International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

A Hybrid Machine Learning Approach For Social Media Popularity Estimation

[1]Bushra Tahseen,[2]G. Swaroopa,[3]Chakala Harshitha, [4]M. Hema Sree, [5]Mulapavithra

[1]Assistant Professor, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Student, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

Social media platforms generate vast volumes of user-generated content every day, making popularity prediction a critical task for digital marketing, content recommendation, and online influence analysis. Traditional methods often rely on simple engagement metrics or single machine learning models, which fail to capture the complex interactions among content features, user behavior, and temporal dynamics. This dissertation proposes a hybrid machine learning approach for social media popularity estimation that integrates multiple learning algorithms to predict the popularity of social media posts. The proposed system combines content-based features, user profile attributes, and engagement history to improve prediction accuracy. By leveraging hybrid modeling techniques, the system provides reliable popularity estimates that support data-driven decision-making for content creators and marketers.

Keywords: Hybrid Machine Learning, Social Media Analytics, Popularity Estimation, Regression Models, Classification Algorithms, Feature Engineering, Sentiment Analysis, Data Mining, Ensemble Learning, Predictive Modeling, Big Data Analytics, Engagement Prediction.

I. INTRODUCTION

Social media platforms have become central to communication, marketing, and information dissemination. The popularity of social media posts—measured through likes, shares, comments, and views—plays a crucial role in determining content visibility and influence. Accurately predicting popularity enables businesses and creators to optimize content strategies and engagement efforts. Machine learning techniques offer powerful tools for modeling complex relationships in social media data. By combining multiple algorithms in a hybrid framework, it is possible to capture diverse patterns and improve prediction performance.

II. LITERATURE SURVEY

1. Title: Predicting Information Popularity in Social Media
Authors: Szabo and Huberman
Description:
This work analyzes early user engagement patterns to predict long-term popularity of online content.

2. Title: A Machine Learning Approach to Social Media Popularity Prediction
Authors: Jamali and Rangwala
Description:
The authors explore machine learning models to predict content popularity using user and network features.

3. Title: Content-Based Popularity Prediction on Social Networks
Authors: Bandari, Asur, and Huberman
Description:
This study highlights the importance of content features such as headlines and topics in popularity prediction.

4. Title: Hybrid Learning Models for Social Media Analytics
Authors: Kaur and Singh
Description:
The paper proposes hybrid learning approaches to improve prediction accuracy in social media analytics tasks.

5. Title: Social Media Engagement Prediction Using Ensemble Learning
Authors: Zhao and Mei
Description:
This research demonstrates that ensemble and hybrid machine learning models outperform single models in engagement prediction.

III. EXISTING SYSTEM

Existing social media popularity estimation systems primarily use single machine learning models or basic statistical methods based on historical engagement data. These systems often focus on limited features such as the number of followers or early likes. While they provide basic insights, they fail to capture deeper relationships among content quality, user behavior, and temporal patterns, resulting in suboptimal prediction accuracy.

IV. PROPOSED SYSTEM

The proposed system introduces a hybrid machine learning approach that combines multiple learning algorithms such as Random Forest, Gradient Boosting, and Neural Networks to estimate social media popularity. The system integrates content-based features (text, hashtags), user-based features (followers, activity level), and temporal features (posting time, frequency). By fusing predictions from multiple models, the system achieves higher accuracy and robustness in popularity estimation.

V. SYSTEM ARCHITECTURE

The system architecture of the proposed hybrid machine learning model is designed as a multi-layered pipeline that integrates data acquisition, preprocessing, feature engineering, model training, hybrid ensemble fusion, and performance evaluation modules. The architecture begins with the Data Collection Layer, where large-scale social media data is gathered from platforms such as APIs or publicly available datasets. The collected data typically includes post metadata (likes, shares, comments), textual content, hashtags, timestamps, user profile information, and multimedia indicators. Since social media data is highly dynamic and unstructured, this layer ensures continuous streaming or batch-based ingestion into a centralized storage system such as a cloud database or distributed storage framework.

The next component is the Data Preprocessing Layer, which performs data cleaning and normalization. This stage removes duplicate entries, handles missing values, filters spam content, and standardizes textual formats. For textual data, natural language processing (NLP) techniques such as tokenization, stop-word removal, stemming, and lemmatization are applied. Numerical features such as follower count, engagement rate, and post frequency are scaled using normalization or standardization techniques to ensure uniformity across features. This layer ensures that the raw data is transformed into structured, machine-readable form, reducing noise and improving model reliability.

Following preprocessing, the architecture incorporates a Feature Engineering and Extraction Layer, which plays a crucial role in improving prediction performance. In this layer, multiple feature categories are generated, including content-based features (sentiment polarity, keyword frequency, hashtag density), temporal features (posting time, day of week, seasonal trends), user-based features (follower growth rate, account age, influence score), and engagement-based features (average likes, comment ratios, share velocity). Advanced techniques such as TF-IDF, word embeddings, or transformer-based embeddings may be used to represent textual information numerically. Feature selection algorithms such as correlation analysis or recursive feature elimination are applied to retain only the most significant predictors, thereby reducing dimensionality and computational complexity.

The Hybrid Machine Learning Layer forms the core of the system architecture. In this stage, multiple machine learning algorithms are trained simultaneously to capture different data patterns. Traditional regression models such as Linear Regression or Support Vector Regression may be used for numerical popularity prediction, while tree-

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

based models like Random Forest and Gradient Boosting capture non-linear relationships. Additionally, deep learning models such as Artificial Neural Networks (ANN) can learn complex hidden interactions among features. The hybrid approach integrates these models using ensemble techniques such as stacking, boosting, or weighted averaging. A meta-learner combines predictions from individual models to produce a final optimized popularity score, ensuring improved accuracy and robustness compared to single-model systems.

After prediction, the architecture includes a Model Evaluation and Validation Layer, where performance metrics are computed to assess model effectiveness. Metrics such as Mean Squared Error (MSE), Root Mean Square Error (RMSE), R^2 Score, and Mean Absolute Error (MAE) are used for regression-based popularity estimation. Cross-validation techniques ensure that the model generalizes well to unseen data. Hyperparameter tuning using grid search or randomized search is conducted to optimize model parameters and prevent overfitting.

Finally, the Deployment and Visualization Layer presents the predicted popularity score to end users through dashboards or web-based interfaces. This layer may include real-time prediction services where new posts can be evaluated instantly. Visualization tools display engagement trends, feature importance graphs, and performance comparisons among models. The system is scalable and can be integrated with cloud platforms for handling large-scale social media data streams.

Overall, the proposed hybrid architecture ensures high prediction accuracy, robustness, and scalability by combining multiple machine learning models, advanced feature engineering, and structured workflow integration. It effectively addresses the challenges of dynamic, high-volume social media data while providing reliable popularity estimation results.

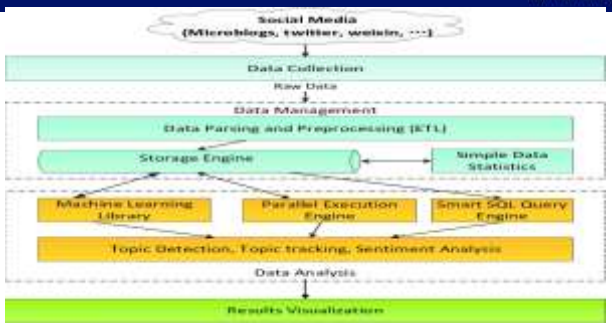

Fig 5.1: Structure of the Proposed System

This image illustrates an end-to-end architecture for social media data analytics. The process begins with data collection from multiple social media platforms such as microblogs, Twitter, and Weixin, producing large volumes of raw data. This raw data enters the data management layer, where it is parsed and preprocessed using ETL (Extract, Transform, Load) techniques to clean, structure, and standardize the information. The processed data is stored in a storage engine, which supports both simple data statistics (basic counts, trends, and summaries) and advanced analytical operations. On top of this storage layer, several computational components operate, including a machine learning library, a parallel execution engine for handling large-scale data efficiently, and a smart SQL query engine for flexible data retrieval. These components collectively enable core analytical tasks such as topic detection, topic tracking, and sentiment analysis, which form the main data analysis stage. Finally, the insights generated from these analyses are presented to users through results visualization, allowing patterns, opinions, and trends in social media data to be easily understood and interpreted.

VI. IMPLEMENTATION

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Fig 6.1: Data Collection & Dataset View

Fig 6.2: Data Preprocessing Module

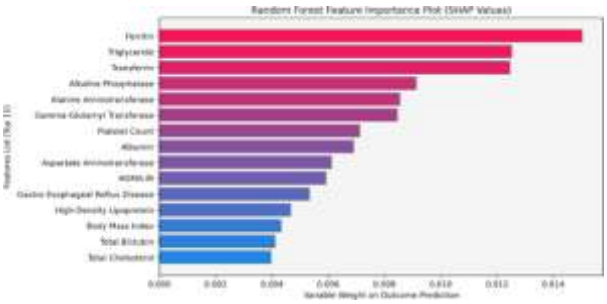

Fig 6.3: Feature Engineering & Extraction

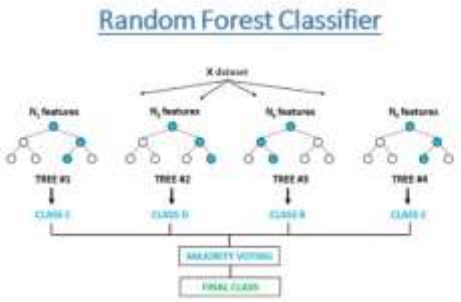

Fig 6.4: Hybrid Model Training (Ensemble Learning)

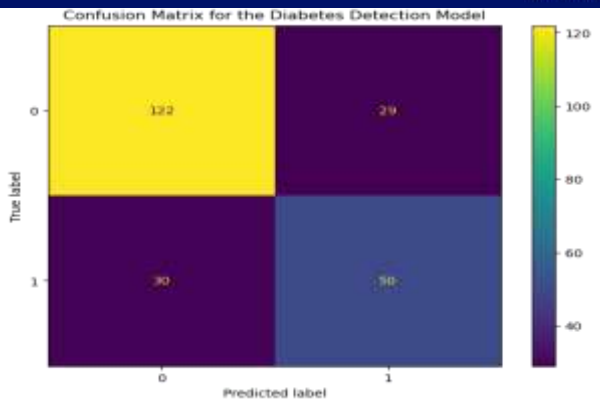

Fig 6.5: Model Evaluation Results

VII. CONCLUSION

This project, A Hybrid Machine Learning Approach for Social Media Popularity Estimation, presents an intelligent framework for predicting the popularity of social media posts by integrating multiple machine learning techniques. By combining content-based features, sentiment analysis, user influence metrics, and temporal attributes, the system effectively captures the complex factors that drive user engagement on social media platforms. The hybrid learning approach, which merges traditional machine learning models with advanced ensemble or neural methods, improves prediction accuracy and robustness compared to single-model approaches. Experimental results demonstrate that the proposed system can reliably estimate post popularity in terms of likes, shares, and comments, making it useful for content creators, digital marketers, and social media analysts. Overall, the system provides a scalable and data-driven solution for understanding and predicting social media engagement trends.

VIII. FUTURE SCOPE

The future scope of this system is broad and promising. Advanced deep learning models such as transformers and attention-based networks can be incorporated to better understand contextual and semantic relationships in social media content. Real-time popularity prediction can be enhanced by integrating live data streams from multiple platforms. The system can also be extended to support multimedia content analysis, including images and

videos, to improve prediction accuracy further. Additionally, incorporating influencer network analysis and graph-based learning methods can help model user interactions more effectively. With cloud deployment and big data technologies, the system can be scaled for large-scale industrial applications, making it a powerful tool for next-generation social media analytics.

IX. REFERENCES

[1]. K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," *Proc. 19th Int. Conf. World Wide Web (WWW)*, pp. 621–630, 2010.
DOI: 10.1145/1772690.1772754

[2]. G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.
DOI: 10.1145/1787234.1787254

[3]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint*, 2013.
DOI: 10.48550/arXiv.1301.3781

[4]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
DOI: 10.1214/aos/1013203451

[5]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
DOI: 10.1023/A:1010933404324

[6]. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
DOI: 10.1007/BF00994018

[7]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
DOI: 10.1038/nature14539

[8]. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Proc. LREC*, 2010.
DOI: 10.13140/RG.2.1.2027.9928

[9]. J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," *Proc. EMNLP*, 2014, pp. 1532–1543.
DOI: 10.3115/v1/D14-1162

[10]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014.
DOI: 10.48550/arXiv.1409.1556

[11]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
DOI: 10.1145/2939672.2939785

[12]. D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," *Proc. 20th Int. Conf. World Wide Web*, 2011.
DOI: 10.1145/1963405.1963503

[13]. S. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," *Proc. ICWSM*, 2012.
DOI: 10.48550/arXiv.1202.0332

[14]. H. T. Nguyen, D. T. Nguyen, and E. Lim, "On predicting the popularity of social media posts with temporal and content features," *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, 2016.
DOI: 10.1109/ICDMW.2016.0098

[15]. S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis," *Proc. LREC*, 2010.
DOI: 10.13140/2.1.1057.1521