WP$	Frequency of possessive WH-pronouns
WRB	Frequency of WH-adverbs
baseform	Frequency of infinitive verbs (base form verbs preceded by â€œtoâ€□)
Quotes	Frequency of quotation pairs in the entire article
questionmarks	Frequency of questions marks in the entire article
exclamationmarks	Frequency of exclamation marks in the entire article
fullstops	Frequency of full stops
commas	Frequency of commas
semicolon	Frequency of semicolons
colon	Frequency of colons
ellipsis	Frequency of ellipsis
pronouns1st	Frequency of first person pronouns (personal and possessive)
pronouns2nd	Frequency of second person pronouns (personal and possessive)
pronouns3rd	Frequency of third person pronouns (personal and possessive)
compsupadjadv	Frequency of comparative and superlative adjectives and adverbs
past	Frequency of past tense verbs with 1st and 2nd person pronouns
imperative	Frequency of imperative verbs
present3rd	Frequency of present tense verbs with 3rd person pronouns
present1st2nd	Frequency of present tense verbs with 1st and 2nd person pronouns
sentence1st	First sentence class
sentencelast	Last sentence class
txtcomplexity	Text complexity score

3. Model and Analysis of Results

In this part, we will use Support Vector Machine (SVM), K Nearest Neighbor (KNN) algorithm, and an integrated algorithm to construct a classification model to accurately distinguish the objectivity and subjectivity of sports news articles. Support Vector Machine is a powerful supervised learning algorithm that constructs hyperplanes in high-dimensional space for classification with good generalization ability. K Nearest Neighbor algorithm, on the other hand, performs classification based on the closest neighboring training samples, and its simplicity and intuitionistic nature makes it perform well in some cases. The integration algorithm, on the other hand, combines the predictions of multiple classifiers to improve the overall classification performance. Through the application of these algorithms, we will be able to deeply analyze the objectivity characteristics of sports news articles and evaluate the performance of different algorithms on the classification task. This will provide important clues to our understanding of the metrics of objectivity in news reporting and provide useful insights for future media regulation and self-media content auditing.

3.1. SVM Classification Model

In this study, we will use the Support Vector Machine (SVM) algorithm to construct a classification model to accurately distinguish the objectivity and subjectivity of sports news articles. SVM is a powerful supervised learning algorithm that constructs hyperplanes in high-dimensional space for classification with good generalization ability. The key idea of the algorithm is to achieve classification by finding the optimal hyperplane that can effectively divide the samples of different categories, and at the same time maximize the interval between the samples, so that the model has strong robustness. Compared to other classification algorithms, SVM performs well in handling small samples, nonlinearly differentiable and high-dimensional data, and is not susceptible to overfitting. In addition, SVM is able to utilize the kernel function to map samples to higher dimensional spaces for classification, thus improving the applicability and performance of the model. By applying the SVM algorithm, we will be able to dig deeper into the objectivity features of sports news articles and evaluate their performance in classification tasks. This will provide important empirical support for our research, which will help to understand the objectivity of sports news in a deeper way and provide useful insights for future media regulation and self-media content auditing. The subsequent discourse delineates the intricate modeling process employed in the SVM algorithm [2-3].

Sample data set:

$$S = \{(x_1, y_2), \cdots, (x_i, y_i)\} \in (X \times Y)^n.$$

$$min \frac{1}{2}\|W\|^2 \ \ s.t. \ \ y_i(W^T x_i + b) \geq 1, i = 1,2, \cdots n.$$

Where b is the threshold value and W^T is the transpose of the weight vector W. If S is linearly inseparable, it can be solved by adding a new constraint:

$$min \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n} \varphi_i \ \ s.t. \ \ y_i(W^T x_i + b) \geq 1 - \varphi_i, (\varphi_i \geq 0, i = 1,2, \ldots n),$$

Where φ_i is the relaxation variable and C is a constant. In the following, with the help of Lagrangian duality, the dual problem equivalent to the above problem is solved, given the Lagrangian multiplier vector, which has the dual expression:

$$max \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \ \ s.t. \ \ \sum_{i=1}^{n} a_i y_i = 0, a_i \geq 0, i = 1,2, \ldots n.$$

$$max \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \ \ s.t. \ \ \sum_{i=1}^{n} a_i y_i = 0, 0 \leq a_i \leq C, i = 1,2, \ldots n.$$

$$W = \sum_{i=1}^{n} a_i y_i x_i, b = y_i - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_j.$$

The above two formulas show that only one α variable is included, and both W and b can be found by α. Vapnic introduced the kernel function $k(x_i, x_j)$by using kernel space theory, and then the above two equations are expressed as a pairwise equation:

$$max \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \ \ s.t. \ \ \sum_{i=1}^{n} a_i y_i = 0, a_i \geq 0, i = 1,2, \ldots n.$$

$$max \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \ \ s.t. \ \ \sum_{i=1}^{n} a_i y_i = 0, 0 \leq a_i \leq C, i = 1,2, \ldots n.$$

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \, k(x_i, x) + b, b = y_i - \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j).$$

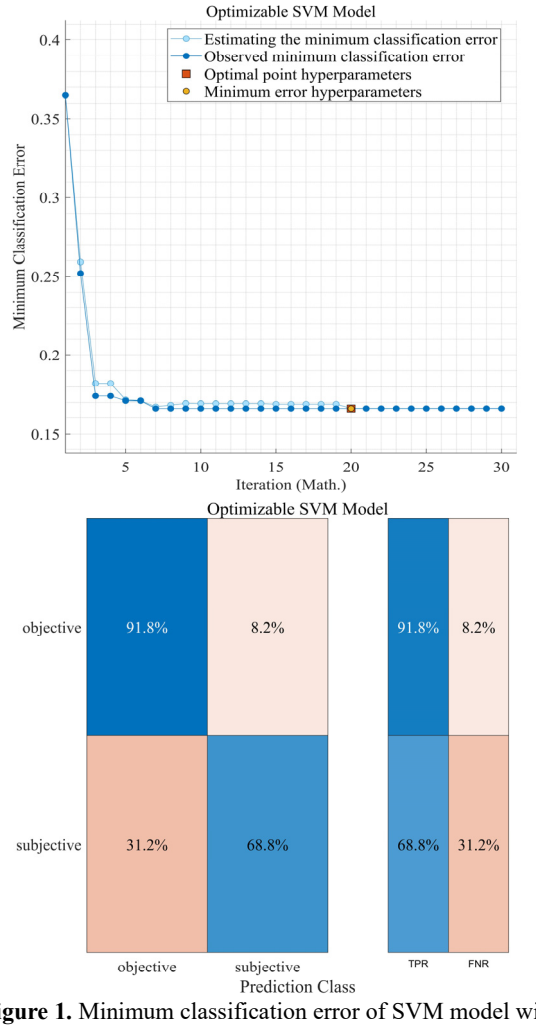

Figure 1. Minimum classification error of SVM model with confusion matrix

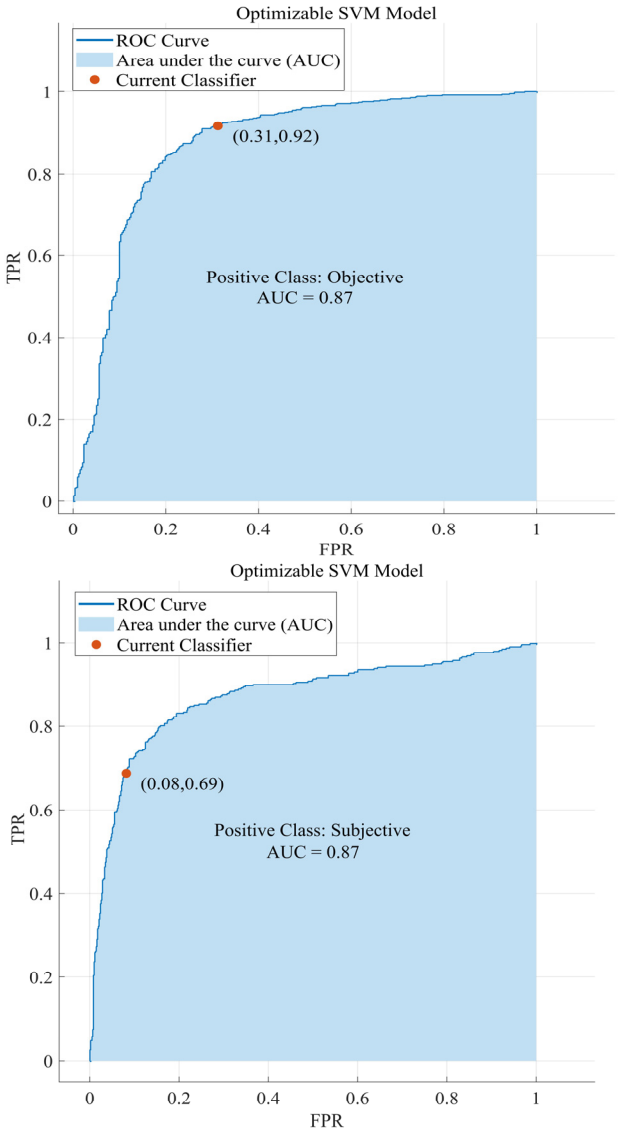

Figure 2. ROC curve for the SVM model.

3.2. KNN Classification Algorithm

In this part, we will use the K-Nearest Neighbors (KNN) algorithm to construct a classification model to accurately distinguish the objectivity and subjectivity of sports news articles. The KNN algorithm is a simple and intuitive supervised learning algorithm, whose basic principle is to classify samples based on the distance between them. Specifically, when classifying a new sample, the KNN algorithm finds the closest K neighboring samples in the training set and classifies the new sample into the category that occupies the majority of these K neighbors. This neighbor-based classification approach makes the KNN algorithm perform well in certain situations, especially when the dataset has complex decision boundaries.

One of the advantages of the KNN algorithm is that it does not need to train the model explicitly, but instead utilizes the training data directly when performing classification. This makes the KNN algorithm highly flexible and simple for dealing with multiclass problems and nonlinear datasets. In addition, the KNN algorithm makes no assumptions about the data distribution and is robust to outliers and noise, and thus performs well in some practical application scenarios. However, the performance of the KNN algorithm is also affected by the choice of K-value and the distance metric, which requires appropriate tuning in practical applications.

By applying the KNN algorithm, we will be able to dig deeper into the objectivity features of sports news articles and evaluate their performance in classification tasks. This will provide a more comprehensive and diverse perspective for our research, which will help to better understand the objectivity of sports news and provide a more comprehensive consideration for future media regulation and self-media content auditing [4-6].

The specific modeling process is as follows:

Step 1. Calculate the distance of each sample point in the training and test samples (common distance metrics are Euclidean distance, Mahalanobis distance, etc.);

Step 2. sort all the above distance values;

Step 3. select the first k samples with the smallest distance;

Step 4. vote based on the labels of these k samples to get the final classification category.

Input training dataset:

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}.$$

Where x_i is the feature vector, y_i is the category, $i = 1, 2, 3, \ldots \ldots N$.

Based on the given distance metric, find the nearest neighbor to x in the training set T. The neighborhood of x covering this k point is denoted as $N_k(x)$.

Decide the category of x in $N_k(x)$ based on a categorical decision rule (e.g., majority decision) y.

$$y = \underset{c_j}{argmax} \sum_{x_i \in N_k(x)=1} I(y_i = c_j)$$

$$i = 1, 2, 3, \ldots \ldots N;\ j = 1, 2, 3, \ldots \ldots K$$

Where I is the indicator function and instantly $y_i = c_j$ is 1, otherwise I is 0.

There are various common distance measures in distance class models such as KNN. Such as Euclidean distance, Manhattan distance, Minkowski distance, Chebyshev distance and cosine distance. In this algorithm cosine distance is used:

$$cos(\theta) = \frac{\Sigma_{i=1}^{n} x_{ia} x_{ib}}{\sqrt{\Sigma_{i=1}^{n} x_{ia}^2} \sqrt{\Sigma_{i=1}^{n} x_{ib}^2}})$$

Figures 3 to 4 illustrate the model visualization results. According to the visualization results, the classification accuracy of the KNN algorithm is 83.4%.

3.3. Ensemble Classification Algorithms

In this section, we will model the problem using the integration algorithm. For such a dataset, the integration algorithm shows some advantages in some aspects. The core of the integrated algorithm lies in the "set", that is, combining multiple classifiers together for training and decision making, in order to obtain better model performance and superior training results. In the face of multi-dimensional complex data, the integrated algorithm will make use of the advantages of multiple models to compensate for the limitations of a single classifier, which is conducive to reducing the training variance of the model and improving the generalization ability of the model. For data analysis and modeling in the financial, entertainment, and media fields, integration algorithms are widely favored for their unique data processing advantages. Common integration algorithms include Bagging, Boosting and Random Forests. By applying integration algorithms, we will be able to further improve the accuracy of identifying the objectivity of sports news articles, thus providing more reliable and robust classification models for our research. This will help deepen the understanding of the

objectivity of sports news coverage and provide more depth and breadth of thinking for future media regulation and self-media content censorship. The introduction of the comprehensive algorithm will provide a new perspective for our study and further expand our methods and understanding of sports news objectivity analysis.

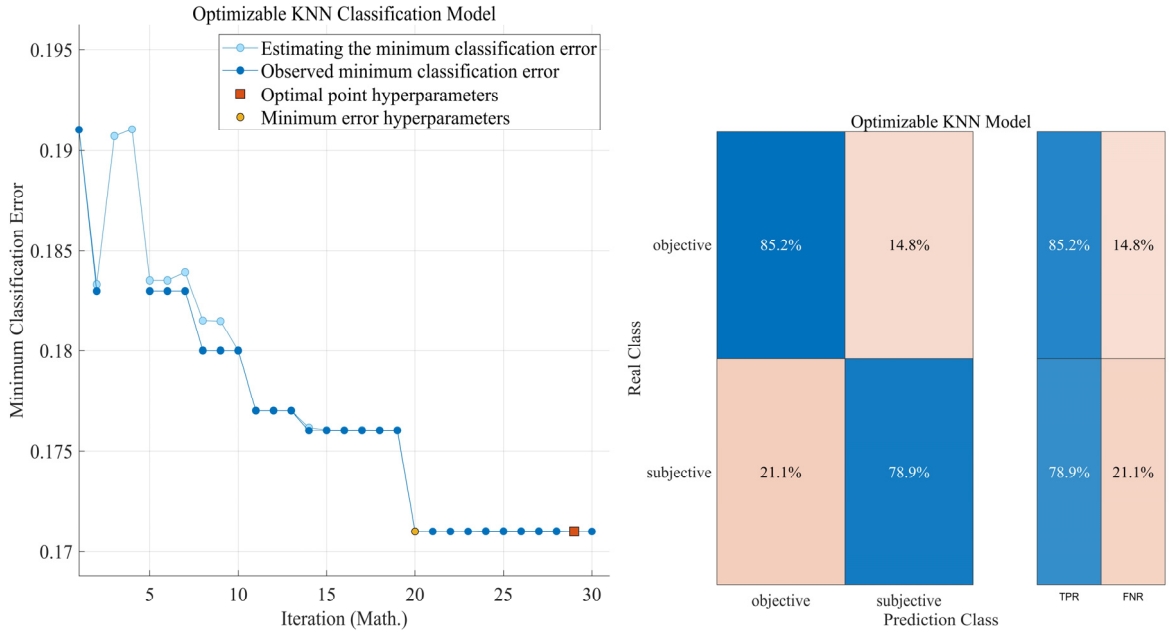

Figure 3. Minimum classification error of KNN model with confusion matrix

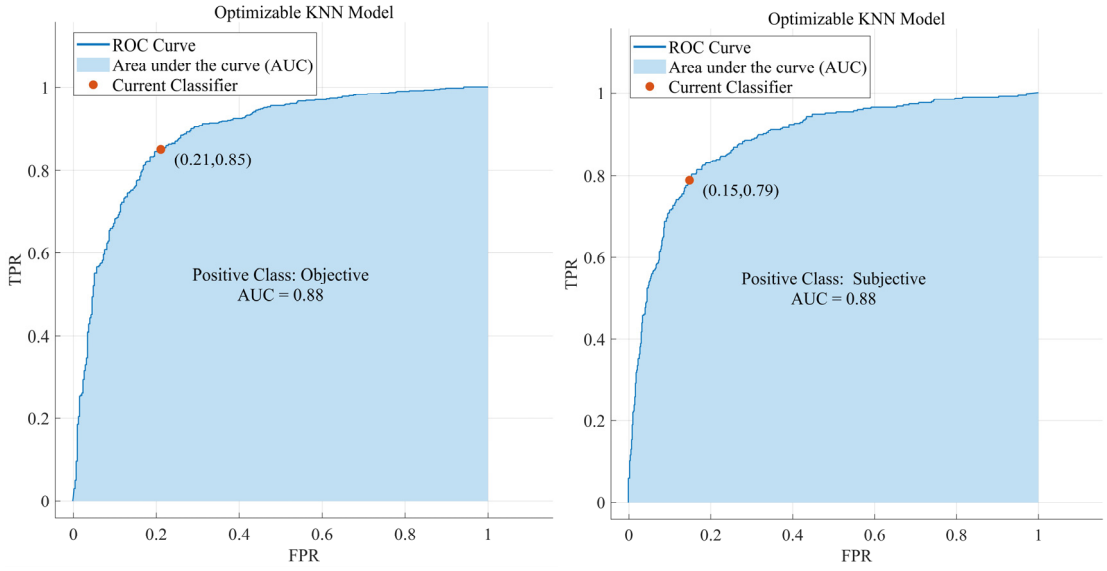

Figure 4. ROC curve for the KNN model

The basic principle of this algorithm is as follows:

Adaboost (adaptive boosting: boosting + single-level decision tree) is a more representative algorithm in boosting, the basic idea is to construct a classifier through the distribution of the training data, and then find out the weight of this if the weak classifier through the error rate, and iteratively carry out by updating the distribution of the training data until reaching the number of iterations or the loss function is less than a certain threshold [7-8].

The training dataset is:

$$S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}.$$

Where there is $Y_i = \{-1, 1\}$.

Step 1. Initialize the distribution of the training data; the weights of the training data are equally distributed $D = \{W_{11}, W_{12}, W_{13}, W_{14}, W_{15}\}$, $W_{1i} = \frac{1}{N}$.

Step 2. Selection of basic classifiers. After the classifier has been selected, minimizing the classification error allows the parameters to be derived.

Step 3. Calculate the coefficients of the classifier and update the data weights; the error rate can also be derived as e_1. The coefficients of this classifier can also be derived. The basic Adaboost gives the formula for the coefficients as:

$$\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}.$$

The training weight distribution is then updated to:

$$D_{m+1} = \{w_{m+1,1}, \cdots, w_{m+1,i}, \cdots, w_{m+1,N}\}.$$

$$w_{m+1,i} = \frac{w_{m,i}}{Z_m}\exp\left(-\alpha_m y_i G_m(x_i)\right).$$

The normative factor is $Z_m = \sum i = 1^N w_{m,i}\exp\left(-\alpha_m y_i G_m(x_i)\right)$.

The combination of classifiers is derived as:

$$f(x) = \sum_{i=1}^{N} e\, xp\left(-\alpha_m y_i G_m(x_i)\right).$$

It can be shown mathematically that the AdaBoost method

is the process of continuously fitting a strong learner and minimizing the objective function $J(F) = E\left(e^{-yF(x)}\right)$ using some optimization method.

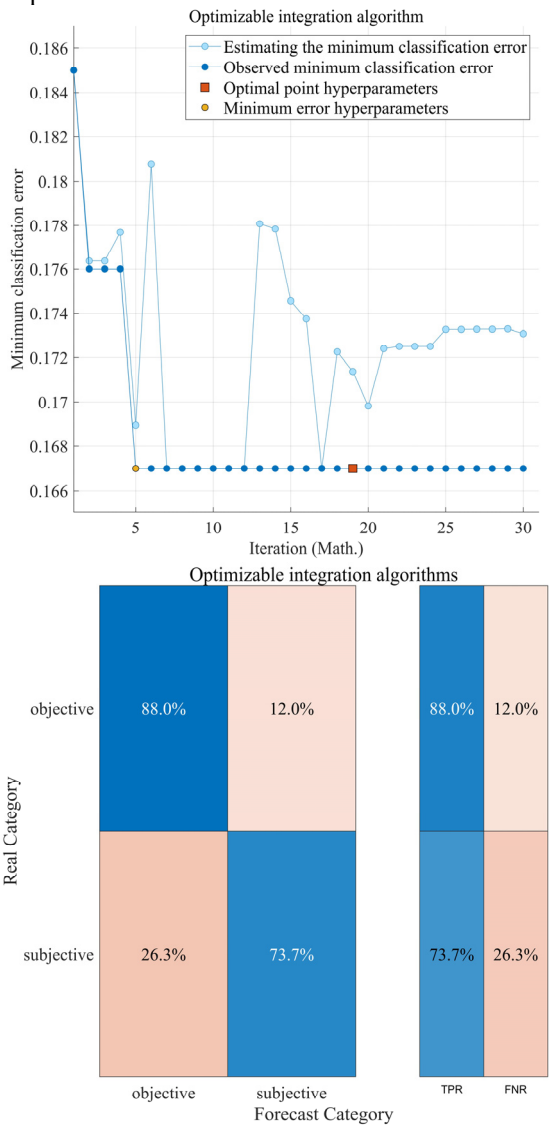

Figure 5. Minimum classification error of integration algorithm model with confusion matrix

Figures 5 to 6 illustrate the model visualization results. According to the visualization results, the classification accuracy of the integration algorithm is 82.8%.

Based on the above results of model building and training, the classification accuracy of SVM algorithm is 83.3%, the classification accuracy of KNN algorithm is 83.4%, and the classification accuracy of integrated algorithm is 82.8%. The evaluation index of classification accuracy demonstrates the performance of these algorithms in objectively classifying sports news and provides an important basis for understanding the performance of different algorithms. However, it is not sufficient to evaluate algorithm performance based on prediction accuracy alone. When considering algorithm selection, we also need to consider the model's generalization ability, robustness, and adaptability to different data features. Not only that, factors such as computational complexity and training time of the algorithm are also important considerations. From the current results, although the KNN algorithm is slightly higher than the SVM classification algorithm and the integration algorithm in terms of prediction accuracy, we still need to focus on how to further improve the classification accuracy of the KNN algorithm in the follow-up, and further explore why the integration algorithm, which is highly adapted to the financial and media data, exhibits the lowest training effect under this problem.

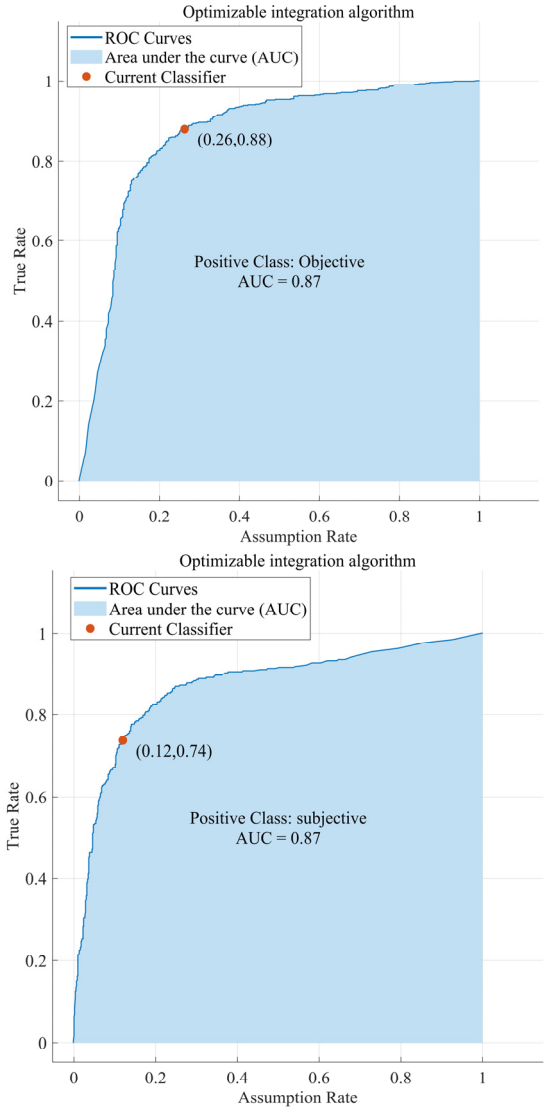

Figure 6. ROC curve for the integration algorithm model.

4. Results and Discussion

To address the issue of whether the self-media platform has objectivity to the published content in the era of artificial intelligence, this study utilizes a variety of machine learning algorithms for comparison, using a large amount of data for training, and concludes that the machine learning algorithms have shown excellent performance and effectiveness in this issue. the classification accuracy of SVM algorithm, KNN algorithm, and integrated algorithm are all over 80%, with the KNN algorithm having the highest classification accuracy of 83.4%. algorithm has the highest classification accuracy of 83.4%. Undoubtedly, such results are satisfactory. However, there are still some problems that have not been discussed in this paper, that is, how to further improve the prediction accuracy and how to avoid some shortcomings of the KNN algorithm in future classification problems. In addition, for the integration algorithm, this kind of entertainment as well as media data tends to show high adaptability as well as classification accuracy, why the effect in this environment is not dominant, and what other methods can be further

optimized? What new self-published content will need to be analyzed objectively in the future? This is an area of great interest for the authors of this paper in the future.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

I would like to express my sincere gratitude to the editors and reviewers for their helpful comments in improving the quality of the original manuscript.

References

[1] Rizk, Yara and Awad, Mariette. (2018). Sports articles for objectivity analysis. UCI Machine Learning Repository. https: // doi. org/10.24432/C5801R.

[2] Chen, R.; Herskovits, E.H. Machine-learning techniques for building a diagnostic model for very mild dementia. Neuroimage, 2010, 52, 234–244.

[3] Raza, M.; Awais, M.; Ellahi, W.; Aslam, N.; Nguyen, H.X.; Le-Minh, H. Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques. Expert Syst. Appl. 2019, 136, 353–364.

[4] Hajj, N., Rizk, Y. & Awad, M. A subjectivity classification framework for sports articles using improved cortical algorithms. Neural Comput & Applic 31, 8069–8085 (2019). https:// doi.org/10.1007/s00521-018-3549-3.

[5] Young Joon Park, Hyung Seok Kim, Donghwa Kim, Hankyu Lee, Seoung Bum Kim, Pilsung Kang, A deep learning-based sports player evaluation model based on game statistics and news articles, Knowledge-Based Systems, Volume 138, 2017.

[6] Jeronimo, Caio Libanio Melo, et al. "Fake news classification based on subjective language." Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. 2019.

[7] Rahman, Moqsadur, Summit Haque, and Zillur Rahman Saurav. "Identifying and categorizing opinions expressed in Bangla sentences using deep learning technique." International Journal of Computer Applications 176.17 (2020): 13-17.

[8] Nassif, Ali Bou, et al. "Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities." Applied Soft Computing 98 (2021): 106836.