International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Probability Calibration Using Platt Scaling in SVM Based Fake News Detectors

Aditi Ganapati Karki[1], Ms. Dimpy Parashar[2], Dr. Yatu Rani[3]

[1]AI-DS Scholar, AI&DS, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India
[2]Assistant Professor, AI&DS, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India
[3]Associate Professor, AI&DS, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India

Abstract
The rapid growth of unverified digital media has made fake news a growing concern in recent years. As content spreads quickly through social media and digital platforms, distinguishing between reliable and misleading information has become increasingly difficult. This not only defeats the purpose of core journalism but also has broader societal consequences. A confidence estimation mechanism is incorporated using the SVM classifier to display the certainty of each prediction. This feature is enabled so that a confidence score is provided alongside the classification result, improving the reliability of the model's decisions that refine the digital medium. These findings suggest that even simple ensemble techniques can provide meaningful improvements, particularly in settings where data and resources are limited.

Keywords: Digital Media; Fake News; Platt Scaling; Classifier

I. INTRODUCTION

A. Background

Fake News contains misleading information that could be checked. This could be a lie about a certain commercial product or can include creating unnecessary panic by spreading fake information regarding a plausible World War 3, which may cause unrest for the common people who solely rely on digital media. There are regulatory bodies whose main role is to deal with reliability by checking if authors are accountable. However, their scope is limited because they depend on human manual detection. In a globe with millions of articles being published every minute, this cannot be accountable or feasible manually.
A solution could be the development of a system to provide a credible automated index scoring or rating for credibility of different publishers and news context. This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on the analysis of the content of the articles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, using the probability=true feature of the support vector machine algorithm. We propose to create the model by calculating the probability of belonging to a particular class using the concept of Platt scaling.
The product model will test unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and gives a confidence score as well indicating how confident the model is. Usually, models like these use natural language programming (NLP) approaches like deep learning, but this method often requires a large amount of data to predict accurately (millions of examples).

Algorithms like Support Vector Machine (SVM) can perform reasonably well with smaller datasets because they rely on feature engineering rather than learning embeddings from scratch.

II. LITERATURE REVIEW

Numerous researchers have explored various approaches for fake news detection. Jahan et al. (2024) [1] suggested that analysis of textual and contextual patterns can help in distinguishing news articles as true or fake by implementing machine learning models through Python and NLP techniques. Tiwari and Jain (2024) [13] compared different machine learning algorithms such as decision tree, random forest, and logistic regression for fake news detection and showed that ensemble and hybrid models can improve classification accuracy. Chen et al. (2024) [3] recommended multimodal fake news detectors where both textual and visual features are implemented together to improve the reliability of detection systems. Karim et al. (2023) [4] analyzed different text vectorization techniques for fake news detection using Support Vector Machines and found that linear-kernel SVM combined with Bag-of-Words features achieved accuracy close to transformer-based models.

The previous work done suggests that using SVM classifiers is the most reliable and effective model compared to other classifiers, due to its suitability for high-dimensional space and better metric scores (i.e., Accuracy, Precision, F1-scores).

III. OBJECTIVES OF THE PAPER

The study seeks to evaluate the viability of linear SVM classifiers as a robust and interpretable baseline for binary fake news classification in a high-dimensional linguistic feature space. With multiple fake detectors available on the internet, it becomes difficult to rely on one model.

Although the Support Vector Machine model provides better metric scores compared to other ML models, it is important to introduce a feature that is part of the user interface so that clients without any technical knowledge can access the precision and accuracy scores with interactive web design.

The model is further configured with probability calibration through Platt scaling, enabling the generation of class posterior probabilities alongside discrete classifications, extending the system's utility to scenarios requiring confidence-aware predictions, threshold adjustment, or integration within ensemble frameworks. To fulfill this criterion, a feature termed as 'Confidence Score' was introduced so that it gives a numerical value which serves as a quantitative entity informing the user how confident the model is with its prediction.

IV. METHODOLOGY

A model was created using fundamental principles of Natural Language Processing (NLP), implemented by Support Vector Machine (SVM). The entire pipeline is structured as follows:

1. Load the data
2. Provide the labels to each data file
3. Combine the datasets and shuffle using the ignore index
4. Clean the text using a function and remove the word 'reuters' from the dataset to reduce publisher bias
5. Set the train-test split
6. Fit and transform the train texts, and only transform the test texts
7. Train the Support Vector Machine model
8. Evaluate the metrics

9. Save the model

The methodology is divided into two sections: (A) Model Training and (B) Working of Confidence Score.

A. Model Training

A.1 Dataset and Preprocessing

This study uses a binary-labelled dataset comprised of two groups: a set of verified real news articles and a set of fake or misleading news articles. Ground-truth labels were assigned in a straightforward manner. Real articles received a positive label (1), while fake articles received a negative label (0). The two datasets were combined into a single dataset and shuffled with a fixed random seed to eliminate ordering bias and ensure reproducibility.

Before feature extraction, all article texts were cleaned using a standardized process. Each document was converted to lowercase for consistency. Publisher-specific tokens, particularly the word "reuters," were removed to avoid source-attribution bias, which could inflate classification performance. HTML code, hyperlinks, and extra spaces were also stripped using regular expressions, yielding clean, normalized plain-text inputs.

A.2 Feature Extraction

Textual features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. The vocabulary was limited to the top 6,000 most distinct terms, including both unigrams and bigrams (n-gram range: 1–2) to capture local context between nearby tokens. Standard English stop words were excluded from the feature set to reduce noise.

To avoid data leakage, the TF-IDF vectorizer was fitted only on the training data and applied unchanged to the test data, ensuring that test set statistics did not influence the vocabulary or IDF weights learned during training.

A.3 Experimental Setup

The dataset was split into training and testing subsets with an 80/20 ratio. Stratified sampling was used during the split to maintain the original class distribution in both subsets, reducing the risk of class imbalance affecting evaluation. A fixed random seed (42) was used for all stochastic processes to ensure reproducibility.

A.4 Classification Model

A Support Vector Machine (SVM) with a linear kernel was chosen as the classification method. SVMs with linear kernels work well with high-dimensional, sparse text data produced by TF-IDF vectorization, optimizing a maximum-margin boundary that generalizes efficiently. Probability calibration was enabled through Platt scaling (probability=True), allowing the model to provide class probabilities as well as hard predictions. This feature is useful for adjusting thresholds or integrating with other models.

A.5 Evaluation

The model's performance was evaluated on the held-out test set using four common classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures overall correctness; precision and recall assess the balance between false positives and false negatives; and the F1-score provides a balanced view that handles class imbalance well. A confusion matrix was also computed to detail true positives, true negatives, false positives, and false negatives.

TABLE I.
SVM Classification Performance Metrics

Accuracy	Recall	Precision	F1-Score
0.9846	0.9890	0.9788	0.9840

Confusion Matrix: [[4604, 92], [46, 4238]]

A.6 Model Persistence

After training and testing, both the trained SVM classifier and the TF-IDF vectorizer were saved to disk using Python's pickle module. Saving the vectorizer with the model is crucial, as any future predictions must use the same vocabulary and IDF weighting scheme learned during training to maintain feature space consistency.

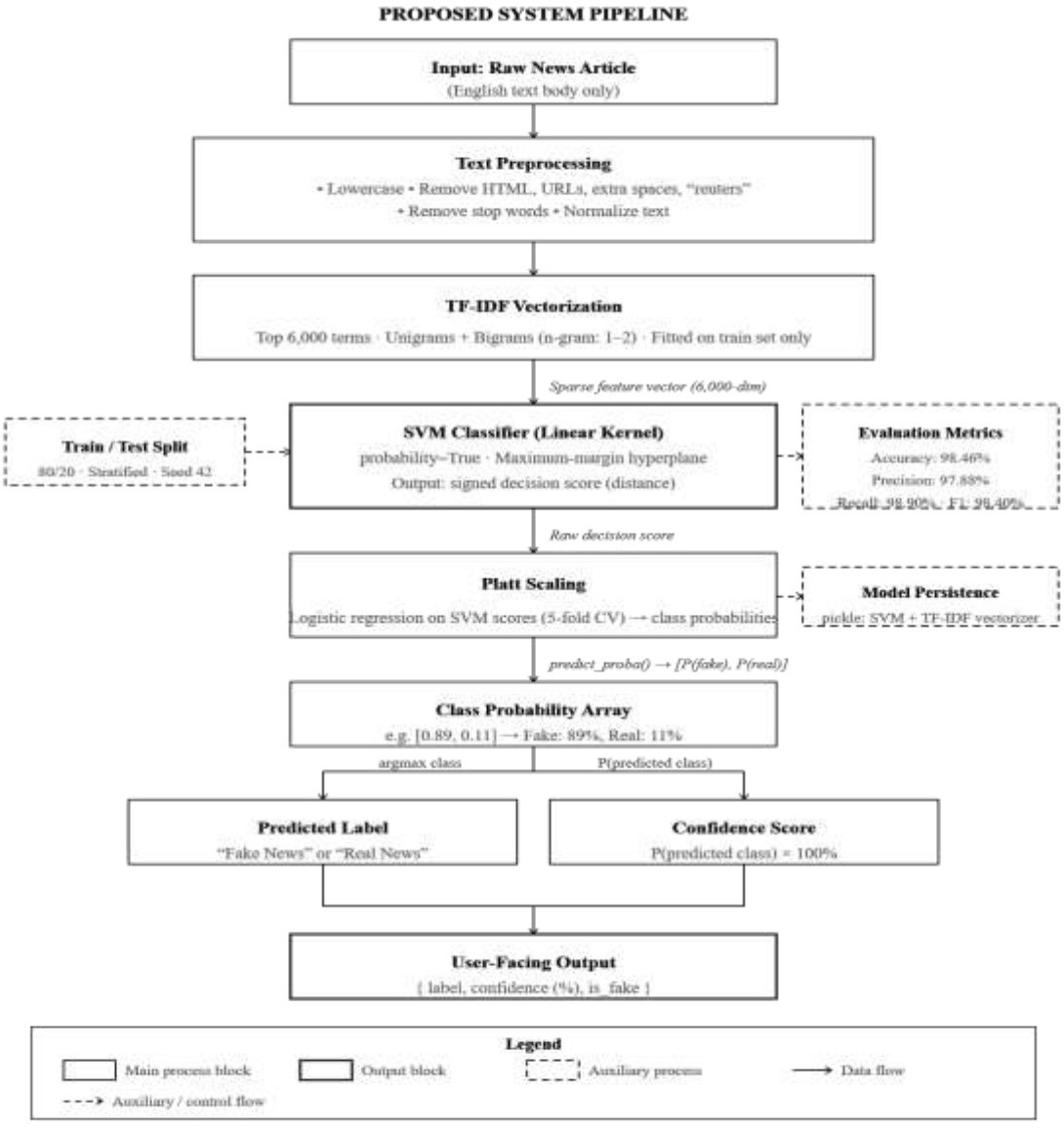


Fig. 1. Block diagram of the proposed SVM-based fake news detection system with Platt scaling probability calibration.

B. Working of Confidence Score

B.1 Need for probability=True

SVM is a distance-based classifier, not a probability-based one—it finds a hyperplane and classifies which side the point falls on. Hence, SVM does not naturally give probabilities but gives a decision score, which is the distance from the hyperplane, not a probability score like '85% true.' Therefore, probability=True was required.

B.2 Internal Platt Scaling

In the code, setting probability=True on the SVC constructor causes sklearn to internally run Platt Scaling, taking the raw SVM decision scores and fitting a logistic regression on top using cross-validation. The complete flow is:

Raw Text
↓ TF-IDF Vectorization
SVM Decision Score (e.g., -1.8 → fake side)
↓ Platt Scaling (Logistic Regression)
Probability: [0.89, 0.11] (89% fake, 11% real)

The trained Platt scaler is saved inside the pickle file along with the SVM.

B.3 Probability Array and Confidence Score

The predict_proba() method returns a 2D array with one row per input. For a single input:

predict_proba(X) = [[0.89, 0.11]]
↑ ↑
Fake% Real%

The prediction index (0 for Fake, 1 for Real) is used to extract the correct probability:

confidence_score = probabilities[prediction] * 100

This always extracts the probability of whichever class was predicted.

B.4 Human-Readable Output

The format_prediction() function packages the result into a structured output:

```
def format_prediction(prediction, confidence_score):
is_fake = (prediction == 0)
label   = 'Fake News' if is_fake else 'Real News'
return {
'prediction': label,
'confidence': round(confidence_score, 2),
'is_fake'   : is_fake
}
```

Complete end-to-end example for "BREAKING: Vaccines contain microchips!":

Input → 'breaking vaccines contain microchips'
TF-IDF → sparse vector (6000 features)
predict() → 0 (Fake)
proba() → [0.93, 0.07]
Output → { prediction: 'Fake News', confidence: 93.0 }

V. SCOPE OF THE PAPER

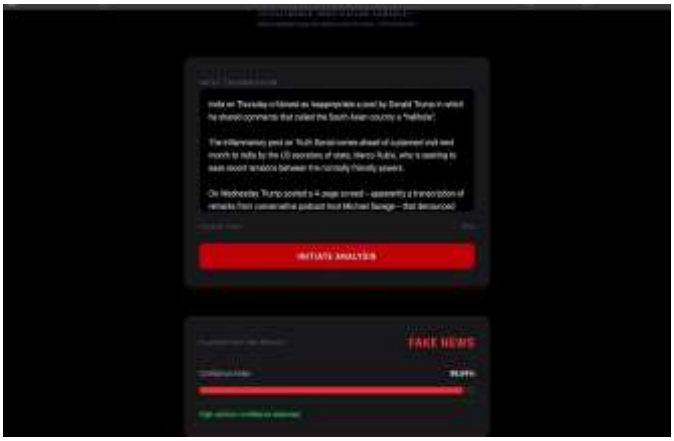
This research is restricted to binary classification of textual news articles as real or fake using a machine

learning approach. The dataset used is English language only, and the article body text alone is considered. The dataset is pre-labelled and static; the methods and results reported in this work cannot be generalized for real-time applications. The research is limited to the linear kernel SVM model and is not applicable to other machine learning methods or deep learning architectures. The research does not include any multimodal signals such as images, metadata, or social media propagation.

VI. RESULT ANALYSIS

The results show that the approach builds upon hard binary classification. In particular, the SVM pipeline is extended by the incorporation of Platt scaling to obtain calibrated posterior probabilities for every prediction. This represents a meaningful improvement over the standard SVM, which only provides a decision boundary.

This is useful in that the decision sensitivity can be adapted to specific scenarios. For instance, a stricter classification threshold (above the default 0.5 decision boundary) may be preferred in publishing contexts where false positives could damage a publisher's reputation. In addition, the calibrated probabilities are useful when constructing an ensemble or stacking approach. The current approach therefore functions not only as a standalone binary classifier but also as a calibrated expert that can contribute to a broader ensemble framework for misinformation detection



VII. CONCLUSION

This paper presented a fake news detection pipeline based on a linear SVM classifier enhanced with Platt scaling for probability calibration. By enabling the probability=True parameter, the model produces confidence-aware predictions that go beyond simple binary decisions, providing end users with a quantitative confidence score alongside each classification. Experimental results demonstrated strong performance, with an accuracy of 98.46% and an F1-score of 98.40% on the held-out test set. The approach is lightweight, interpretable, and effective even with smaller datasets, making it a practical alternative to deep learning-based methods. Future work could explore extending the system to multimodal inputs, real-time data streams, and ensemble configurations to further enhance detection robustness.

REFERENCES

1. I. Jahan, M. N. Hasan, S. N. Islam, L. Akter, M. K. R. Onik, and M. Adnan, "Advanced machine learning techniques for fake news detection: A comprehensive study," Magna Scientia Advanced Research and Reviews, vol. 12, no. 2, pp. 203–212, 2024.

2. A. Nachaithong and K. Wisaeng, "Improved SVM with hyperparameter tuning for fake news detection," Journal of Computer Science, vol. 20, no. 10, pp. 1357–1375, 2024.
3. R. Anandan, T. Nalini, S. Chiwhane, and M. Shanmuganathan, "Detection of fake news from social media using support vector machine learning algorithms," Journal of King Saud University – Computer and Information Sciences, vol. 36, no. 8, Art. no. 102160, 2024.
4. H. I. Gungbias, V. O. Waziri, I. Ismaila, J. Ojeniyi, M. Olalere, and O. Adebayo, "A machine learning approach to fake news detection using Support Vector Machine (SVM) and unsupervised learning model," in Proc. Cyber Secure Nigeria Conf., Jul. 2023, pp. 11–18.
5. I. Ahmad, M. Yousaf, and M. Gawo, "Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives," Computers, vol. 14, no. 9, Art. no. 394, 2025.
6. H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," Frontiers in Big Data, vol. 6, Art. no. 1320800, 2024.
7. M. Nadeem, P. Abbas, W. Zhang, S. Rafique, and S. Iqbal, "Enhancing fake news detection with a hybrid NLP-machine learning framework," ICCK Transactions on Intelligent Systematics, vol. 1, no. 3, pp. 203–214, 2024.
8. S. V. Balshetwar, A. RS, and D. J. R, "Fake news detection in social media based on sentiment analysis using classifier techniques," Multimedia Tools and Applications, vol. 82, pp. 35781–35811, 2023.
9. S. A. Al-obaidi and T. Çağlıkantar, "Automated fake news detection system," Iraqi Journal for Computer Science and Mathematics, vol. 5, no. 4, pp. 12–26, 2024.
10. H. Murti, S. Sulastri, D. B. Santoso, and D. A. Diartono, "Performance comparison of SVM, Naive Bayes, and Random Forest models in fake news classification," Engineering and Technology Journal, vol. 9, no. 8, pp. 4799–4804, 2024.
11. A. Saeed and E. Al Solami, "Fake news detection using machine learning and deep learning methods," Computers, Materials and Continua, vol. 77, no. 2, 2023.
12. S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments," Sensors, vol. 23, no. 3, Art. no. 1748, 2023.
13. S. Tiwari and S. Jain, "Fake news detection using machine learning algorithms," in Proc. KILBY 100 7th Int. Conf. Computing Sciences (ICCS 2023), Phagwara, India, 2024.
14. M. V. Rampurkar and D. D. Thirupurasundari, "An approach towards fake news detection using machine learning techniques," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, pp. 2868–2874, 2024.
15. A. Khalil, F. Metti, M. M. Rahhal, and D. Azar, "Ensemble based high performance deep learning models for fake news detection," Scientific Reports, vol. 14, Art. no. 26591, 2024.
16. E. M. Mahir, S. Akhter, and M. R. Huq, "Detecting fake news using machine learning and deep learning algorithms," Neurocomputing, vol. 530, pp. 91–103, 2023.
17. T. W. Teo, H. N. Chua, M. B. Jasser, and R. T. K. Wong, "Incorporating LLM judgment with conventional machine learning in fake news detection," in Proc. 20th IEEE Int. Colloquium Signal Processing and Its Applications, 2024.
18. E. S. Hamsheen and L. R. Flah, "Fake news detection using machine learning approaches," ZANCO Journal of Pure and Applied Sciences, 2023.

19. D. Mouratidis, M. N. Nikiforos, and K. L. Kermanidis, "Comparative analysis of classical and deep learning classifiers for fake news detection," Information Processing & Management, vol. 60, no. 3, 2023.
20. L. Shen et al., "GAMED: Knowledge adaptive multi-experts decoupling for multimodal fake news detection," in Proc. 18th ACM Int. Conf. Web Search and Data Mining (WSDM '25), 2025, pp. 586–595.