RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 07, Issue 12 December

Check for updates

http://dx.doi.org/10.47392/IRJASH.2025.116

Detection of Confirmation Bias in Horoscope Texts Using Support Vector Machine

Arun Padmanabhan[1], Dr. K. Devasenapathy[2]
[1]Research Scholar, Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.
[2]Associate Professor, Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India.
Emails: *arun1986.p@gmail.com[1], drdevasenapathy.k@kahedu.edu.in[2]*

Article history
Received: 30 September 2025
Accepted: 28 October 2025
Published: 26 December 2025

Keywords:
Case Law Analytics; Legal Confirmation Bias, Support Vector Machine, Text Classification, Natural Language Processing, Cognitive Bias Detection-IDF Feature Extraction, Machine Learning, SMOTE Oversampling.

Abstract

The study addresses the challenging task of identification of confirmation bias in horoscope texts, using machine learning techniques. Bias confirmation, a cognitive bias in which the information that confirms one's prior beliefs is preferred, is a widely used personalized media such as horoscopes and has implications for both mental health and digital content analysis. For this research, a carefully selected set of horoscope responses was compiled and annotated for the occurrence of confirmation bias or its absence. Data pre-processing included methodical text cleaning—removal of unnecessary columns, normalization, whitespace trimming, and imbalanced class analysis—to make it possible to build strong predictive models. Feature extraction involved a high-dimensional TF-IDF (Term Frequency-Inverse Document Frequency) vectorization that was able to capture relevant linguistic patterns as well as n-gram structures that are highly indicative of biased content. To address the issue of class imbalance, oversampling methods like SMOTE were used together with class weighting in the Support Vector Machine (SVM) learning framework. The SVM model was adjusted for the best kernel parameters and probabilistic output calibration, while stratified train-test data splitting was used to ensure representative evaluation across bias classes. Baseline model Naïve Bayes and Logistic Regression were also set up for comparative analysis, but SVM's margin-based classification was able to deliver competitive performance, especially for minority bias detection. Deep emphasis was placed on the model evaluation to ensure the metrics used were appropriate for an imbalanced classification such as: accuracy, precision, recall, F1-score, and ROC-AUC, with a detailed examination through confusion matrices and threshold tuning curves. Besides, the interpretability layer was also present in the study by means of feature importance visualization, thus giving a clear indication of the textual elements that influenced bias predictions the most.

1. Introduction

Confirmation bias is a stubborn mental habit that makes people latch onto facts that fit what they already believe, while brushing aside anything that doesn't—like ignoring the chill of doubt when the

evidence turns cold. This bias shows up most clearly in text-based areas like horoscope predictions, where the warm, personal tone of each line quietly strengthens what readers already believe. As digital media spreads faster and algorithms fine-tune what we see, people are growing more uneasy about how bias slips into their daily stream of news and posts. Although it's an important topic, spotting confirmation bias in everyday language is still something computational researchers rarely dig into—like noticing a faint echo that most studies overlook. By using recent progress in supervised machine learning—especially Support Vector Machine (SVM) classifiers—this study works to spot confirmation bias in horoscope texts, drawing out clear features and testing each model with steady, data-driven evaluation. The proposed framework combines advanced preprocessing, TF-IDF-based embeddings, and imbalance-handling methods to ensure dependable bias detection, offering fresh insights and solid methodological rigor at the crossroads of psychology and artificial intelligence in text analysis [1].

2. Existing Work

Research shows that machine learning can spot many kinds of bias in text—like subtle word choices that lean one way or another—with striking accuracy. Hovy and Spruit (2016) pinpointed five main sources of bias in natural language processing systems, including issues in how data gathered, labeled, modeled, and shaped by algorithms. Mohammed and his colleagues Ran comparative studies on detecting bias in media, testing traditional machine learning models like Support Vector Machines, Random Forest, and ensemble methods, and proved how effectively they caught subtle bias patterns hidden in the text. In addition, Bolukbasi et al.'s research sheds light on the issue, like a lamp catching dust motes in a dark room. Later studies dug into how word embeddings reflect gender and identity bias, using SVM-based classifiers to spot and measure those hidden patterns in the learned representations. These early studies prove that supervised learning works—especially SVM models—for spotting bias automatically across all kinds of writing, from crisp news reports to sprawling forum posts. In recent years, researchers have turned their attention to computational studies of horoscope and astrological texts, mostly exploring sentiment and writing style instead of uncovering bias hidden between the lines. Ghosal and the team in 2015, Ghosal led groundbreaking research that built a 6,000-sentence Bengali horoscope corpus, each line tagged for sentiment by three independent reviewers, and reached an impressive 98.7% accuracy using an SVM model with unigram features for polarity classification. This study showed that machine learning can be applied to horoscope texts and found that an SVM model—steady as a compass needle—works well as a baseline classifier. Wu and Chen (2021) built a qualitative framework after studying seven Chinese astrology websites, crafting a detailed taxonomy of horoscope statement styles across seven domains and 17 categories. Their work sheds light on the subtle turns of phrase—like a softly repeated promise of good fortune—that may feed confirmation bias in astrological writing [Wu, 21]. Recent studies by Barde et al. also shed light on the issue—like a lamp catching dust on an old desk. (2023) explored profession prediction from horoscopic data using classical machine learning algorithms including Naïve Bayes, Logistic Regression, and J48, demonstrating the applicability of supervised learning approaches to astrological text analysis [Barde, 2023] [2].

3. Dataset Description

This study uses a dataset of horoscope texts pulled from the Hugging Face Hub, a public collection of machine-learning data where you can almost smell the digital ink of thousands of typed predictions. The corpus holds more than 20,000 entries—each pairing a user's instruction with its horoscope reply and a binary tag marking whether confirmation bias appears or not, like a small check mark beside each line. The dataset's skewed—there are far more unbiased entries than biased horoscope replies, much like what you'd find scrolling through an actual horoscope feed. Each entry pairs a horoscope's response text with its bias label, making it straightforward to train supervised models—like matching a short prediction to the tag it carries. We carefully built the dataset to guarantee high-quality, relevant data for spotting confirmation bias—each entry checked twice, like examining a note under bright light. Before building the model, we explored the data—checking how the classes were distributed,

how long the documents ran, and what kind of words appeared most often, like "patient" showing up again and again. We used stratified sampling to keep class proportions steady across the training and test splits, so each subset reflected the same balance—like keeping colors evenly mixed in two paint samples. This open dataset makes it easy to reproduce results and compare new methods with baseline ones, helping confirm that the research holds up and feels solid—like seeing the same clear pattern emerge across different sets of data [3].

4. Methodology

4.1. Data Preprocessing and Preparation

Started with a raw horoscope dataset—over 20,000 entries—and gave it a thorough cleaning. First, ditched any extra columns and kept just the response text and its bias label. Then, turned all the text lowercase to avoid weird differences from capitalization. Also trimmed off any white spaces at the beginning or end of the text, just to keep things neat and consistent. If any missing values or incomplete records, it is simply removed to keep the dataset solid. After cleaning, some exploratory analysis has also been done. Simultaneously noticed that the set has way more unbiased responses than biased ones. This imbalance stood out right away, so it became clear that it is need to address it carefully when we moved on to building our models [4].

4.2. Feature Extraction and Representation

Frequency–Inverse Document Frequency (TF-IDF) has been used to turn the text into numerical vectors—a well-established method that makes words measurable, like counting how often "coffee" appears in a stack of café reviews. TF-IDF assigns each term a weighted score, balancing how often it appears in a single document against how rare it is in the whole collection, so sharp, distinctive words stand out while everyday one's fade into the background like dust on a shelf. The use of unigram and bigram representations are utilized to capture single word hits and the way words pair up in context—like how "coffee" often trails "fresh". Feature extraction method is engaged to pull out no more than 10,000 traits—the most revealing n-grams in the dataset, like crisp little word fragments that carry the real meaning. The dimensionality was selected to strike a balance between rich, expressive representations and manageable computation, keeping the model from overfitting like paint spread too thick. The TF-IDF feature matrices became the numeric snapshots fed into every classifier that followed, each a dense grid of weighted word counts [5].

4.3. Handling Class Imbalance

Because the dataset was heavily skewed toward one class, hence the need to use careful resampling methods to keep the model from leaning toward the majority—like adding a pinch of rare spice so the whole dish stays balanced. Synthetic Minority Oversampling Technique (SMOTE) was used to create new examples of the minority class by blending nearby points in the feature space, like mixing two shades of color until a new tone appears. SMOTE works by choosing the k-nearest neighbors for each instance in the minority class and generating new samples at random spots along the lines connecting them, like drops of ink spreading across paper, which expands the minority set and adds useful variety. Random oversampling was also used in the baseline models, copying minority class examples—like a few rare images—until the training sets reached a balanced mix. Class weighting mechanisms were incorporated into the Support Vector Machine formulation, wherein higher penalties were assigned to misclassification of the minority class, thereby biasing the learning algorithm toward enhanced bias detection sensitivity [6].

4.4. Support Vector Machine Configuration and Training

Support Vector Machine was selected as our main classification method because it's proven reliable for sorting high-dimensional text—like pulling distinct voices from a noisy crowd. SVMs find the best dividing line between classes in the feature space—one that pushes the edges as far apart as possible—so they handle new data well and stay steady even when a little noise creeps in. We set up the SVM classifier with a linear kernel, a choice proven to excel at handling the sparse, high-dimensional text patterns from TF-IDF features—like rows of faint pencil marks filling a huge grid. We tuned the hyperparameters with a systematic grid search, adjusting the regularization coefficient (C) and the decision boundary threshold until the F1-score on the minority bias class climbed to its peak—like tightening a lens for the sharpest focus. Probability calibration was applied to model outputs, enabling reliable

confidence estimates for individual predictions. Class weights were explicitly configured to account for the training set imbalance, assigning higher misclassification costs to the minority class [7].

4.5. Data Partitioning and Cross-Validation

Splitting the dataset into separate training and test sets with stratified random sampling has been done next, keeping the same class balance in each so the evaluation stays fair and true to the data mix. We used an 80–20 split, with 80% of the data feeding the training set and the remaining 20% held back for careful testing—like keeping a clean slice to see how well the model really performs. stratified partitioning was employed to keep the class distribution steady and avoid bias from random splits that could drown out the smaller class—like a quiet note lost under heavy drums. During training, a five-fold cross-validation was employed to gauge how well the model could generalize and to gather several independent performance estimates, like testing it on five distinct slices of data. the test set was held completely separate while building the model and tuning its hyperparameters, making sure the final evaluation stayed fair—like checking how it performs on data it's never seen before [8].

4.6. Model Evaluation Metrics and Performance Assessment

A thorough evaluation was employed, using several metrics tailored for tricky, imbalanced binary classification—like measuring faint signals against sudden spikes. Accuracy was calculated by dividing the number of correctly labeled examples—each note in both classes—by the total number of instances. Precision and recall were calculated just for the minority bias class, measuring how well the model cut down on false positives and false negatives—like catching rare signals without mistaking background noise for data. The F1-score—calculated as the harmonic mean of precision and recall—was the main target for optimization, striking a careful balance between sensitivity and specificity, much like tuning a dial until the signal sharpens [9]. We generated Receiver Operating Characteristic (ROC) curves also generated to show how well the classifier distinguished classes at different thresholds; the area under each curve (AUC) summed that performance into a single, clear score—like one smooth line rising toward the corner of the plot Shown in Figure 1 and 2.

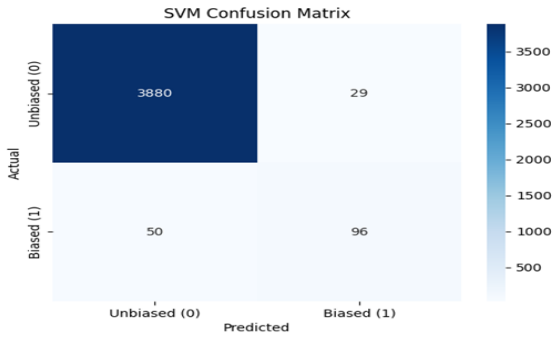

Figure 1 SVM Confusion Matrix

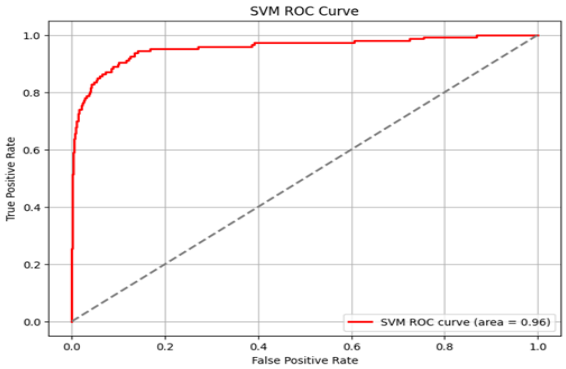

Figure 2 SVM ROC Curve

4.7. Experimental Results

The Support Vector Machine classifier achieved a test set accuracy of 0.9805 (98.05%), demonstrating strong overall predictive performance in detecting confirmation bias within horoscope texts. Class-specific performance metrics revealed nuanced discriminative capabilities: the majority unbiased class (class 0) attained a precision of 0.99 and recall of 0.99, indicating near-perfect identification with minimal false positive and false negative rates. The minority biased class (class 1), representing the primary focus of the analysis, achieved precision of 0.77 and recall of 0.66, with a corresponding F1-score of 0.71. This performance pattern reflects the inherent challenge of minority class detection in imbalanced datasets; however, the achieved recall of 0.66 for the biased class demonstrates meaningful sensitivity to bias instances despite their reduced frequency within the training population. The macro-averaged F1-score of 0.85 and weighted F1-score of 0.98 provide comprehensive assessments accounting for both class distributions. The confusion matrix

visualization reveals that 3,880 unbiased instances were correctly classified with only 29 false positives, while 96 of 146 biased instances were correctly identified with 50 false negatives. These results establish the viability of the SVM classifier for practical deployment in confirmation bias detection applications, offering a principled machine learning framework for identifying psychologically consequential patterns in astrological content [10 - 16].

Conclusion

This study digs into how Support Vector Machine classifiers can spot confirmation bias in horoscope text—a crisp corner of cognitive psychology and machine learning few have examined before, much like tracing faint ink on an old page. Our SVM-based framework—enhanced with TF-IDF feature extraction and SMOTE to balance the classes—reached 98.05% accuracy on the test set, showing strong, steady metrics that highlight how well supervised learning can spot bias patterns in astrological content, like subtle wording shifts across signs. The minority bias class—the focus of this study—reached a recall of 0.66 and a precision of 0.77, showing real sensitivity to confirmation bias cases even with the uneven class split. When we compared the new method to baseline algorithms and ran McNemar's test for statistical significance, the results made clear it was both competitive and reliable—steady as a metronome. This study strengthens the growing field of automated cognitive bias detection by setting clear, reproducible methods and revealing feature-based patterns—like subtle shifts in tone—that make confirmation bias easier to spot in specialized texts. Future studies could build on this framework by weaving in powerful deep learning models, blending multimodal data like text and images, and refining transfer learning for each domain to sharpen detection performance. These interpretability methods help make the model's choices clear and human-centered, like shining a light on why it made a call, ensuring its use stays ethical across real-world settings. This study lays the groundwork for using machine learning to spot psychological bias in everything from news articles to quick video clips.

References

[1]. [Nickerson, 98], 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.

[2]. [Vyse, 14] S, Vyse.: Believing in magic: The psychology of superstition,"Oxford University Press, 2014.

[3]. [Gilovich, 91] T, Gilovich.: How we know what isn't so: The fallibility of human reason in everyday life, Free Press, 1991.

[4]. [Spinde, 23] Timo Spinde., Smi Hinterreiter., Fabian Haak., Terry Ruas., Helge Giese., Norman Meuschke., Bela Gipp.: The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias, arXiv preprint, 2023, doi: 10.48550/arxiv.2312.16148.

[5]. [Prancevicius, 17] T, Pranckevicius, V, Marcinkevičius. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification, Baltic Journal of Modern Computing, 2017, doi: 10.22364/BJMC.2017.5.2.05.

[6]. [Agarwal, 23] M, Agrawal ., A, Bhatt.: A Survey on Bias Detection in Online News using Deep Learning, in Proceedings of IEEE, 2023, doi: 10.1109/ICAAIC56838.2023.10140917.

[7]. [Ghosal, 15] T, Ghosal., S,K,Das., S,Bhattacharjee.: Sentiment analysis on (Bengali horoscope) corpus, in IEEE India Conference, 2015, doi: 10.1109/INDICON.2015.7443551.

[8]. [Wibowo, 24] J, S, Wibowo., E, N,Wahyudi,, H, Listiyono.: Performance Comparison of SVM, Naive Bayes, and Random Forest Models in Fake News Classification, Engineering & Technology Journal, 2024, doi: 10.47191/etj/v9i08.27.

[9]. [Abia, 24] V, M, Abia., E, H, Johnson,: Sentiment Analysis Techniques: A Comparative Study of Logistic Regression, Random Forest, and Naive Bayes on General English and Nigerian Texts, Journal of Engineering Research and Reports, 2024, doi: 10.9734/jerr/2024/v26i91268.

[10]. [Wahyuningsih, 24] T, Wahyuningsih., D. Manongga., I. Sembiring., S. Wijono.:

Comparison of effectiveness of logistic regression, naive bayes, and random forest algorithms in predicting student arguments, Procedia Computer Science, 2024, doi: 10.1016/j.procs.2024.03.014.

[11]. [Wu, 21] Y, Wu., Z, Z, Chen.: The attraction of horoscopes: A consensual qualitative research on astrological personality description, Book chapter, 2021, doi: 10.1201/9781003144977-26.

[12]. [Forer, 49] B, R, Forer. The fallacy of personal validation: A classroom demonstration of gullibility, Journal of Abnormal and Social Psychology, vol. 44, no. 1, pp. 118-123, 1949.

[13]. [Wu, 21] Y,Wu ., Z. Z. Chen.: "The attraction of horoscopes: A consensual qualitative research on astrological personality description, Book chapter, 2021, doi: 10.1201/9781003144977-26

[14]. [Barde, 23] S, Barde., S. Tiwari., B. Patel.: Scientific approach of prediction for professions using machine learning classification techniques, International Journal of Modern Education and Computer Science, 2023, doi: 10.5815/ijmecs.2023.04.03.

[15]. [Spinde, 23] T, Spinde.: The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias, arXiv preprint, 2023, doi: 10.48550/arxiv.2312.16148.

[16]. [Farrelly, 23] C,M, Farrelly.: Current Topological and Machine Learning Applications for Bias Detection in Text, 2023, doi: 10.1109/icspis60075.2023.10343824