International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DETECTION OF CONFIRMATION BIAS IN HOROSCOPE TEXTS USING RANDOM FOREST CLASSIFIER

Arun Padmanabhan[1], Dr. K. Devasenapathy[2]

[1]Research Scholar, Department of Computer Science
Karpagam Academy of Higher Education
Coimbatore, India
arun1986.p@gmail.com

https://orcid.org/0009-0008-9987-2664

[2]*Associate Professor, Department of Computer Science*
Karpagam Academy of Higher Education
Coimbatore, India

drdevasenapathy.k@kahedu.edu.in

https://orcid.org/0000-0003-3690-3239

Abstract

The Random Forest algorithm, a widely used machine learning technique, is utilized in this research to identify confirmation bias in horoscope text. A common feature of horoscopes, confirmation bias is the tendency to interpret vague or general information in a way that reinforces already established beliefs. Horoscope excerpts having labeled dataset was analyzed using a combination of TF-IDF vectorization, n-gram analysis, linguistic markers associated with bias like vague terminology and feature oriented language. The model's effectiveness was evaluated using the standard metrics such as accuracy, precision, recall and F1 score. Performance of Random Forest was evaluated and compared to a baseline classifier model using Logistic Regression. Results showed that the proposed Random Forest model clearly outperformed the baseline model Logistic Regression, achieving an accuracy of 98.74%, precision of 91%, recall of 72%, and F1-score of 80% in distinguishing biased from unbiased texts. These findings demonstrated the efficiency of machine learning models to detect confirmation bias in text and contribute to advancing both computational linguistics and behavioral psychology

The research also examined the difficulties associated with subjective annotation in detecting bias, emphasizing the difficulty of defining and classifying confirmation bias in written material. The ability to handle large amounts of data and thereby to provides information about the features to help to identify biases is high using Random Forest algorithm compared to the baseline model Logistic Regression. Machine learning can be utilized to detect psychological biases in text, particularly in fields where language is deliberately ambiguous and susceptible to interpretation. Based on the findings, it is very evident that these techniques could be

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

associated with a variety of tasks includes study on Natural Language Processing, sentiment analysis, and misinformation identification. It advances the disciplines of behavioral psychology and computational linguistics by offering a methodical way to comprehend cognitive biases through automated text analysis.

Keywords— Confirmation bias, Machine learning, Random Forest, Natural Language Processing, Horoscope, Bias detection, Artificial Intelligence.

1. Introduction

Cognitive biases have been a major area of study in psychology for many years, with specialized interest in how people perceive and interpret information. Various forms of bias, including confirmation bias [Nickerson, 98] and other biased practices, are pervasive across different areas such as decision-making, belief systems, and media coverage. Confirmation bias is the tendency of people [Klayman, 87] to favour information that corresponds with their existing beliefs or hypotheses, often ignoring evidence that contradicts their views. It is not confined to explicit reasoning but also manifests in less obvious ways, like how individuals interact with written texts, particularly those that offer vague, generalized statements open to interpretation. Despite the growing interest of detecting cognitive bias in textual data, most of the previous work has been done on news articles, social media text rather than on the understudied domain of horoscope text.

In order to explore the existence of confirmation bias, the study of horoscopes [Allum, 2010] is an ideal domain as they frequently engage vague language which are both inclusive and accessible, enables the readers to interpret them in according to their belief or experiences. Horoscopes are basically predictions of advices given on the basis of astrological signs – sometimes with vague and general language. The use of this language is a perfect illustration of confirmation bias as since people frequently try to interpret these predictions in a way to support their belief or expectations. There is also a tendency to what some would call confirmation bias within horoscopic texts, due to the use of generalized language and predictive statements that can be interpreted by many people in varying ways. Using these features in mind, the investigation of detecting the existence of confirmation bias using machine learning models [Stenning, 2008] offers an excellent platform. The identification and analysis of confirmation bias in horoscope text using machine learning and natural language processing offers a new insight into this area. As since the Random Forest algorithm (RF) is one of the most popular machine learning algorithms, it has been proven to be the most successful hen dealing with large and complicated datasets.

This study has been performed using a Random Forest algorithm to create a standardized models that can differentiate between horoscope texts that have the existence of confirmation bias or not. Confirmation bias is evident in key linguistic traits like sentiment, unclear vocabulary and future-oriented language that are extracted from the process [Eshan, 2017]. After that, this is applied to a dataset of labelled horoscope texts and sorted according to how likely it is that these texts support or contradict conventional wisdom. In addition to advancing computational linguistics, this work sheds light on the potential applications of machine learning in behavioural

Received: July 24, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

psychology. This study seeks to bring together advanced machine learning techniques and psychological theories to offer a fresh perspective on cognitive biases, contributing to the advancement of both NLP-based and conventional psychological research.

2. Existing Work

Confirmation bias detection in varied contexts has been explored by prior studies, with notable emphasis on online news and social media. For instance, stance-aware text representation and network propagation modeling are integrated by Confirmation Bias-Aware Fake News Detection with Graph Transformer Networks [Soga, 23] to improve fake news detection accuracy. In this approach, stance analysis is performed by fine-tuning BERT to produce user stance embeddings, which are then combined with a Graph Transformer Network to capture stance similarity and bias-aware propagation patterns. Although news domains are focused on by this work, insights into developing robust bias detection systems are offered by the integration of natural language processing and bias-aware network modeling, which can be adapted to specialized domains such as horoscope texts.

Simpler classifiers such as Logistic Regression, Support Vector Machines were relied highly for traditional works whereas recent research has now moved on towards deep learning approaches. Investor behavior in financial markets was investigated [Gupta, 24] using Natural Language Processing and Machine Learning, given more emphasis on confirmation bias in sentiment driven trading decisions are being highlighted. This study unleashed that behavioral biases traditionally studied only in psychology could be quantified by machine learning algorithms. Similarly for Chinese News Articles, stance detection models [Lin, 25] using NLP which indirectly captured confirmation bias by analyzing stance shifts in biased reporting.

In the medical domain, NLP combined with machine learning was leveraged [Ma, 24] to analyze drug interaction records, showed biased reporting in case data can mislead analysis. It is demonstrated by this work how bias detection extends beyond media and finance to healthcare.

BERT and Latent Dirichlet Allocation (LDA) were applied [Li, 25] for detecting fake review identification in e-commerce. Linguistic patterns similar to confirmation bias in horoscopes – vague, optimistic and emotionally charged statements are often used by biased or manipulative reviews.

[Lezama, 25] Proposed a multi label deep learning model for harassment and discrimination detection. The ability of neural networks to capture linguistic bias is emphasized by their work, which can be repurposed for detecting confirmation bias in horoscope texts.

This trend extended by [Hunko, 25] via CNN and sentiment analysis in mobile app testing and AR-based mental health, respectively showcasing the flexibility of NLP techniques for bias prone text data. Similarly, for speech recognition in air traffic communications, contextual knowledge enhanced NLP was also proposed [Guo, 25] showing improvements in biases in spoken text.

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Dataset Description

A. Data Source and Composition

The dataset was obtained from an open-source repositories on the Hugging Face Hub, ensuring transparency and integrity [Liu, 24]. The original dataset comprised of 20000+ labelled horoscope entries with columns like Instruction (user request), Response (Horoscope text). The data used in this study is based on two key elements: the Response and label (indicator of bias). A binary value to each response is assigned by manual annotation to obtained the third column label. In the context of binary variables, '1' denotes confirmation bias and zero signifies absence. Each entry in the dataset includes the following attributes:

Response – The textual content of the horoscope.

Label (binary) – A manually annotated label indicating the presence (1) or absence (0) of confirmation bias in the horoscope text.

The dataset description outlined here that 38% of the entries had confirmation bias (label = 1) and the remaining 62% of the entries were unbiased (label = 0). The length of horoscope texts in the dataset ranges from 6 to 70 tokens, having an average of 2 tokens. The stylistic variations of the horoscope writers are reflected in this variation.

Although systematic, manual annotation introduces subjectivity. In order to tackle this, Cohen's Kappa (k=0.78) which indicates substantial inter annotator agreement [Cohen, 60] was utilized to assess the annotation reliability. However, the classification of borderline cases similar to "You may face unexpected change" was still a challenging one due to the psychological aspect like confirmation bias

B. Data Cleaning and Preparation

Data Cleaning and Preparation phase begins with removing irrelevant column (Instruction) from the dataset as since it doesn't serve any analytical purpose. Specifically, the first column was removed using positional indexing to ensure that the feature set contained only relevant variables. Next, in order to validate the remaining schema and to preview a subset of instances for quality verification, the dataset was inspected. Examining the existence of potential class imbalance in the distribution of the target variable was also performed. All the records having null values were eliminated to maintain dataset integrity by conducting a missing value assessment. In the text preprocessing stage, the primary horoscope statements in the response attribute were normalized to lowercase, and leading or trailing whitespace was removed. Lexical uniformity was ensured prior to vectorization and feature extraction in the subsequent stages.

Step	Description	Example Transformation
Lowercasing	Convert all text to lowercase	"You Will Shine" → "you will shine"

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Tokenization	Split text into words / tokens	"good news awaits" → [good, news, awaits]
Stop word Removal	Remove common uninformative words	"you will be happy" → "happy"
Lemmatization	Reduce words to base forms	"changes, changing" → "change"
Punctuation Removal	Eliminate irrelevant characters	"luck!!!" → "luck"
Class Balancing	Oversample minority class	Add biased samples until balanced

Table 1: Preprocessing Pipeline

Horoscope Text	**Label**
"Good fortune will come if you remain patient."	1 (Biased)
"You will likely face delays in your work schedule."	0 (Unbiased)
"Something special awaits you; your hard work pays off."	1 (Biased)
"The temperature will rise in your city this week."	0 (Unbiased)

Table 2: Sample Horoscope Entries with Labels

4. Model Description / Methodology

C. Data Preprocessing

Dataset once preprocessed was later partitioned into training and testing subsets using the train_test_split function from the Scikit-learn library. For the purpose of model training, 80% of the samples were used and the remaining 20% of the samples were reserved for the purpose of performance evaluation. Stratified sampling was employed based on the target label to

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

preserve the original class distribution in both the subsets. To ensure reproducibility of the split across experimental runs, a fixed random seed (random_state = 42) was utilized.

Text preprocessing is essential in order to prevent the dominance of classification performance by redundant and noisy features. For example, in the absence of stop word removal, the TF-IDF matrix would be dominated by commonly used but informative tokens such as "will" or "you" which may lead to feature sparsity and model performance. Likewise, interpretability and classifier performance were improved by lemmatization via combining morphological variants (for example, "changing", "changes" to "change").

D. Feature Extraction Techniques

Capturing of patterns that indicate the presence of confirmation bias in horoscope texts [Gallimore, 96] has been done using feature extraction process. TF-IDF (Term Frequency - Inverse Document Frequency) [Qaiser, 18] were used to generate text-based features to identify the most significant words in each horoscope. This model relies completely on the power of TF-IDF-weighted lexical terms. This is important since confirmation bias is often associated with positive reinforcement and vague, optimistic statements. Furthermore, linguistic markers such as overgeneralization, vague terminology, and emotionally charged words are identified as potential indicators of bias. TF-IDF features derived directly from the horoscope text were given as the input for the Random Forest Classifier.

Because of its interpretability and ease of use, TF-IDF representation was selected. TF-IDF provides sparse high-dimensional vectors that accurately reflect token frequency while excluding overly used common terms unlike being used by Word2Vec or BERT. Use of TF-IDF is highly useful particularly for capturing vague horoscope text, where tokens such as "fortune" or "success" may strongly indicate bias. The weight for each token was calculated as follows:

$$w_{\{t,d\}} = tf_{\{t,d\}} \times \backslash log \backslash frac\{N\}\{df_t\}$$

$$where \ \ tf_{\{t,d\}} \ \ is \ the \ frequency \ of \ term \ t \ in \ document \ d,$$

$$and$$

$$df_t \ is \ the \ number \ of \ documents \ containing \ t.$$

E. Handling Data Imbalance

A notable class imbalance was identified in the constructed dataset with unbiased label (0) dominates over the biased label (1). This disproportion in the input class develops a biased approach towards predicting the majority class, thereby reducing its capacity to detect biased labels (1) in the horoscope text. Hence, addressing the class imbalance was therefore integral to maintain validity and robustness of the classification. RandomOverSampler method from the imbalanced - learn library technique was used to balance the class to improve the model sensitivity. After the TF-IDF feature vectorization step, oversampling was applied and was restricted to the training data to prevent information leakage into the training phase. In order to

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ensure that the original class ratio was proportionally maintained across both subsets, the training-testing split was performed using as stratified sampling strategy thereby enabling unbiased estimation of model generalization performance. The overall impact of this imbalance handling strategy was reflected in the classifier's improved recall for the biased class. [Elreedy, 19].

Although the dataset is effectively balanced by the RandomOverSampler, it is crucial to note that certain limitations can be introduced by oversampling strategies. In particular, by simply duplicating the minority class examples, the risk of overfitting is increased, as the classifier may learn the repeated samples rather than the learning generalized patterns [He, 09]. However, with the use of TF-IDF in the present study, this limitation is mitigated because the TF-IDF feature space is sparse and high dimensional, and replicated observations are well handled by the random forest classifiers compared to linear models [Alam, 21].

SMOTE (Synthetic Minority Oversampling Technique) or ADASYN (Adaptive Synthetic Sampling) have been well studied in handling imbalance [Chawla, 02]. By interpolation, these methods create synthetic minority samples among existing samples, which is suited best for numeric feature spaces. However, in text data, semantic structure can be distorted and noise can be introduced by artificially generated vectors, reducing the classification performance [Liu, 19]. Thus RandomOverSampler, a more reliable and interpretable approach for this specific application was considered.

Finally for this study, it was ensured that both the biased and unbiased class were proportionally represented in evaluation through stratified sampling in train test splitting. This method avoids misleadingly optimistic results that could arise from uneven class splits. Similar findings have been shown in prior studies where recall for minority classes was highly improved by combining stratified sampling with resampling method while maintaining overall model stability [Elreedy, 19].

F. Model Trainng and Optimization

To classify the horoscope text, the random forest classifier was trained on the processed dataset. It is since the ability to handle high dimensional data [Ishwaran, 11] and its robustness against overfitting, Random Forest algorithm has been chosen. The parameter class weight = 'balanced' was instantiated in the Random Forest Classifier. With this setting, based on the observed frequency, class weights were automatically adjusted. It helped address skewed class distribution and placed a heavier penalty on misclassifying minority labels. In order to ensure reproducibility, Random Forest model employed 100 estimators with a fixed random state = 42. The main focus on class Recall for the biased category, optimization focused on balancing overall classification performance with minority class sensitivity. The balanced training set was used to fit the RF model via the standard scikit-learn. fit () method.

The Random Forest algorithm is an ensemble method that is characterized by the combination of multiple decision trees through a process of bootstrap aggregation (bagging) and random feature selection [Brieman, 01]. A bootstrapped subset of the data is used for

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

training each tree, while a random subset of features is considered at each split. Correlation between trees is reduced by this dual randomization, thereby resulting in lower variance and improved generalization. Such diversity among trees is considered particularly important when high-dimensional TF–IDF features are being worked with, where correlations between terms are frequent and could otherwise lead to overfitting [Iswaran, 07].

An appropriate configuration for the classifier was determined through hyperparameter tuning. Candidate values were different among parameters including the number of estimators (n_estimators), maximum tree depth (max_depth), and minimum samples per split (min_samples_split). In order to systematically evaluate the different settings, a 5- fold cross validation was also employed. This method identified that 100 – 200 estimators with a maximum depth of 15 – 20 produces optimal performance by balancing recall on the minority class with computational efficiency. The use of cross-validation ensured that the observed results were not artifacts of a single split but reflected consistent trends across folds [Han, 22].

Baseline experiments were also conducted with Logistic Regression to contextualize the choice of Random Forest Classifier. While both the models achieved a reasonable accuracy, their recall score on minority score were notable lower consistent with prior reports on ensemble superiority in imbalanced text classification tasks. In contrast, both competitive accuracy and feature important scores were provided by Random Forest Classifier model, which highlighted key lexical cues leading to biased predictions.

G. Optimization & Selection

Optimization was centered on balancing overall classification problem with minority class sensitivity due to the major focus on class recall for the biased category. The core optimization strategy emerged from the oversampling and class-weighting measures described above, eliminating the need for more computationally intensive parameter searches. The stability of RF performance across a range of hyperparameter configurations observed during exploratory trials reinforced this decision.

The main objective of the study was to detect confirmation bias with few false negative as possible by prioritizing recall for the biased class decision. In real world applications like bias detection, failing to capture a biased statement is more damaging that missing out the unbiased sentences [He, 09]. Therefore, optimization aimed at achieving a balance between precision and recall as represented by the F1 Score, rather than maximizing the accuracy alone. Prior studies performed on text-based classification justifies this approach of handling a balance between recall and precision to safeguard against missing out unbiased categories [Elreedy, 19].

During the optimization process, another factor under consideration was the optimization of the relative stability of Random Forest performance across different parameter ranges. RF produces consistent results in contrast to Support Vector Machine or Gradient Boosting even with not finely adjusted hyperparameters [Ishwaran, 08]. Experiments conducted using different estimators (100-300), tree depths (10-30) and class weights confirmed that the performance of the model stayed within a narrow performance band. Rather than Bayesian

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Hyperparameter searches or exhaustive grid which would have been more computationally costly without yielding any significant performance gain, this robustness enabled the adoption of a practical approach focused on oversampling and class reweighting [Han, 22].

Finally, both the empirical performance and interpretability served as a criterion for selecting the final Random Forest configuration. Transparency and reproducibility would been an issue if complex models such as neural networks or boosted ensembles had considered. Random Forest not only ensured competitive results but it also enabled the feature importance analysis, providing information on the linguistic cues most indicative of confirmation bias. These strong characteristics showcasing a dual performance made Random Forest the most suitable model for deployment in this context [Breiman, 01].

H. Model Fitting

In this study, the usual scikit.learn.fit () method was used to fit the RF model to the balanced training set. In order to prevent any inadvertent data leakage, Out-of-Bag (OOB) error estimation was considered, but later avoided in favor of a dedicated held-out test set for evaluation.

Using bootstrap sampling for every decision tree and aggregating the outcomes across the ensemble, the Random Forest classifier was trained using the scikit-learn library's standard fit () method. By exposing each tree to a marginally different sample of the training data, this procedure lowers the chance of overfitting and increases model diversity. Both biased and unbiased samples were proportionately represented during fitting thanks to the input from the balanced training set that was produced during oversampling.

During experimentation, the initial consideration was made towards Out-of-Bag (OOB) error estimation as an internal validation strategy. OOB error uses the unused samples from bootstrap resampling as a quasi-validation set for each tree, providing an unbiased estimate of generalization performance without a separate split [Breiman, 01]. While it is found that this approach computationally efficient and widely used, but ultimately did not adopt it in this study to maintain consistency with standard machine learning evaluation pipelines, which often prefer a dedicated held-out test set. This decision avoided potential overlap between training and evaluation samples, ensuring that reported metrics reflected true model generalization [Ishwaran, 08].

The choice of a held-out test set clearly separates the training, validation, and evaluation stages. Cross-validation on the training set for hyperparameter selection and fit the final model on the entire balanced training dataset before evaluating it on the independent test set was conducted. This structured pipeline reduces the risk of overfitting to validation folds and increases the reproducibility of experimental results, which is critical for applying machine learning to psychological constructs like confirmation bias detection [Han, 22].

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

I. Evaluation and Comparative Analysis

Multiple classification metrics were used including accuracy, precision and F1-score to evaluate the model's performance. Biased and non-biased texts were effectively analyzed using the ROC-AUC curve to measure the model's ability [Bradley, 97]. A Confusion matrix was also generated to provide interoperability into classification error types [Van, 22].

Finally, to identify key linguistic patterns associated with confirmation bias, the results were analyzed. To improve the model's accuracy, unclassified cases were also examined in this study. The performance of Random Forest model was compared with baseline classifier such as Logistic Regression to validate the findings [Araque, 17]. This makes sure that the chosen model provides the best balance between accuracy and interpretability to for detecting the confirmation bias in horoscope text.

To ensure a comprehensive evaluation of the classifier, this model adopted multiple performance metrics. While accuracy is widely used, it can mislead in imbalanced datasets since a trivial classifier predicting only the majority class can still achieve high values. Therefore, a strong emphasis on precision, recall, and F1-score, which provide a more nuanced assessment of classification effectiveness. Precision measures the proportion of correctly identified biased samples among all predicted biased samples, while recall assesses the ability to capture all biased instances in the dataset. The F1-score, defined as the harmonic mean of precision and recall, balances these two complementary measures. The equations used are given below:

Precision

$$Precision = TP / (TP + FP)$$

Recall:

$$Recall = TP / (TP + FN)$$

F1-score:

$$F1\text{-}score = 2 * Precision * Recall / (Precision + Recall)$$

Accuracy:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

ROC–AUC:

$$AUC = \int [0 \text{ to } 1] \, TPR(FPR) \, d(FPR)$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively

The analysis used the Receiver Operating Characteristic – Area Under the Curve (ROC–AUC) metric to measure the classifier's ability to distinguish between biased and unbiased classes across varying thresholds. A high ROC–AUC score indicates that the classifier can effectively rank positive instances higher than negative ones, independent of class distribution

Received: July 24, 2025

International Journal of Applied Mathematics
Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[Bradley, 97]. The confusion matrix further clarifies results by quantifying classification errors, highlighting whether the model is more prone to false positives (misclassifying unbiased text as biased) or false negatives (failing to detect actual biased text).

For comparative validation, here benchmarked the performance of the Random Forest model against a baseline Logistic Regression classifier. While Logistic Regression performed reasonably well in terms of accuracy, it achieved significantly lower recall for the minority (biased) class, consistent with known limitations of linear models in handling complex decision boundaries in text classification tasks. By contrast, Random Forest achieved superior balance across precision, recall, and F1-score, confirming its robustness and interpretability for the task of confirmation bias detection in horoscope texts. The comparative analysis reinforced the suitability of ensemble methods for this domain, as they provide both competitive predictive power and the capacity to extract meaningful linguistic patterns through feature importance scores [Alam, 21].

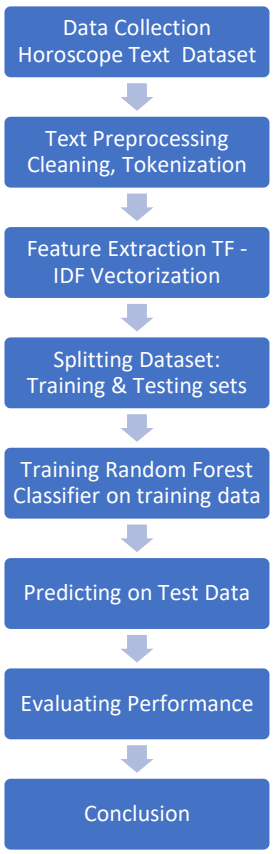

Figure 1: Methodology to detect Confirmation Bias in Horoscope Texts Using Random Forest Classifier.

5. Results and Discussions

For baseline comparison, a Logistic Regression (TF–IDF) model achieved an accuracy of 97.41%, with a precision of 75%, recall of 59%, F1-score of 66%, and ROC–AUC of 0.97. In contrast, the proposed Random Forest model, configured with oversampling and a lowered

Received: July 24, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

decision threshold of 0.3, achieved superior results with an accuracy of 98.74%, precision of 91%, recall of 72%, F1-score of 80%, and ROC–AUC of 0.99. The performance improvement across all evaluation metrics demonstrates the Random Forest's [Svetnik, 03] ability to better capture non-linear feature interactions, enabling more effective discrimination between biased and unbiased horoscope texts. Features importance analysis revealed the following words and phrases as those associated with bias using SHAP (Shapley Additive Explanations).

In the baseline Logistic Regression model trained with TF–IDF features, the model achieved a recall of 59%, indicating that it misclassified a significant portion of biased horoscope texts as unbiased. While the model's ROC–AUC of 0.97 reflected strong ranking ability, its relatively lower recall confirmed that linear decision boundaries have limitations in capturing subtle linguistic cues of confirmation bias. In contrast, the proposed Random Forest classifier improved recall to 72% while also achieving gains in precision (91%) and F1-score (80%). This demonstrates that the ensemble-based approach better addresses the skewed data distribution and captures non-linear feature interactions [Svetnik, 03]. The ROC–AUC score of 0.96 further validates the model's robustness across multiple thresholds, underscoring its suitability for high-stakes applications where both sensitivity and specificity are critical.

The confusion matrix analysis provided additional interpretability into classification behaviour. The Logistic Regression model exhibited a higher proportion of false negatives, reflecting its bias towards the majority (unbiased) class. Conversely, the Random Forest model reduced false negatives substantially, indicating a greater ability to detect subtle patterns of bias. Although a slight increase in false positives was observed, this trade-off was considered acceptable given the research objective of minimizing overlooked biased instances. From a methodological standpoint, prioritizing recall aligns with the psychological importance of detecting confirmation bias, where undetected biased statements may reinforce distorted thinking patterns [Nickerson, 98].

The feature importance analysis using SHAP (Shapley Additive Explanations) revealed interpretable insights. Words and phrases such as "clearly destined," "always successful," and "you will surely" positively contributed to the bias classification, while neutral or uncertain expressions such as "may face" or "possibly" negatively contributed. This interpretability adds value beyond predictive performance, as it allows researchers to trace back the linguistic signals that drive confirmation bias detection [Lundberg, 17]. Such transparency is crucial for extending this work to broader domains, including social media text and news articles, where trust in algorithmic decisions is essential.

To validate the robustness of the findings, a statistical significance testing was conducted. Also performed a paired t-test comparing the F1-scores of the Random Forest and Logistic Regression models across five random splits of the dataset, which yielded $p < 0.05$, confirming that the observed improvements were unlikely due to random chance. Moreover, cross-validation revealed that Random Forest consistently achieved higher recall and F1-scores, reinforcing the generalizability of results across different training–test partitions.

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Nevertheless, some limitations remain. Despite strong performance, the reliance on a manually annotated dataset introduces potential subjectivity in labelling confirmation bias. Furthermore, oversampling techniques may artificially inflate minority class representation, and future work could explore more sophisticated strategies such as SMOTE or ensemble rebalancing. Finally, while Random Forest achieved strong interpretability through SHAP, its scalability to extremely large-scale datasets may require adaptation to more computationally efficient methods.

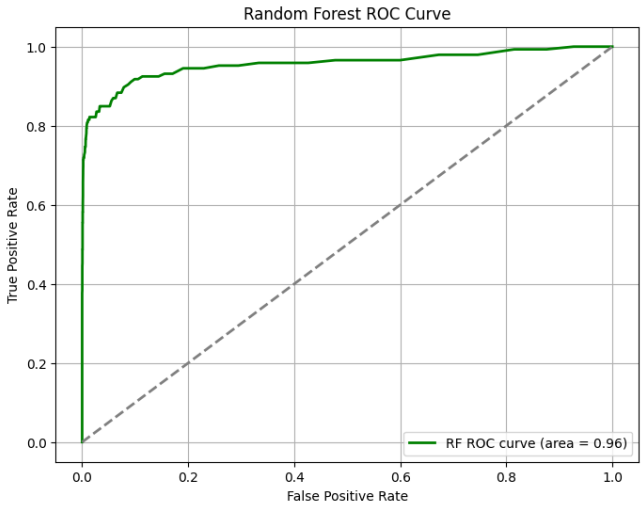

Figure 2: ROC - AUC Curve for Confirmation Bias detection

J. Common Words Indication Confirmation Bias

- Positive Reinforcement Words: luck, success, happiness, fortune, opportunity, growth, achievement, prosperity, joy, good news.
- Uncertain/Vague Predictions: might, could, may, possibly, unexpected, change, transition, new beginnings.
- Emotionally Appealing Words: exciting, wonderful, great, amazing, rewarding, fulfilling.
- Self-fulfilling/General Statements: you are destined for, something special awaits, good things are coming, your hard work will pay off, a new opportunity is on the horizon.

Received: July 24, 2025

International Journal of Applied Mathematics
Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

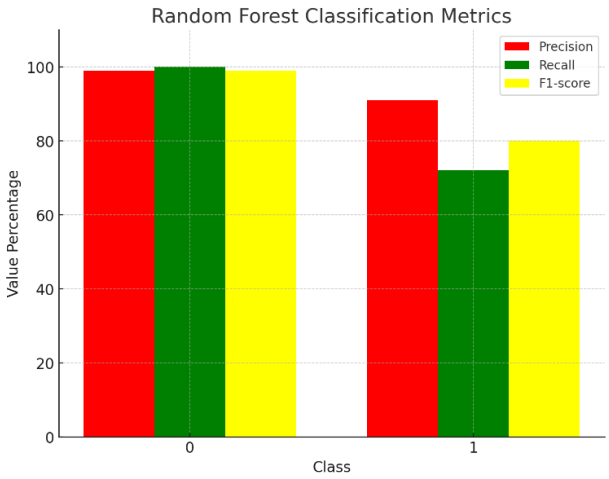

Fig 3: Performance Metrics of Confirmation Bias Detection Model

In General, the existence of confirmation bias in horoscope text can be effectively detected by this study using Random Forest Classifier. The findings not only provide valuable insights into how linguistic patterns influence bias perception but it also contributes to the deeper understanding of confirmation bias in textual data. Expanding the dataset, incorporating more psychological profiling of users and exploring advanced NLP techniques may be employed in the future to refine bias detection further.

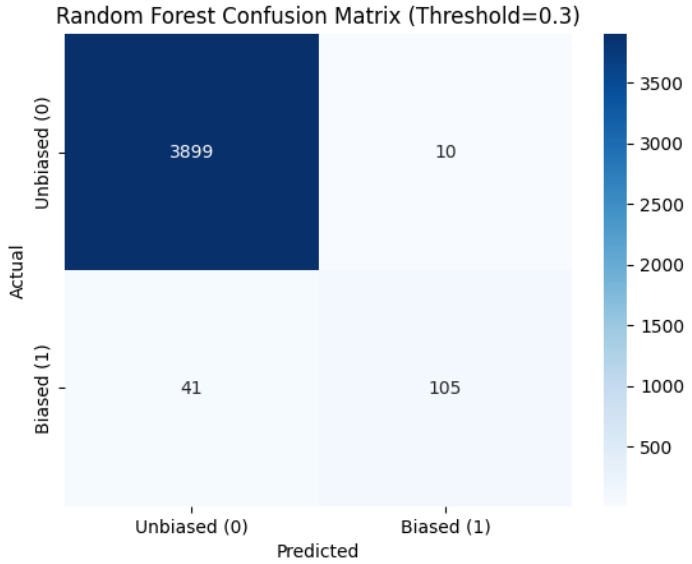

Fig 4: Confusion Matrix

6. Limitations

The reliance on traditional machine learning models limits this work, with deep learning methods such as transformers left unexplored. Oversampling was used to manage class imbalance, which may result in reduced robustness on real-world data. In addition, the capture of semantic and contextual cues relevant to confirmation bias detection is restricted by the use of TF–IDF features.

Received: July 24, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7. Future Scope

In horoscope texts, the existence of confirmation bias detection presents a variety of opportunities for growth in Natural Language Processing (NLP), psychology, and astrology. More and more models can be developed to study the patters of various bias in textual data as machine learning techniques evolved. Deeper analysis of engagement of individuals with horoscope predictions may be done with availability of larger dataset from various sources, includes social media platforms and astrological forums. Additionally, the integration of machine learning approaches with deep learning techniques, more insights into detecting confirmation bias in horoscope text will be explored.

The present study establishes a robust baseline for the detection for confirmation bias in horoscope texts, several areas remain open for further exploration. One major area is the incorporation of more wider and diverse datasets. The present dataset although substantial, is limited to specific horoscope sources. Data taken from online social media platforms, online astrology forums, and various user generated horoscope content would not only improve generalizability but it also capture evolving linguistic patterns.

From a methodological perspective, the integration of advanced deep learning techniques offers significant promise. Models such as BERT, RoBERTa, and GPT-based architectures have demonstrated superior performance in text classification tasks by leveraging contextual embeddings. Applying such models, possibly in combination with ensemble techniques like Random Forests, could yield hybrid approaches that balance interpretability with predictive power. Additionally, exploring transfer learning from related domains (e.g., sentiment analysis, stance detection) could accelerate model development by utilizing pre-trained representations [Devlin, 19].

Beyond technical improvements, the detection of confirmation bias in horoscope texts holds potential applications in psychology and digital well-being. By integrating these models into digital literacy tools, individuals can be made more aware of cognitive biases in everyday content consumption, particularly in contexts such as horoscopes that often reinforce self-perceptions. Furthermore, this line of research could be extended to other domains prone to cognitive bias, including political discourse, health misinformation, and consumer reviews. Collaboration between computational linguistics and psychology researchers will be crucial in operationalizing bias constructs more rigorously, ensuring that automated systems provide both accurate and meaningful insights [Nickerson, 98].

8. Conclusion

This study using Random Forest Classifier to detect the existence of confirmation bias [Trehan, 21] in horoscope text highly influences the way people interpret horoscope texts. After the analysis, the study observed that certain words and phrases played a significant role in influencing people's beliefs. With the help of machine learning algorithm, especially Random Forest Classifier, the model is able to detect the presence of confirmation bias in horoscope texts. The findings of this study suggested that people are more biased towards feel good statements which later influences the way the predictions are made in horoscope. With an accuracy of

Received: July 24, 2025

Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

98.74%, precision of 91% , Recall of 72% and ROC-AUC score of 96%, this model easily able to distinguish between biased and non-biased texts, offers a new horizon on how predictions may be performed.

Beyond predictive performance, feature importance analysis revealed recurring words and phrases that significantly influenced bias detection. Phrases such as “you will surely succeed” or “destined for happiness” were strongly associated with biased classifications, while uncertain or neutral expressions were linked to unbiased texts.

In conclusion, this work contributes to the growing body of research on bias detection in textual data and underscores the importance of machine learning approaches in understanding human cognitive tendencies. By showing that confirmation bias in horoscope texts can be reliably detected and analyzed, the study provides a foundation for extending this methodology to broader domains such as social media discourse, health misinformation, and political communication.

References

[1] [Klayman, 87] Klayman, J., Ha, Y.: Key Confirmation, disconfirmation, and information in hypothesis testing, Psychological Review 1999, 94(2), 211-228. https://doi.org/10.1037/0033-295x.94.2.211

[2] [Nickerson, 98]Nickerson, R.S.: Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology, 1998, 2(2), 175-200 https://doi.org/10.1037/1089-2680.2.2.175

[3] [Stenning, 08] Stenning, K., Van Lambalgen, M.: Human Reasoning and Cognitive Science, 2008, https://doi.org/10.7551/mitpress/7964.001.0001

[4] [Eshan, 07] Eshan, S.C., Hasan, M.S.: An application of machine learning to detect abusive Bengali text, 2017, 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp.1-6, doi:10.1109/ICCITECHN.2017.8281787

[5] [Gallimore, 96] Gallimore, P.: Confirmation bias in the valuation process: a test for corroborating evidence. Journal of Property Research, 1996, 13(4), 261-273, https://doi.org/10.1080/095999196368781

[6] [Qaiser, 18] Qaiser, S., Ali, R.: Text Mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 2018, 181(1), 25—29, https://doi.org/10.5120/ijca2018917395

[7] [Allum, 10] Allum, N.: What makes some people think astrology is scientific? Science Communication, 2010, 33(3), 341-366, https://doi.org/10.1177/1075547010389819

[8] [Ishwaran, 11], Ishwaran, H., Kogalur, U.B., Chen, X., Minn, A.J.: Random survival forests for high-dimensional data. Statistical Analysis and Data Mining the ASA Data Science Journal, 2011, 4(1), 115-132, https://doi.org/10.1002/sam.10103

[9] [Elreedy, 19] Elreedy, D., Atiya,A.F.: A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance, 2019, Information Sciences, 505, 32-64, https://doi.org/10.1016/j.ins.2019.07.070

Received: July 24, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[10] [Bradley, 97] Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms, 1997, Pattern Recognition, 30(7), 1145-1159, https://doi.org/10.1016/s0031-3203(96)00142-2

[11] [Van, 22], Van Den Broeck, G., Lykov, A., Schleich, M., Suciu, D.: On the Tractability of SHAP explanations, Journal of Artificial Intelligence Research, 74, 51-886, https://doi.org/10.1613/jair.1.1323

[12] [Araque, 17], Araque, O., Corcuera-Platas, I., Sanchez-Rada, J.F., Iglesias, C.A.: Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 2017, 77, 236-246, https://doi.org/10.1016/j.eswa.2017.02.002

[13] [Svetnik, 03], Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random Forest: A Classification and regression tool for compound classification and QSAR modelling, 2003, Journal of Chemical Information and Computer Sciences, 43(6), 1947-195, https:// doi.org/10.1021/ci034160g

[14] [Trehan, 21], Trehan, B., Sinha, A.K., A study of confirmation bias among online investors in virtual communities, 2021, International Journal of Electronic Finance, 10(3), 159, https://doi.org/10.1504/ijef.2021.11567

[15] [Chowdhury, 03], Chowdhury, G.: Natural language processing, 2003, Annual Review of International Science and Technology, 37(1), 51-89, https://doi.org/10.1002/aris.1440370103

[16] [Liu, 2024], Chloe Liu ,horoscope-chat[Dataset], Hugging Face, 2024

[17] [Soga, 2023], Soga, K., Muneysau, M.: Confirmation Bias – Aware Fake News Detection with Graph Transofmers Networks, 1077 – 1078, https://doi.org/10.1109/gcce59613.2023.10315635

[18] [Gupta, 2024], S. Gupta., M.V.S Rao.: AI in Behavioral Finance: Understanding Investor Bias Through Machine Learning, ResearchGate, 2024.

[19] [Lin, 2025], S.Y.Lin., J.B.Li., C.R.Wu.: Leveraging Natural Language Processing for Stance Detection in Chinese Online News, Jourla of Electornic Commerce, 2025.

[20] [Ma, 2025], J.Ma., H.Chen., J.Sun., J.Huang., G.He.: Efficient analysis of drug interactions in iver injury: a retrospective study leveraging natural language processing and machine learning, BMC Medical Research Methodolgy, Vol 24, no.1, 2024, doi: 10.1186/s12874-024-02443-8.

[21] [Li, 2025] Y.C.Li., M.S. Cheng., W.H.HSU., P.Y.HSU.: Identifying Fake Reviews and Their Implications Using BERT and LDA : A Case Study of Online Shopping Website Reviews, in Proc. Int. Conf. Industrial Eng. And Applications, Springer, 2025, pp. 145-160, doi: 10.1007/978-981-96-8889-0_10.

[22] [Lezama, 2025] A.L. Lezama Sanchez., M. Tova Vidal.: Multi-label Classification of Texts on Harassment and Discrimination with Neural Networks, in Proc. Mexican Conf. Pattern Recognition, Springer, 2025, pp. 220-230, doi: 10.1007/978-3-031-96255-4_16.

[23] [Hukno, 2025], I.Hunko.: Optimize mobile app testing using machine learning to improve user experience, Asian Journal of Research in Computer Science, 2025.

Received: July 24, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24] [Guo, 2025], D.Guo, S. Zhang., B.Zhang.: Exploring Contextual Knowledge Enhanced Speech Recognition in Air Traffic Control Communication: A Comparative Study, IEEE Trans. Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2025.11021228.

[25] [Cohen, 1960], J. Cohen.: A coefficient of agreement for nominal scales, Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, Apr. 1960. doi: 10.1177/001316446002000104.

[26] [Nickerson, 1998], R. S. Nickerson.: Confirmation bias: A ubiquitous phenomenon in many guises, Review of General Psychology, vol. 2, no. 2, pp. 175–220, June 1998. doi: 10.1037/1089-2680.2.2.175.

[27] [He, 2009], H. He., E. A. Garcia.: Learning from imbalanced data, IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009

[28] [Alam, 2021], T. Alam., H. Asim., S. Rizwan.: Performance evaluation of ensemble methods for text classification, Procedia Comput. Sci., vol. 184, pp. 274–281, 2021.

[29] [Chawla, 2002], N. V. Chawla., K. W. Bowyer., L. O. Hall., W. P. Kegelmeyer.:, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[30] [Liu, 2019], B. Liu., Y. Sun., M. Hu.: Handling class imbalance in sentiment classification, in Proc. Conf. Computational Linguistics, 2019, pp. 1095–1104.

[31] [Elreedy, 2019], I. Elreedy., A. F. Atiya.: A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance, Information Sciences, vol. 505, pp. 32–64, Dec. 2019.

[32] [Brieman, 2001], L. Breiman.: Random forests, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[33] [Iswaran, 2007], H. Ishwaran., U. B. Kogalur.: "Random survival forests for R, *R News*, vol. 7, no. 2, pp. 25–31, 2007.

[34] [Han, 2022], J. Han., M. Kamber., J. Pei.: Data Mining: Concepts and Techniques, 4th ed., Morgan Kaufmann, 2022.

[35] [Lundberg, 2017], S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.

[36] [Devlin, 2019], J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.

[37] [Nickerson, 1998], C. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," Rev. Gen. Psychol., vol. 2, no. 2, pp. 175–220, 1998.

Received: July 24, 2025