AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 58(1) (2026) 125-136
DOI: 10.22060/miscj.2026.24962.5444

Generalized Content-Aware Diffusion Models for Twitter

Farzaneh Ghayour-Baghbani *

Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.

ABSTRACT: Influence and information diffusion are important fields in social network analysis. Diffusion models play a key role in predicting the propagation behaviour of messages and are a crucial prerequisite in solving other social network problems. In this research, we focus on propagation models applicable to the Influence Maximization Problem, particularly the widely used Threshold and Cascade models, along with their various extensions. These models attempt to simulate how users adopt information based on network structure and peer influence. We evaluate the effectiveness of these models in predicting the propagation behavior of tweets within the Twitter network. Our analysis reveals that while each model can accurately predict the spread of certain tweets, they fail to generalize across diverse content types. This inconsistency suggests that message content significantly influences propagation dynamics. To address this limitation, we propose two novel content-aware generalized propagation models that incorporate message semantics into the prediction process. These models dynamically select the most suitable propagation mechanism based on the tweet's content characteristics. Experimental validation using real-world Twitter datasets demonstrates that our approach significantly improves prediction accuracy and offers a more robust framework for modeling information diffusion. Our findings highlight the importance of integrating content analysis into diffusion modeling for more effective influence strategies.

Review History:

Received: Oct. 22, 2025
Revised: Jul. 10, 2026
Accepted: Jul. 12, 2026
Available Online: Jul. 20, 2026

Keywords:

Social Networks
Influence Flow
Propagation Models
Content
Influence Maximization

1- Introduction

Social network analysis has attracted many researchers in recent years. Influence and information diffusion are two of the main fields in this area. There are different diffusion models that try to describe the flow of information and the influence of users on each other within social networks. These diffusion models are the basis for solving other important problems within social networks.

A social network is modelled as a graph $G = (V, E)$ called a social graph. V is the set of vertices (nodes) showing the members of social networks, and E denotes the set of edges that represent relations between members, which could be directed or undirected. Each node could be active or inactive. Being (not) active means that the corresponding member has (not) received a piece of information, has (not) adopted an idea, has (not) bought an item, has (not) caught a virus, or other similar situations. Diffusion models show how nodes become active or inactive in time due to interactions between nodes. If a node remains active till the end of time after the node has become active, the diffusion model is called progressive; but if a node could change between being active and inactive, the model is called non-progressive. (Chen, Lakshmanan, and Castillo 2013). Progressive models are suitable for applications such as receiving data or buying goods, whereas non-progressive models are appropriate for applications such as adopting an idea. In this paper, we focus on progressive models.

A diffusion model could be seen as a binary classifier predicting which nodes would become active. In other forms, it outputs the expectation that a node would become active. A diffusion model also contains activation time and information about the influence of nodes on the newly active nodes.

The Linear Threshold (LT) model and Independent Cascade (IC) models are two classic probabilistic diffusion models. (Kempe, Kleinberg, and Tardos 2005). The LT(IC) model receives the social graph G, the weight (probability) of each edge, and a seed set, and outputs S_t for all $t>0$. In the LT model, each inactive node could be activated if the sum of the weights of its active in-neighbours is greater than a random threshold. In the IC model, each active node activates its outgoing neighbours with a specific probability.

The Influence Maximization problem is the problem of finding a fixed-size set of users so that the spread of information beginning from this set is maximized on the network, and it has many applications, such as viral marketing (Chen, Lakshmanan, and Castillo 2013). Many

*Corresponding author's email: farzane.ghayour@gmail.com

recent researches try to model the propagation of information in social networks (Yan, F., Liang, W. and Hirota, K. 2022; Hong, Z., Zhou, H., Wang, Z., et al. 2023; Li, Y., & Han, Z. 2024) but only the Threshold and Cascade models like LT and IC and their extensions could be used for Influence Maximization problem (Chen, Lakshmanan, and Castillo 2013).

In this paper, we analysis twitter data and the propagation behaviour of the tweets. We observe that each of the existing propagation models capable of expressing the Influence Maximization problem could predict the behaviour of some of the tweets but fails to predict the behaviour of the others. In other words, tweets may have different behaviours, even if they start propagation from the same point in the network structure. One main property of each tweet is its content, which is ignored in classic Threshold and Cascade propagation models. We analyse the data to observe how the content of the tweets and its relation to the content of the network could affect their propagation behaviour. Finally, we suggest two Content-Aware Generalized propagation models which could express different propagation behaviour of different tweets.

The proposed generalized propagation models assign a content score to every node in a network. This score states how much the node is similar or related to the tweet for which we want to predict its propagation behaviour. Using content scores, we will predict the popularity of the tweet and select an appropriate propagation model for it.

For computing the content similarity scores between a node and a message, we use Information Retrieval techniques. Information retrieval is a broad field with many applications such as web searching, expert finding, recommender systems, summarization, spam filtering, document classification, etc. (Manning, Raghavan, and Schütze 2013). We concatenate all the text messages of each user into one document. Each test message is considered a query, and each document is assigned a score by performing the query on the index of documents. In this paper, we use the basic scoring method of Lucene, which is based on TF-IDF.[1].

There are related studies that try to model the sharing behaviour in networks based on user messages and message content, but none of them presents a propagation model considering both network structure and content of the messages. (Macskassy and Michelson 2011; Petrovik, Osborne, and Lavrenko 2010; Pramanik and Guillaume 2017; Ye and Wu; Lin, Mei, and Han 2011; Kupavskii et al. 2012; Challenge 2010; Anh et al. 2018) . These studies show that the sharing behaviour of individuals is related to the similarity between the content of the message and the messages of the user. (Macskassy and Michelson 2011).

Therefore, the core contribution of this work lies in addressing the temporal and semantic limitations of current popularity prediction frameworks through a dual-pronged approach:

We employ information retrieval methods to measure content similarity. This dynamic metric serves as a key predictor of tweet popularity.

We introduce a Temporal-Context-Aware (TCA) model that accounts for tweet aging. It also captures the dynamic nature of users' publishing behavior over time.

The paper is organized as follows: Section 2 describes the preliminaries and reviews previous related works. In Section 3, we study the fitness of existing propagation models to predict the propagation of tweets. Sections 4 and 5 describe our methods for diffusion prediction of a message in a social network. Section 6 presents the experimental results of the proposed models. The paper is finalized by a conclusion and discussion about future work.

2- Preliminaries And Related Work

The social network is modelled here as a directed graph. $G=(V,E)$ where V is a finite set of vertices and $E \subseteq V \times V$ is the set of directed edges. Vertices show the individuals in the network, and edges show a social relation (e.g., influence, following) among vertices. Each edge *(u,v)* has a weight (or probability) parameter w_{uv} (or p_{uv}) showing the strength of influence of u on v (Chen, Lakshmanan, and Castillo 2013). Diffusion occurs in discrete time steps $t=0,1,2,\ldots$. Each node $v \in V$ has two possible states: active or inactive. Active nodes at time t are shown by $S_t \subseteq V$ and called the active set at time t. S_0 is called the seed set.

Influence or information diffusion in networks is described by diffusion models. In this paper, we deal with the progressive type of diffusion models in which once a node is activated, it remains active forever. Kempe et al. proposed two progressive diffusion models: the Independent Cascade (IC) model and the Linear Threshold (LT) model. (Kempe, Kleinberg, and Tardos 2003). IC and LT are stochastic models.

In the IC model, each edge *(u,v)* has a probability parameter p_{uv}. If each node u is activated at time t, it could attempt to activate each of its outgoing neighbours, v, at time $t+1$. The node u has a one-time chance to activate its neighbour, v. It is activated with probability $p_{uv,}$ and u does not have any other chance to activate v in the future. (Kempe, Kleinberg, and Tardos 2005).

In the LT model, each edge *(u,v)* has a weight parameter w_{uv}. The sum of the weights of incoming edges for each node is at most 1. Each node v is given a random threshold τ_v from [0, 1] with a uniform distribution. In each time step t, if the sum of the weights of the active incoming edges of an inactive node v is greater than or equal to its threshold τ_v, the node v is activated. IC and LT are not equivalent, but there are the General Independent Cascade model and the General Threshold model, which are equivalent (Kempe, Kleinberg, and Tardos 2005).

In progressive diffusion models, the active set is non-decreasing, and because the number of nodes is finite, diffusion stops in the final time steps. The expected size of the final active set for a specific seed set is called the spread function. (Kempe, Kleinberg, and Tardos 2005).

There are extensions to the LT and IC models that consider different topics on the network; these are called Topic-aware diffusion models (Barbieri, Bonchi, and Manco 2012).

1. https://lucene.apache.org

Barbieri et al. propose the Topic-aware Linear Threshold (TLT) and the Topic-aware Independent Cascade (TIC) diffusion models. These models are extensions of the LT and IC models. These models have the beneficial property of monotone and submodular diffusion, so they are suitable for studying problems like viral marketing. The main weakness of these models is their large number of parameters, which makes them vulnerable to overfitting. Barbieri et al. reported that the TIC and TLT diffusion models do not improve the diffusion prediction of the IC and the LT models; they just improve the prediction of the time of activation (Barbieri, Bonchi, and Manco 2012).

Note that the topic-aware diffusion model does not select between the models based on the topics, and only the parameters of the models are tuned. Barbieri et al. carried out their experiments on networks like Digg[1] and Flixter[2], which are voting networks, and used the common action of users (eg, voting for a common item) to form topics (Barbieri, Bonchi, and Manco 2012). However, we are studying a network containing text messages. So, direct comparison between the Topic-aware model and the Content-Aware model is possible if we can find network data with both topics and text messages, or suitable clustering methods to find topics on network data containing text messages. On the other hand, Barbieri reported that the prediction of Topic-Aware models is as precise as the prediction of classic LT and IC models in specifying whether a node is active or not. Also, their improvement in predicting activation time is rather similar. In this paper, we do not deal with activation time and only deal with the prediction of activation, so it is enough to compare our proposed model with the classic models.

Xiang et al. proposed PageRank with priors as an influence diffusion perspective and showed that it is an approximation of the IC model. Random Walks (RW) could simulate this diffusion process (Xiang et al. 2013).

We compare diffusion models via the F-measure of classification of nodes as active or inactive for specific messages. In the next section, we present data analysis results.

Table 1 shows a comparison of the basic IC and LT models with Topic-aware TIC and TLT models and our proposed model (Temporal Context-aware Generalized Propagation model), which is discussed in the next sections.

Some works analyse propagation on the Twitter network. Pramanik et.al. present an analytical framework for cascade formation in Twitter and compute cascade size (Pramanik and Guillaume 2017). Some works analyse the impact of contextual information on retweets (Suh B et al. 2010; Malhotra A, Malhotra CK and See A 2012). Petrovic et. al. use a machine learning approach based on the passive-aggressive algorithm to predict retweets (Petrovik, Osborne, and Lavrenko 2010).

To the best of our knowledge, none of these works analyse the fitness of the LT and IC propagation model on the Twitter network .Most of the previous works use a structural view or contextual view to analyse propagation. Kupavskii et. al. use contextual information in addition to PageRank, which is a structural view. But they use features like length of the tweet, number of mentions, hashtags, URLs, etc.(Kupavskii et al. 2012) But we use information retrieval methods to compute the content similarity score between tweets. We suggest a hybrid method to differentiate between the tweets by content similarity score and then use an appropriate structural propagation model.

3- Fitness of Existing Propagation Models for Twitter

Due to privacy issues, there is a limited number of datasets that contain both network structure and network content accessible to the public. Hence, we had to crawl Twitter data for our own research work. Due to privacy issues, publishing the text data of Twitter is not allowed, but everyone can collect data. Table 2 shows the specification of our collected dataset.

In order to prepare training data and test data, tweets are sorted by date and time; 10% of the tweets, the latest ones, are selected as test tweets, and the others form the training data.

3- 1- Evaluation Criteria

As we described in the previous section, propagation prediction is a binary classifier that determines for each tweet which nodes (users) become active or do not. There are many criteria for the evaluation of binary classifiers; most of them are based on precision and recall[3]. We use the F-measure as an evaluation criterion. Here, the true positive is the number of users (nodes) that the propagation model predicts will become active, and in gold, data is active too.

3- 2- Learning edge weights

We need to compute the weight of the edges of the network in order to be able to use them for predicting propagation. The weight of edges is learned via the available information of the network. In this task, we use common basic methods for learning edge weights that use the degree of the nodes and other information (Goyal, Lu, and Lakshmanan 2011). These weighing methods are described here:

Degree Weigh (DW): In this method, we learn the weight of edge *(u,v)* by using the degree of the target node *v* as in Eq. (1):

$$w_{uv} = \frac{1}{d\_in_v} \tag{1}$$

The weight of edge *(u,v)* is shown by w_{uv} and d_in_v is the degree of node *v*.

Tweet Weight (TW): The weight of edge *(u,v)* is computed as the number of retweets that node *v* has from node *u*, divided by all tweets and the retweets user *u* has

1. www.isi.edu/~lerman/downloads/digg2009.html
2. http://www.cs.sfu.ca/~sja25/personal/datasets/
3. https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Table 1. Comparison of basic and recent methods with proposed method.

Method	Task	Input Data	Evaluation Metrics	Advantages	Key Limitation / Gap Addressed by Our Work
Independent Cascade Model	activation prediction	Network only (graph)	F1	1) one of the first models, modeling the information and influence propagation in networks; 2) Just needs the graph and probability on the edges	1) Ignores content semantics; 2) Ignores temporal features; 3) Prediction is far from the behaviour of tweets with low retweet count;
Linear Threshold model	activation prediction	Network only (graph)	F1	1) one of the first models, modeling the information and influence propagation in networks; 2) Just needs the graph and weights on the edges	1) Ignores content semantics; 2) Ignores temporal features; 3) Prediction is far from the behaviour of tweets with high retweet count;
Barbieri et al. (Topic-aware)	activation prediction	Network + **Topic**	F1, AUC	1) Using topic information; 2) using temporal features; 3) More accurate in predicting the "time" of activation	1) No improvement in activation prediction compared to IC and LT models; 2) A lot of parameters because of different topics and a higher risk of overfitting; 3) Needs access to both topics and the graph of the network
Our Proposed Method (Temporal Content-Aware Generalized Propagation Model)	activation prediction	Network + **Text**	F1	**1) using content semantics; 2) using temporal features; 3) successful in activation prediction for all tweet groups with high, low, and medium retweet counts; 4) High improvement in the F1-measure of activation prediction**	1) needs access to both the text and the graph of the network

published. Eq. (2) shows this method in which RT_{uv} shows the number of retweets *v* has from *u,* and T_u shows all the tweets and retweets that *u* has published:

$$w_{uv} = \frac{RT_{uv}}{T_u} \quad (2)$$

This method of learning edge weights dramatically suffers from data sparsity, and many edges receive a weight of zero. For example, in our dataset, there are 973,474 retweets, and therefore, at least 7,041,645 edges receive a weight of zero, which are 88% of the edges. In these situations, for most of the test tweets, no new node could become active, and precision and recall became zero. So we exclude this method and instead use the next method described, **TW-S**, which uses smoothing.

Tweet Weight with Smoothing (TW-S) is similar to **TW** described before, except it uses Laplace smoothing according to Eq. (3):

Table 2. Twitter dataset specification.

Number of tweets	6,756,887
Number of users	2,029,108
Number of users whose tweets were fully crawled	9,442
Number of edges of the retweet network	8,015,119

$$w_{uv} = \frac{RT_{uv}+1}{T_u+d\_in_v} \quad (3)$$

where the constant 1 in the numerator is a fixed additive smoothing term.

3- 3- Computing Content Scores

For weighting nodes, we use the basic scoring method of Apache Lucene open-source search software, which is based on TF-IDF. First, we index documents corresponding to users. Then, for each test tweet, all documents are scored by considering the test tweet as a query.

3- 4- Methods to compare

LT: In Fig. 1, LT-DW is the LT propagation model using the DW method to compute edge weights, and LT-TW-S is the LT propagation model using the TW-S method to compute edge weights via retweet count and smoothing as described before.

IC: In Fig. 1, IC-DW is the IC propagation model using the DW method to compute edge weights, and IC-TW-S is the IC propagation model using the TW-S method to compute edge weights via retweet count and smoothing as previously described.

TLT and TIC: As described in the previous section, the TLT and TIC propagation models are reported to be the same as the LT and IC models in predicting the final activation of nodes, and so the curves for the LT and IC models represent the results of the TLT and TIC models, and we do not repeat them.

RW: In Fig. 1, RW-α-DW is the Random Walk method with jump parameter α using the DW method to compute edge weights.

3- 5- Results of classic propagation models

Fig. 1 shows results for the classic models, LT, IC, and RW, with different methods for learning the weight of edges. The horizontal axis shows Retweet Count on a logarithmic scale, and the vertical axis shows F-measure on a logarithmic scale. Tweets are divided into bins of length 10, based on retweet count, in order to reduce the distortions of the diagrams. In other words, the average F-measure for test tweets having retweet counts of 1 to 10 in the gold data is used as the first point for forming the curve, and similarly for the next points.

We observe, according to Fig. 1, that the F-measure for the classic propagation model LT with different learning methods for edge weight decreases, whereas the retweet count in the real world increases. In other words, the classic LT model has very good prediction for tweets with low retweet potential, but is not successful in predicting the behaviour of tweets with higher retweet potential. This is because in the LT model, each node usually needs some active neighbours to become active. So propagation is hard and stops early. This kind of propagation is the behaviour of tweets with low retweets, and therefore, the LT model is very good for the prediction of tweets with low retweet count. However, it is far from the behaviour of tweets with broad propagation, and so the LT model fails in predicting the propagation of this kind of tweet.

The decrease in LT's F-measure with increasing retweets can be explained by the weighted nature of the threshold mechanism. In LT, activation depends on the cumulative influence weight of active neighbors exceeding a node's threshold. As the number of retweets grows, many of these retweets originate from peripheral or low-influence users (e.g., casual followers rather than hubs). While they increase the raw count of exposures, they contribute only marginal weight, making it harder for the cumulative sum to cross the threshold—especially if thresholds are calibrated on high-influence users.

Conversely, in the IC model, each retweet acts as an independent propagation attempt with a fixed transmission probability p. Even if retweets come from low-influence users, they still provide additional independent chances to activate neighbors. Therefore, more retweets monotonically increase the probability of at least one successful transmission, leading to an upward trend.

In Fig. 1, it is observed that the IC model and the RW model have very similar behaviours, and changes in methods of learning the edge weights and the jump parameter in the RW model do not make any noticeable difference. This similarity in behaviour was not far from expectation because, as we mentioned before, Xiang et al. stated that the RW model could be an approximation of the IC model.

We could conclude from Fig. 1 that propagation in the IC model (without a jump parameter) reaches a wide area of the

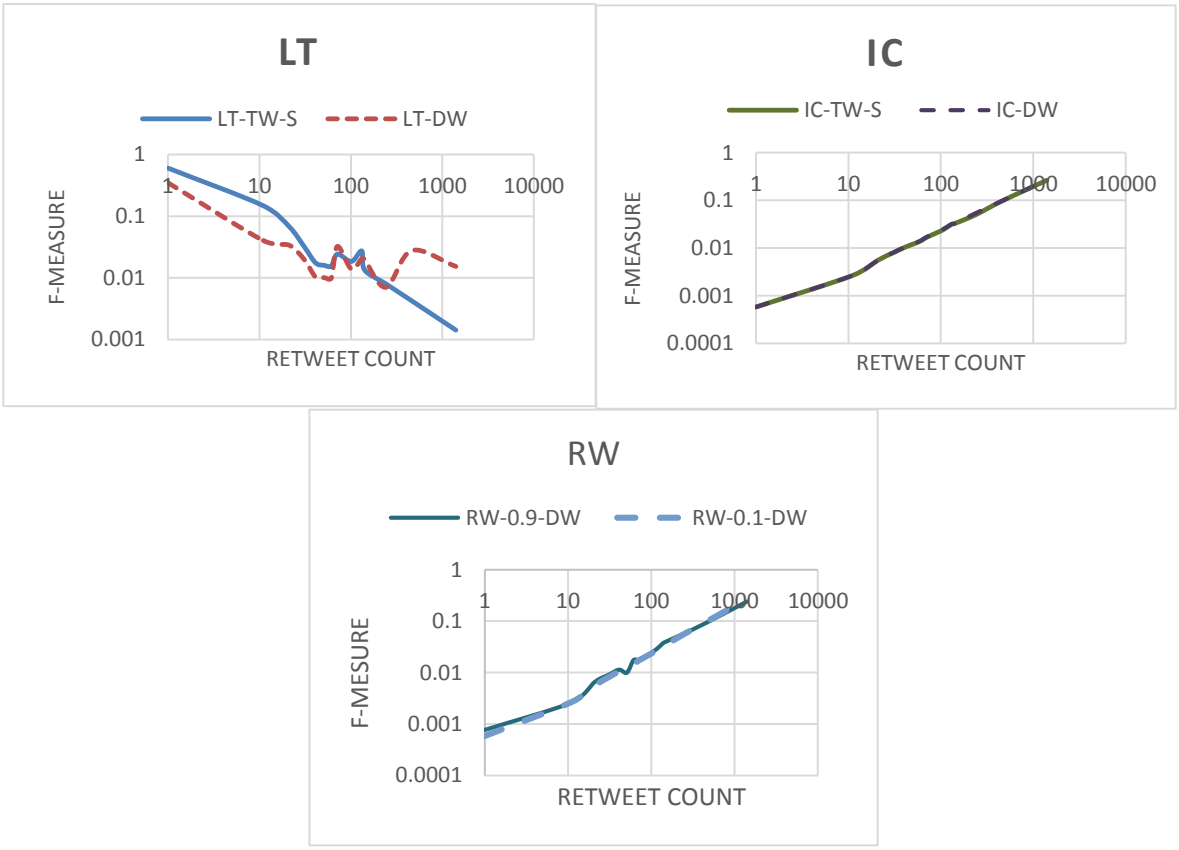

Fig. 1. F-measure of propagation prediction using basic propagation models, LT, IC, RW.

network, and adding jump parameters in the RW model does not remarkably change the size of the area. That is why the IC and the RW models have good propagation for tweets with high retweet count, and changing the jump parameter in the RW model does not cause significant changes in behaviour.

4- Content-Aware Generalized Propagation Model

In this section, we describe the Content-aware approach for predicting diffusion behaviour in social networks. The content-aware diffusion model is composed of three main phases: First, we compute a content-similarity score between each user (node) of the network and the message whose diffusion we want to predict. Then, based on the computed score, we predict how big the message diffusion would be in the network. This prediction classifies the messages into two classes. The third and final phase is selecting a suitable diffusion model based on the popularity of the message.

These three phases are described in more detail:

4- 1- Content Score Computation

A measure called "content-similarity score" is assigned to each user, u, of the network and every message, m, whose diffusion model we want to predict. Tweets of each user u in the network are concatenated together and considered as a document d_u. These are all available tweets of the user, published until now, and they are going to help us predict the number of users who would potentially retweet the query tweet. The message m is referred to as the query tweet hereafter. The idea is that content similarity between the query tweet and the user's tweets means the user is interested in the query tweet.

For the computation of similarity scores, we used the Lucene search engine. When the query tweet m is given to the search engine, all documents (here tweets of users) are searched efficiently, and similar documents, sorted according to a ranking method appear in the output. This score, shown as $S(m,u)$, is the basis of our judgement. The content similarity score between a query message m in the social network, SN, is defined as the sum of the similarity scores of all users u with query message m:

$$CS(m, SN) = \sum_{u \in SN} S(m, u) \quad (4)$$

The content score computed in this phase is used in the popularity prediction phase.

4- 2- Popularity Prediction

The second phase is in fact a classifier. This step uses the

content scores computed in the previous phase and specifies which propagation model should be used in the next step. In the simplest form, it is a binary classifier. It computes the sum of the content scores of all nodes, and if this sum is lower than a threshold, outputs are low spread; and if it is higher than the threshold, outputs are high spread. Eq. (5) shows how we choose between the two propagation models.

$$I_{m,SN} = \begin{cases} \text{IC} & CS(m,SN) \geq \tau \\ \text{LT} & CS(m,SN) < \tau \end{cases} \quad (5)$$

$I_{m,SN}$ shows, for message *m* in social network *SN,* which propagation model predicts the propagation behaviour of the message. *CS(m,SN)* is defined in Eq. (4) and τ is a tuned threshold. Threshold selection is described in the next subsection. The IC and the LT model in Eq. (5) could also be replaced with other suitable propagation models. The selection of a model is based on experiments that demonstrate which model describes which kind of propagation. Based on our experimental results, the IC model describes broad propagation and the LT model describes limited propagation.

4- 3- Model Selection

The third and final phase is using the suitable propagation model for each class. As mentioned before, based on our experiments, the LT model is suitable for the low spread class, and the IC or RW models are suitable for the high spread class.

Phases B and C could be formulated by General LT propagation model as Eqs. (6-7):

$$f(CS(m,SN) = \begin{cases} 1 & CS(m,SN) \geq \tau \\ 0 & CS(m,SN) < \tau \end{cases} \quad (6)$$

$$Inf_v^t(m,SN) = f\big(CS(m,SN)\big) \times (1 - \prod_{u \in in_v(t)}(1 - w(u,v)) + (1 - f(CS(m,SN)) \sum_{u \in in_v(t)} w(u,v) \quad (7)$$

CS(m, SN) is the Content Score of message *m* in social network *SN* as defined in (4). $f(x): R^+ \rightarrow \{0,1\}$. Therefore if *f(CS(m,SN))=1* it is equal to IC propagation model and if *f(CS(m,SN))=0* it is equal to LT propagation model[1]. $Inf_v^t(m,SN)$ is the influence node *v* receives at time *t* for message *m* in social network *SN*. $in_v(t)$ is the set of all active in-neighbours of node *v* at time *t*.

1 We examined other values in (0,1) for *f(CS(m,SN))* and the prediction received lower f-measures.

4- 4- Threshold Selection

Before analysing the behaviour of the Content-Aware propagation model, we should describe how we determine the threshold of decision between the two propagation models.

Train data is also divided into two parts. The first and main part is used to form user content. The second part is used to determine the threshold in popularity selection, and we show this set of tweets by *MT*[2] . Then the *ICM* is defined as the set of tweets in *MT* that gain a higher F-measure in predicting their behaviour by the IC propagation model compared to the LT propagation model:

$$ICM = \{m \in MT \text{ such that } F\text{-}measure(m,IC) > F\text{-}measure(m,LT)\} \quad (8)$$

The threshold τ is computed as the 5th percentile of the CS(m, SN) values over the ICM set, rather than the minimum, to mitigate the influence of outliers:

$$\tau(m,SN) = \underset{m \in ICM}{Q_{0.05}}\,(CS(m,SN)) \quad (9)$$

where $Q_{0.05}$ is the 0.05 quantile. This choice preserves weak signals in the lower tail while excluding extreme anomalous values. We suggest this procedure for selecting the threshold because our goal is to propose a propagation model that describes the behaviour of tweets with high retweet count in the first place, and in the second place, could give a good prediction for other tweets with low retweet count. Notice that the number of tweets with high retweet count is much lower than the number of tweets with low retweet count, but these tweets are usually the most important ones in the social networks.

5- Temporal Content-Aware Generalized Propagation Model

In the previous section, we proposed a content-aware propagation model based on the popularity prediction of the message. The model used a binary classifier selecting between two propagation models, each of which was appropriate either for limited or broad propagation. In this section, we expand this idea and notice that limited or broad propagation are not absolute concepts, but relative ones.

In this section, we propose a propagation model that considers the lifetime of the message. Every message has broad propagation in its early lifetime and limited propagation in its late lifetime. And how it means early or late in its lifetime is determined by its content and popularity detection. Popular messages have longer lifetimes with broad propagation and finally fall into a limited propagation state. Less popular messages have shorter lifetimes with broad propagation and

2 Standing for Mid-Test

soon fall into a limited propagation state. Eq. (10) states the influence of node v for message m at time t in social network SN. In Eq. (10), $f\left(t,CS\left(m,SN\right)\right)$ is a function of time t and the content score of message m in social network SN.

$$Inf_v(t,m,SN) = f\big(t,CS(m,SN)\big) \times \left(1 - \prod_{u \in in_v(t)} (1 - w(u,v)\right) \quad (10)$$

$$+(1 - f(t,CS(m,SN)) \sum_{u \in in_v(t)} w(u,v)$$

Eq. (11) shows a proposed function for $f\left(t,CS\left(m,SN\right)\right)$:

$$f(t,CS(m,SN)) = \begin{cases} 1 & t \le log\ CS(m,SN) \\ 0 & t > log\ CS(m,SN) \end{cases} \quad (11)$$

The next section describes experimental results.

6- Results

In this section, experimental results are described. Evaluation criteria and the configuration of methods for comparison, such as learning edge weights and computing content score, are the same as the ones described in Section III.

The proposed propagation model, the content-aware propagation model with a threshold τ is shown by **CA-τ**. The temporal content-aware propagation model is shown by **TCA**.

Training data is also divided into two parts as described in Threshold Selection. The first part consists of 80% of all of the tweets and is used to form user content. The second part consists of 10% of all of the tweets and is used to determine the threshold in popularity selection. The threshold for our dataset was selected as 300 by the described procedure. Table 3 shows the distribution of content scores of these tweets.

Fig. 2 shows the results for the CA and TCA propagation models and compares them with the LT and IC models, and Table 4 shows the exact F-measure of these methods for each retweet count bin. In Fig. 2, for simplicity of comparison, we include only one curve from each approach: LT, IC, and CA. For the LT model, configurations with more extreme behaviour were selected in order to be able to see the difference between the better approaches. LT curves have decreasing total behaviour (high F-measure at the beginning of the curve and low F-measure at the end), so the **LT-TW-S** configuration is selected. IC curves are vice versa and increase despite the two curves of different configurations for the IC propagation models being almost the same, and therefore, there is not much difference in the selection of each of them.

Table 3. Distribution of content scores

Content Score Range	Count
[0,100]	84842
(100,200]	16710
(200,300]	520
(300,400]	310
(400,500]	53
(500,600]	24
(600,700]	13
(700,1000]	4
(1000,2000]	4
(2000,9000]	5

The weakness of the three models - the LT, IC, and CA propagation models - is related to tweets with middling retweet counts. Neither the LT nor the IC models give good predictions of the behaviour of these tweets, and therefore, the CA model that selects between these two models also has this weakness. However, as we can see in Fig. 2 and Table 4, the CA propagation model can overcome the weaknesses of the two classic propagation models, LT and IC, for tweets with high or low retweet count, and has the advantages of both of them. The CA propagation model gives a good prediction of behaviour for tweets with high retweet count, like the IC model, and a good prediction of behaviour for tweets with low retweet count, like the LT model. This recognition of the behaviour of tweets and selection of the appropriate model has been done by the previously described content analysis.

As observed in Fig. 2, the TCA propagation model overcomes the weakness of the LT, IC, and CA models and enhances the prediction of propagation behaviour in all points of the diagram. This is because we switch from a binary classifier to a scheme considering a spectrum of behaviour from limited propagation to broad propagation.

7- Conclusion

In this paper, we have proposed utilizing contextual information of social networks to improve the modelling of propagation. First, we introduced classical propagation models, including LT, IC, and their extensions.

We analysed the Twitter data and observed that the suitability of classic propagation models like LT, IC, and RW depends on the retweet count. The LT model is suitable for

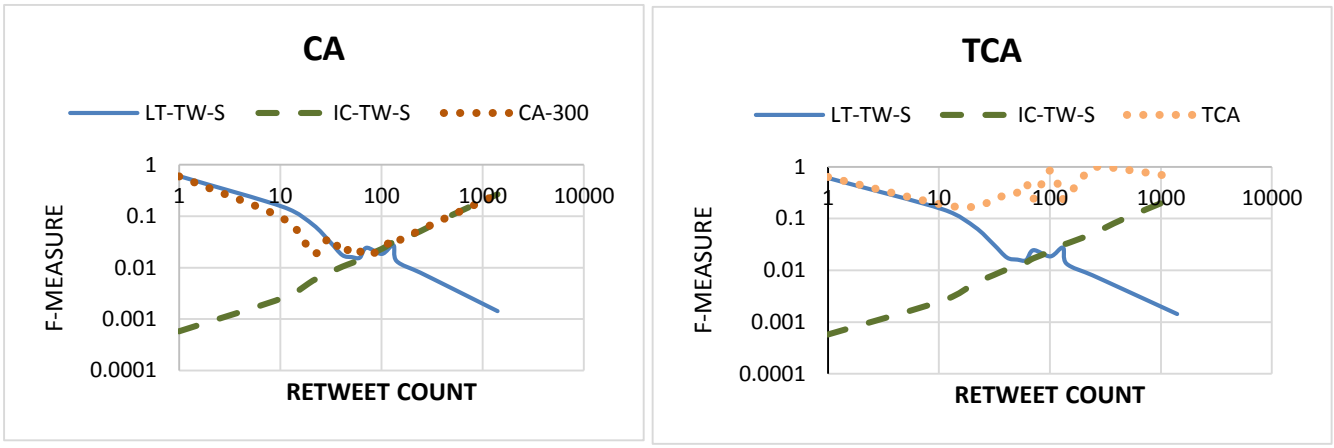

Fig. 2. F-measure of CA propagation prediction models in comparison to basic models.

Table 4. F-measure of CA propagation prediction models in comparison to basic models.

Gold Retweet Count Bin	LT-TW-S	IC-TW-S	CA	**TCA**
1	0.60670842	0.0005769	0.5984354	**0.636779234**
11	0.14862511	0.0026357	0.0871063	**0.184148827**
21	0.06875451	0.0055793	0.0188807	**0.170655246**
31	0.03176438	0.0080735	0.0387426	**0.229143878**
41	0.01756854	0.0104149	0.0188433	**0.295999104**
51	0.01602543	0.0122744	0.0261719	**0.319460226**
61	0.015625	0.014236	0.020915	**0.485010843**
71	0.02439024	0.017059	0.017059	**0.241758242**
101	0.01851852	0.023124	0.023124	**0.847826087**
131	0.02739726	0.0316134	0.0368098	**0.23877095**
141	0.01342282	0.0327203	0.0327203	**0.254545455**
241	0.00801684	0.0506979	0.0534421	**0.921051007**
461	0.00425532	0.098225	0.098225	**0.887039239**
1401	0.00142349	0.2658891	0.2658891	**0.623835213**

predicting the behaviour of tweets with low retweet counts, but fails in predicting the behaviour of tweets with broad propagation. Vice versa, the IC and RW models are suitable for predicting the behaviour of tweets with high retweet counts, but fail in predicting the propagation of messages with limited propagation. Based on the observations, we proposed selecting a propagation model by popularity prediction of the tweet in the network.

Content analysis is done for popularity prediction of the messages, determining how much they get retweeted and propagated. We described using information retrieval methods to predict the resharing of a message by a user

on social networks. All previous messages of a user are considered to be a document. We create an index from all documents corresponding to users. Then each test message is considered a query. All documents are given a score, which shows how much a document is related to the query. The sum of the scores of all documents is a measure to predict how much the message would be propagated. This phase is the popularity prediction.

We proposed two new propagation models based on popularity prediction using contextual information. The CA method, which uses a binary classifier in the popularity prediction phase, and the TCA method, which is based on a context score (popularity score, in other words), decide how long a message has a broad propagation behaviour and after how many time steps its behaviour changes to limited propagation.

Experiments done on data collected from Twitter show the advantages of the CA and TCA propagation models over the previous ones – LT, IC, and RW. Our proposed model, the CA model, overcomes the weaknesses of the LT, IC, and RW models and has good propagation prediction for tweets with high or low retweet counts. This is done by popularity prediction using contextual information, selecting the suitable model before giving a propagation prediction. In fact, the CA model is a combination of the LT and the IC model that determines popularity prediction by using content analysis.

But the CA model could not do more than either the LT or IC model does. The TCA propagation model is a noticeable improvement that, at every point, predicts better than all of the previous models. That is because the TCA model does not stick to the LT or IC model alone, but considers that every message changes its behaviour in its lifetime. Every message has a broad propagation behaviour in its early lifetime, and the more popular the message, the longer this period is, and then it changes its behaviour to limited propagation. This strategy leads to a more accurate propagation prediction.

Despite its strong performance, our proposed TCA model has several limitations that should be acknowledged. First, the method assumes the availability of textual content for all messages in the diffusion cascade. In platforms where text is sparse (e.g., image-centric social networks) or inaccessible due to privacy restrictions, the content similarity component would be ineffective, and the model would degrade to a purely structure-based approach. Second, our current implementation relies on TF-IDF trained primarily on English text; performance may not generalize to multilingual or code-mixed corpora without additional adaptation or multilingual embeddings. Third, the model exhibits sensitivity to network sparsity—in graphs with few observed edges or in cold-start scenarios where new users or topics have no historical data, the diffusion predictions become less reliable. Fourth, in Equation (11), we adopted a logarithmic transformation for $f(t,CS)$ model the saturating effect of content similarity on diffusion. However, we acknowledge that this choice was made without an exhaustive empirical comparison against alternative transformations such as square root, power, or linear functions. While the logarithm is theoretically grounded in the diminishing returns behavior observed in information retrieval, its optimality across datasets with different characteristics remains unverified. Finally, the computational overhead of diffusion simulation may limit real-time applicability in very large-scale settings, though this could be mitigated through approximation techniques. Addressing these limitations constitutes the primary direction of our future work.

Future work will focus on three primary directions. First, we plan to move beyond TF-IDF by exploring deep contextualized embeddings (e.g., BERT or Sentence-BERT) to capture more nuanced semantic similarity between content items. Second, we intend to extend our diffusion model to multilingual networks, investigating how information spreads across communities that operate in different languages, which remains an open challenge in cross-cultural social media analysis. Third, we will conduct a systematic comparative analysis of alternative functional forms for $f(t,CS)$ including square root, power, and linear transformations. This will not only identify the most effective formulation across diverse datasets but also provide deeper insights into the functional relationship between content similarity and diffusion dynamics. We also plan to explore adaptive transformations that learn the optimal functional form directly from the data.

References

[1] Anh, Nguyen Viet, Duong Ngoc Son, Nguyen Thi, Thu Ha, S Kuznetsov, Nguyen Tran, and Quoc Vinh. 2018. "A Method for Determining Information Diffusion Cascades." Eastern-European Journal of Enterprise Technology 6 (2): 61–69. doi:10.15587/1729-4061.2018.150295. http://journals.uran.ua/eejet/article/view/150295.

[2] Barbieri, Nicola, Francesco Bonchi, and Giuseppe Manco. 2012. "Topic-Aware Social Influence Propagation Models." In IEEE International Conference on Data Mining, 81–90. Ieee. doi:10.1109/ICDM.2012.122. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6413913.

[3] Challenge, Network. 2010. "Social Contagion : An Empirical Study of Information Spread on Digg and Twitter Follower Graphs." ACM Transactions on Embedded Computing Systems V (Section 3): 1–15. https://arxiv.org/abs/1202.3162.

[4] Chen, Wei, Laks V.S. Lakshmanan, and Carlos Castillo. 2013. Information and Influence Propagation in Social Networks. Synthesis Lectures on Data Management. Morgan & Claypool. doi:10.2200/S00527ED1V01Y201308DTM037. http://www.morganclaypool.com/doi/abs/10.2200/S00527ED1V01Y201308DTM037.

[5] Goyal, Amit, Wei Lu, and Laks V.S. Lakshmanan. 2011. "SIMPATH: An Efficient Algorithm for Influence Maximization Under the Linear Threshold Model." In IEEE International Conference on Data Mining, 211–220. IEEE. doi:10.1109/ICDM.2011.132.

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6137225.

[6] Kempe, David, Jon Kleinberg, and É. Tardos. 2003. "Maximizing the Spread of Influence through a Social Network." In KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 137–146. doi:https://doi.org/10.1145/956750.956769. http://dl.acm.org/citation.cfm?id=956769.

[7] Kempe, David, Jon Kleinberg, and É. Tardos. 2005. "Influential Nodes in a Diffusion Model for Social Networks." In: Caires L., Italiano G.F., Monteiro L., Palamidessi C., Yung M. (eds) Automata, Languages and Programming. ICALP 2005. Lecture Notes in Computer Science, Vol 3580. Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/11523468_91. http://link.springer.com/chapter/10.1007/11523468_91.

[8] Kupavskii, Andrey, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. 2012. "Prediction of Retweet Cascade Size over Time." In CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2335–2338. doi:https://doi.org/10.1145/2396761.2398634. https://dl.acm.org/doi/10.1145/2396761.2398634.

[9] Lin, CX, Qiaozhu Mei, and Jiawei Han. 2011. "The Joint Inference of Topic Diffusion and Evolution in Social Communities." In IEEE International Conference on Data Mining. doi:10.1109/ICDM.2011.144. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6137242.

[10] Macskassy, SA, and Matthew Michelson. 2011. "Why Do People Retweet? Anti-Homophily Wins the Day!" In The International Conference on Weblogs and Social Media, 209–216. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2790/3291.

[11] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2013. An Introduction to Information Retrieval. . . . Information Retrieval. Online ed. Cambridge, England: Cambridge University Press. http://link.springer.com/chapter/10.1007/978-3-642-39314-3_1.

[12] Petrovik, Sasa, Miles Osborne, and Victor Lavrenko. 2010. "RT to Win ! Predicting Message Propagation in Twitter." In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, RT, 586–589. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2754/3209.

[13] Pramanik, Soumajit, and Jean-Loup Guillaume. 2017. "Modeling Cascade Formation in Twitter Amidst Mentions and Retweets." Social Network Analysis and Mining 7 (41). doi:10.1007/s13278-017-0462-1. "http://dx.doi.org/10.1007/s13278-017-0462-1.

[14] Xiang, Biao, Qi Liu, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. 2013. "PageRank with Priors : An Influence Propagation Perspective." In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI), 2740–2746. https://dl.acm.org/doi/10.5555/2540128.2540523.

[15] Suh B, Hong L, Pirolli P, Chi EH .2010 "Want to be retweeted? Large-scale analytics on factors impacting retweets in Twitter network. In: 2010 IEEE second international conference on social computing (socialcom). IEEE, pp. 177–184

[16] Malhotra A, Malhotra CK and See A .2012 "How to get your messages retweeted". MIT Sloan Manage Rev 53(2):61–66

[17] Yan, F., Liang, W. & Hirota, K. An information propagation model for social networks based on continuous-time quantum walk. Neural Comput & Applic 34, 13455–13468 (2022). https://doi.org/10.1007/s00521-022-07168-7

[18] Hong, Z., Zhou, H., Wang, Z., et al. (2023). Coupled Propagation Dynamics of Information and Infectious Disease on Two-Layer Complex Networks with Simplices. Mathematics, 11(24), 4904. DOI: 10.3390/math11244904

[19] Li, Y., & Han, Z. (2024). The propagation dynamics of UAU-SEPIR in two-layered networks. Advances in Continuous and Discrete Models. DOI: 10.1186/s13662-024-03858-9

HOW TO CITE THIS ARTICLE
F. Ghayour-Baghbani, Generalized Content-Aware Diffusion Models for Twitter, AUT J. Model. Simul., 58(1) (2026) 125-136.

DOI: 10.22060/miscj.2026.24962.5444