Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27248

Social Media Diffusion Modeling: Validating Social Media Signals for E-commerce Trend Prediction

Xiaolin Chen

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
t330026022@mail.uic.edu.cn

Abstract. In today's digital age, social media has become an indispensable link between consumers and brands. With the popularity of the Internet and the vigorous development of social media platforms, e-commerce enterprises have set their sights on this new and dynamic marketing field, hoping to enhance brand influence, expand user groups and ultimately achieve sales growth through the power of social media. This article systematically reviews the methods to build social media information diffusion model. Focus on prediction model, network analysis and time series modeling as well as compare the traditional modeling and deep learning method. This article also discusses Hawkes process; survival analysis and other time series modeling techniques apply scenarios. The performance of the model is evaluated through indicators such as computational complexity and interpretability, and a hybrid architecture as well as online learning mechanism are proposed to address the challenges of unsteady data. In addition, it emphasized future directions such as multimodal fusion, cross-platform generalization, and privacy protection, and called for the establishment of standardized tools and benchmark datasets to promote the development of the field.

Keywords: Social media diffusion model, Long Short-Term Memory, Graph Neural Network, Hawkes process, Business intelligent

1. Introduction

Social media's information diffusion has the complex characteristics of high-dimensionality, dynamics and nonlinearity. With the increasing popularity of the platform such as Twitter and Facebook, one single topic's diffusion may cover one hundred thousand users' nodes, which reveal social media's high- dimensionality [1]. Social media's information diffusion also has topological structure, such as the interaction between highly creative users and edge users, which directly influence information spread path and efficiency. Dynamics represents in the multi-modal temporal pattern of the diffusion process and the time-varying characteristics of user activity [2]. What's more, the coupling effect between user behavior and network structure such as the long-term dependency catch by Long Short-Term Memory (LSTM) model is quite different with the traditional logical linear regression model's hypothesis, which shows information diffusion's nonlinearity. These characteristic causes traditional diffusion theories like Two-Step flow of communication hard

DOI: 10.54254/2755-2721/2025.LD27248

to provide countable quantitative framework. Thus, to correctly predict how social media influence commodity sales volume, algorithm and scalable computing should be managed.

To predict precisely, need to deal with three main predicaments: The first one is Dynamics Predicament. Traditional temporal model may produce high-error prediction while a multipeak outbreak shows up [1]. To handle this predicament, may need to introduce self- excitation processes like Hawkes and neural temporal networks. Additionally, Data Defect Dilemma is another challenge that need to be focus on. To deal with annotation bias and sparsity data [3], promote the development of semi-supervised learning and graph data augmentation techniques should be implemented. The last one is Interpretability Problem that balance is required between prediction accuracy and decision-making transparency in business scenarios.

This paper focuses on study of three types of data science methods, which can help to enforce the predict result: Predictive models (such as APR and unsupervised clustering in supervised learning), network analysis which influence maximization and community detection, and time series modeling cover techniques ranging from classical statistics to cutting-edge deep learning. Technical implementation details and empirical verification are greatly emphasized in this article, which is quite different from the theoretical-only paper. The review structure is classified and unfolded by methodology, forming a progressive logic of model design, verification framework and open challenge", aiming to provide researchers with a reusable guide for technology selection. Through integrating a multi-disciplinary perspective, using sociological theory as the background and data science methods as the main thread, this review combs existing technologies as well as reveals the core contradiction: the trade-off between the computational efficiency advantages of traditional models in simple scenarios and the precise modeling capabilities of hybrid models in complex dynamics. This contradiction has driven the development of adaptive frameworks, which dynamically integrate the advantages of multiple models while ensuring real-time performance.

2. Methodological taxonomy

Social media's information diffusion model can be managed systematically into a three-dimensional framework. One is Prediction Model which is used to identify what kind of user will spread information and the information they diffuse will affect which group of people, reaching the aim of stimulating consumption. The second one is Network Analysis whose aim is to enhance the predict result. To reach this aim, information spread route should be analysis. Once the route be correctly parsing, the model can catch the earliest signal of a hit product emerging. The last one is Temporal Series Modeling. Depict dynamic laws to predict when the sales will reach the peak as well as when the sales trend will start to decline.

This section, will integrate algorithm principles, technical implementations, and scenario adaptability, deconstructing the intrinsic logic and evolutionary relationships of various methods.

2.1. Prediction model to identify high-impact communicators

The core mission is to quantify the potential of user dissemination and identify early key nodes.

2.1.1. Supervised learning model

This section will introduce one classic APR model, which is study by Imamori et al. [4]:

$$L = -\sum_{v \in V} y_v \log(p_v) + \lambda \|\vartheta\| \quad (1)$$

DOI: 10.54254/2755-2721/2025.LD27248

The first part of the equation is the cross-entropy loss, which forces the prediction probability p_v approaching the true label, where $y_v = 1$ represent for the early accepter. The second part is the regularization for the L2, keeping feature weight's complex rate θ into control, prevent overfitting. This model with the advantage of strong interpretability such as the contribution rate of Posting frequency quantified by SHAP value reaches 32%. However, this model relying on manual definition of early adopter ratings", resulting in poor generalization of the cold start scenario.

2.1.2. Supervised learning model

To solve the problem, discover in supervised model, some researchers try to use unsupervised method to refine it. This section will introduce the state gradient model [2].

$$f_w(t) = H(g_w(t)) \tag{2}$$

In this equation gw(t) represents the proportion of elite users and H(x) stand for the helveside step function. The mark critical point of the spread explosion will show up when gw(t) > 0.7 at the same time fw(t) = 1. State gradient model is used in identify the elite group during the cold start period. Whereas, this model cannot quantify individual influence. Insert Metric learning to optimize the clustering distance may improve this model.

2.1.3. Deep learning model

To solve the limitations of static network representation that traditional method like Node2Vec only produces static user embedding, cannot reflect the temporal evolution of user behavior. What's more, some model ignore users are in the network topology with structural position. For instance, through two users have same posting frequency, if network centrality is different, their comment's influence may be quite different. Thus, LSTM-DeepWalk framework was created to handle the problem below [3]. DeepWalk Objective Function:

$$\max_f \sum_{u \in V} \log P(N(u) \mid f(u)) \tag{3}$$

LSTM Temporal Series Modeling:

$$h_t = \sigma(W_h [h_{t-1}, x_t] + b_h) \tag{4}$$

LSTM is used in Twitter rumor spread prediction, it can successfully provide the rumor warning two hours in advance, whose accuracy reaches 89%. Applicable to multiple platforms is another advantage that keep LSTM method outstanding. In the migration task from Twitter to Reddit, LSTM's AUC remained at 0.82.

2.2. Network analysis of diffusion paths

Analysis information diffusion path and influence mechanism can greatly help to find the prospective customer and prepare a reasonable amount of goods in reserve.

DOI: 10.54254/2755-2721/2025.LD27248

2.2.1. Graph representation learning

The most common graph representation learing method is Node2Vec algorithm, which can handle traditional centrality metrics (such as degree centrality) cannot capture complex topological patterns problem. This algorithm usually be applicated in 'Structural Hole Spanners' and its accuracy of user classification has increased by 23% compared with manual feature engineering.

2.2.2. Dynamic GNN message passing

Traditional graph neural network (GNN) requires static graph input, while this method is handling topological evolution through timestamp edges (Et).

$$m_u^{(k)} = AGG\left(h_v^{(k-1)} : v \in N\left(u\right)\right) h_u^{(k)} = UPDATE\left(h_u^{(k-1)}, m_u^{(k)}\right) \tag{5}$$

2.2.3. Greedy algorithm

The research on the problem of maximizing influence, from theoretical guarantee to engineering implementation, reflects the complete evolution path of algorithm design from theoretical optimality to computational feasibility. The greedy algorithm, through theoretical proof (equation 6), ensures at least 63% of the optimal propagation range, providing a strict mathematical guarantee for seed selection. However, its complexity of O(kn2) limits its application in large-scale networks.

$$\sigma(S) \geq \left(1 - \frac{1}{e}\right)\sigma(OPT) \approx 0.63\sigma(OPT) \tag{6}$$

2.3. Dynamic law characterization

Using model to predict the information diffusion's explosion, persistence and attenuation process can help marketers making appropriate decisions.

2.3.1. Node process model

Traditional Hawkes need to preset attenuation function, to solve this neural Hawkes process (NHP) was invented:

$$\lambda(t) = \sigma(W_\lambda h_t + b_\lambda) \tag{7}$$

$$h_t = GRU(x_t, h_{t-1}) \tag{8}$$

NHP using deep learning and GRU network dynamic learning of event dependencies, catch multimodal propagation patterns and nonlinear temporal dependencies automatically as well as improving the training efficiency to 40% through stochastic gradient optimization.

2.3.2. Survival analysis

This method uses Cox proportional hazard model, which provides a powerful tool for quantitative analysis of the communication life cycle in information dissemination research. This function baseline risk h(t) and covariate effect were effectively distinguished, which handle the difficulty to separate the influence of time factors and features. while estimating β, researchers can accurate

DOI: 10.54254/2755-2721/2025.LD27248

quantitative characteristics of different degree of contribution to the communication process, which show the characteristic of timeliness is the main driving force for early dissemination.

$$h(t|X) = h_0(t)\exp(\beta^T X) \tag{9}$$

3. Model comparison and validation

In social media's information diffusion modeling area, there are several different methods. Each method has its advantage and shortage. This section through the comparison of technical indicators and the empirical verification framework, the performance, applicable scenarios and verification methods of mainstream models are systematically evaluated. Combined with actual cases and cutting-edge research, a scientific basis is provided for method selection.

3.1. Model technology comparison

In the methodological comparison of social media information diffusion modeling, the three core models exhibit significant technical differences [5-7]. First of all, APR model which is based on supervised learning like logic regression model, has a linear computational complexity of O (n features). APR model can provide explicble SHAP analysis and it is suitable for Small-scale labeled data scenarios such as user behavior prediction. However, APR model with the limitations of rely on artificial feature engineering. Additionally, GNN equipped with computational complexity increases in a square order with the number of edges and it use attention wight visualization technique to analyzed yje interaction patterns between nodes, making it an ideal choice for modeling the diffusion path of dynamic networks. For instance, GNN is quite suitable in Twitter topic dissemination's real-time tracking. At last, although Hawkes process facing high computational burden of O (n events^2), the stimulant effects between events can be effectively quantified through parameter significance tests, which has irreplaceable advantages in the multimodal time series modeling of sudden events such as the outbreak of social media rumors.

3.2. The empirical case

3.2.1. Demand forecasting: sales modeling driven by social media information diffusion

In e-commerce field, one essential predictive indicator is the content that post by social media user (UGC). Wang et al. shows that while using XGBoost ensemble learning method to analysis Amazon product reviews, the importance of feature social signals such as emotional polarity and comment activity is significantly higher than traditional historical sales data [8].

This study uses VADER to analysis users' comment to identify their emotional trend and extract temporal features such as fluctuations in daily comment volume and network features like user interaction relationships to comprehensively capture social signals. During the model building period, XGBoost algorithm integrates 20-dimensional social features such as sentiment score and comment length, as well as traditional business variables. The feature importance analysis shows that the SHAP value contribution of sentiment polarity (35%) far exceeds that of historical sales data (12%), highlighting the predictive value of social data. The performance verification results show that the AUC value of this model in the 30-day sales forecast reaches 0.82, which is 15.5% higher than that of the traditional ARIMA model. Among them, for a certain Bluetooth headset case, it is shown that the model can monitor a 10% sudden increase in negative reviews and warn of 22% sales decline

DOI: 10.54254/2755-2721/2025.LD27248

two weeks in advance (the prediction error is controlled within ±3%). It fully verified the forward-looking guiding role of social media data in business decisions [8].

3.2.2. Stock price prediction: time series modeling of multimodal social data

Research shows that the response of the financial market to social media information shows significant nonlinear characteristics. The BERT-GRU hybrid architecture developed by Ho et al. innovatively integrates the semantic features of Reddit discussion posts (through FinBERT sentiment analysis) with stock price time series data (15-minute window alignment), achieving breakthrough progress in Tesla stock price prediction [9]. This model extracts 768-dimensional text features through the BERT encoding layer, combines the GRU time series layer to process historical stock prices, and uses a 5-day sliding window to capture the influence of social hotspots. Eventually, the prediction error (RMSE) is reduced to $1.87, which is 12% higher than that of the traditional LSTM. Typical cases show that the model can accurately capture the sudden changes in market sentiment triggered by Elon Musk's tweets (negative sentiment surges by 35 percentage points within 24 hours), and successfully predict a 7.2% decline in stock prices the next day (the actual decline is 7.8%).

However, such models are confronted with the challenge of unsteady data. To address this issue, Hawkes process was used. By adjusting the parameters of event excitation intensity (α) and attenuation rate (β) in real time, the prediction error can be precisely controlled within ±1.5% in unexpected events such as Robinhood trading restrictions. It provides an effective solution for dealing with sudden changes in social data in the financial market. These studies collectively demonstrate that a hybrid architecture combining deep learning and time series modeling, along with a dynamic parameter adjustment mechanism, can significantly enhance the accuracy of financial predictions based on social media.

4. Challenges and future directions

Combining social media analysis to business intelligence (BI) facing plenty of challenges, including data dynamic, algorism adoption as well as ethical issues. Solving these questions are crucial for model's robustness and scalability.

4.1. Unsteady distribution and multimodal fusion

TSocial media data represents unsteady temporal dynamics. For instance, some research shows that different topic has different information dissemination model, which requires model equipped with dynamic adaptability. Online learning frameworks (such as dynamic Hawkes processes) can respond to the evolution of data distribution by updating parameters in real time (the sixth reference), which is a useful way to improve adaptability. For instance, one efficiency way to increase accuracy of capture the sudden or attenuating trends of user engagement is employing time-varying infection rates in epidemic models [3].

Otherwise, social media data is essentially multimodal, covering text, images and network structures. Traditional single-modal methods such as those relying solely on sentiment analysis, are difficult to fully utilize the complementarity of multi-source signals. Multimodal fusion architectures can com- bine text embeddings like BERT and graph neural networks (GNN) to jointly model content and user interaction. One example is to, combine tweet text analysis (BERT) with User

DOI: 10.54254/2755-2721/2025.LD27248

Interest Network (GNN), which can predict the information dissemination path or user influence with high accuracy [8].

4.2. Algorithm challenges: generalization and domain adaptation

Nowadays, many models have limitations of high platform dependency, such as the model trained by data from Twitter may be not suitable and output high-error result while used on other platforms such as Reddit or Facebook, with the reason of differences in user behavior, network topology and content specifications [5]. Thus, the method like cross-platform domain adaptive techniques (adversarial training or feature alignment) which can enhance the generalization ability of the model should be apply. For example, through unsupervised domain adaptation methods, the user behavior patterns learned on Twitter were transferred to Reddit while retaining platform-specific features.

Another challenge is the insufficient explanatory power of graph neural networks. Although GNNS perform well in social network analysis especially identifying key propagation nodes, their decision- making process is often regarded as a “black box”. Future research can combine explainable AI technologies such as attention mechanisms or subgraph extraction to reveal key user or content features in information dissemination.

4.3. Ethical challenges: privacy and algorithmic bias

The abuse of social media data may trigger serious privacy issues. For instance, the Cambridge Analytica incident [10] exposed the risk that user data could be used to manipulate political leanings. Federated learning and differential privacy are potential solutions. For example, introducing differential noise in user embedding training can protect sensitive features from being cracked by reverse engineering [11].

Furthermore, algorithms may magnify social biases. One example is that the targeted push of recruitment advertisements may discriminate against specific groups due to historical data bias [12]. Future research may need to develop an optimization framework with fairness constraints to ensure the neutrality of the model in dimensions [13].

4.4. Future direction

In the future model building, the model should combine reinforcement learning and online learning to achieve real-time adjustment. Furthermore, build a knowledge graph that integrates text, images and user relationships to support more complex BI tasks should be executed. Another important issue is ethics problem. The future study needs to promote the formulation of industry standards to ensure that the data usage is complied with regulations.

5. Conclusion

The application of social media analysis in business intelligence (BI) has demonstrated great potential, but an appropriate methodological framework should be selected based on the complexity of the task. This essay reveals the key challenges and technical paths through a systematic review, and provides the following guidance for research and apply:

For simple tasks like monitoring user sentiment trends, models with strong interpretability should be given priority, such as the APR framework based on the attention mechanism. Additionally, supplemented by traditional statistical verification should be used to ensure the robustness of the results.

DOI: 10.54254/2755-2721/2025.LD27248

When facing complex scenarios such as cross-platform information dissemination prediction, it is necessary to combine the hybrid model of GNN and Hawkes Process, and apply an online learning mechanism to adapt to dynamic changes.

Another advice is to standardize the tools, which means to develop benchmark data and construction (integrate the information diffusion trajectories of multiple platforms, covering the triples of text, user relationships and timestamps) as well as lightweight algorithm libraries (merge NetwoekX and PyTorch temporal), building an end-to-end analysis toolchain.

The future of social media analysis is focus on dynamics, multimodality and accountability coordinated develop. On the one hand, a more flexible online learning architecture needs to be developed to deal with data instability; On the other hand, industry ethical guidelines should be established to ensure the mandatory application of technologies. Only merge methodological innovation and standardized tools can the value of social media data in business decision-making be fully unleashed.

References

[1] Guille, A., & Hacid, H. (2012). A predictive model for the temporal dynamics of information diffusion in online social networks. WWW 2012 – MŚND'12 Workshop, 1145-1152.
[2] Rotabi, R., & Kleinberg, J. (2016). The Status Gradient of Trends in Social Media. arXiv.Org.
[3] Stai, E., Milaiou, E., Karyotis, V., & Papavassiliou, S. (2018). Temporal dynamics of information diffusion in Twitter: Modeling and experimentation. IEEE Transactions on Computational Social Systems, 5(1), 256-264. https: //doi.org/10.1109/TCSS.2017.2784184
[4] Imamori, D., & Tajima, K. (2016). Predicting Popularity of Twitter Accounts through the Discovery of Link-Propagating Early Adopters. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 24-28-, 639–648. https: //doi.org/10.1145/2983323.2983859
[5] Choi, J., Yoon, J., Chung, J., Coh, B.-Y., & Lee, J.-M. (2020). Social media analytics and business intelligence research: A systematic review. Information Processing and Management, 57(3), 102279. https: //doi.org/10.1016/j.ipm.2020.102279
[6] Trappey et al. (2018). Consumer driven product technology function deployment using social media and patent mining. Advanced Engineering Informatics 36, 120–129.
[7] Li, C.-T., Lin, Y.-J., & Yeh, M.-Y. (2018). Forecasting participants of information diffusion on social networks with its applications. Information Sciences, 422, 432–446. https: //doi.org/10.1016/j.ins.2017.09.034
[8] Wang et al. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. Electronic Commerce Research and Applications 29, 142–156.
[9] Ho, T. K., Yoon, S., & Lee, J. (2021). FinBERT-GRU: A deep learning approach for stock price prediction using social media sentiment and time-series data. Expert Systems with Applications, 184, 115537. https: //doi.org/10.1016/j.eswa.2021.115537.
[10] Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica. The Guardian.
[11] Abadi, M., et al. (2016). Deep Learning with Differential Privacy. ACM CCS.
[12] Lasmi, H., Lee, C. H., & Ceran, Y. (2021). Popularity Brings Better Sales or Vice Versa: Evidence from Instagram and OpenTable. Global Business Review. https: //doi.org/10.1177/09721509211044302
[13] Mehrabi, N., et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys.