International Journal of Knowledge and Systems Science
Volume 10 • Issue 1 • January-March 2019

Reconstructing Diffusion Model for Virality Detection in News Spread Networks

Kritika Jain, Jaypee Institute of Information Technology, Noida, India
Ankit Garg, Jaypee Institute of Information Technology, Noida, India
Somya Jain, Jaypee Institute of Information Technology, Noida, India

ABSTRACT

In today's competitive world, organizations take advantage of widely-available data to promote their products and increase their revenue. This is achieved by identifying the reader's preference for news genre and patterns in news spread network. Spreading news over the internet seems to be a continuous process which eventually triggers the evolution of temporal networks. This temporal network comprises of nodes and edges, where node corresponds to published articles and similar articles are connected via edges. The main focus of this article is to reconstruct a susceptible-infected (SI) diffusion model to discover the spreading pattern of news articles for virality detection. For experimental analysis, a dataset of news articles from four domains (business, technology, entertainment, and health) is considered and the articles' rate of diffusion is inferred and compared. This will help to build a recommendation system, i.e. recommending a particular domain for advertisement and marketing. Hence, it will assist to build strategies for effective product endorsement for sustainable profitability.

KEYWORDS

Diffusion, Natural language Processing, News Spread Networks, Recommendation System SI model, Social Networks, Virality Detection

DOI: 10.4018/IJKSS.2019010102

1. INTRODUCTION

The Internet has revolutionized the way in which people interact and communicate with each other. Communication and access to information are easier than it ever has been. This has resulted in the creation of a master novel platform, known as the "social network" which has made a magnificent impact on the life of each and every individual in one way or another in the last half-decade. Social networks have revolutionized the entire communication process and how the information used to spread from one person to another. Earlier, newspapers and news reports on television or radio were the only source to know what is happening in the world and stay updated. Since centuries, people relied on newspapers to get the most recent world news. But in the current scenario, newsfeeds on social media pages of Facebook, Twitter, etc. keep us up to date on all the things happening all around the globe. It doesn't matter where we are, we can have access to any information we want. One of the most interesting things is that social networks provide users with the power to instantly record and share an event or write articles directly from their phones instantly simultaneously as they happen. In many cases like news about natural disasters, bombing, shooting, product launch etc. reach social media before journalists could even reach the scene or spread the news through their channels. The thing that happens on social media is that once an information article or news is posted on these social networks, the news takes the life of its own. It is shared by the readers across different social media platforms in different ways and it can reach a growing exponential number of people fast. Information flow in the network occurs by replication of articles (sometimes data is modified by users) by different people regularly over a period of time.

Therefore, the study of spread dynamics of information through the internet is an important and interesting task, as it could be valuable to comprehend the factors which determine the journey of information in a social network. A variety of factors control the spread of the news articles like the reputation of the publisher, domain of the news (business, technology, health, sports, fashion etc.), the timestamp of an article and many more. Such extensive information diffusion provides an enormous opportunity to companies and enterprises to exploit the benefits of social media networks to boost their revenues and amplify their presence in the market. It helps businesses to understand their target audience, increase customer engagement and responsiveness and make better business decisions by building a strong strategy for a product launch, product endorsement, customer services etc.

Hence, the aim of this research is to reconstruct SI (Susceptible-Infected) diffusion model to detect virality of news in dynamic news spread network. The overall approach is to analyze the virality of news of four different domains by comparing their respective infectious diffusion rates for product endorsement and extent of virality of news articles. The rest of the paper is structured as follows: section II will provide related work and identified gaps, section III discusses the overall proposed methodology and framework, section IV shows the findings and results and section V will provide the conclusion and the future scope of research.

2. RELATED WORK IN THE AREA

In this section light is thrown on earlier work done related to it with the research gaps identified from the existing literature. Diffusion models, virality of tweets and news, news spread networks and prediction by analyzing rate of information diffusion has been investigated in various studies. The conception of structural virality is studied to quantify the structure of information cascades in (Goel et al., 2015). It was proposed by lack of replication of the observed diversity of structural virality. Usage of K-core and Page Rank algorithm for finding out important seeds in social networks is presented in (Sarkar et al., 2017). Comparative analysis against true seeds from the two-phase algorithm is also given. Dynamic patterns of online news are analyzed by the authors and then comparison against the spreading patterns of the epidemic model in (Wang et al., 2009). A method for prediction of the interest gain by news articles is proposed in (Tatar et al., 2014). The work is followed by ranking data based on the prediction using a linear log popularity prediction model. Examination of the effects over controversial news story and analyzing the supporters' and opponents' reaction is done in (Fang & Ben-Miled, 2017). Reaction over news to check if bad news spread faster than good news was also demonstrated. Some authors focus on the spread of news across social networks and understand its dynamics to see how the news spread follows SIR (Susceptible-Infected-Recovery) epidemic model (Mussumeci & Coelho, 2018). Analysis of the information spreading in Twitter and temporary dynamics of information spreading in Twitter is presented in (Zaman et al., 2010).

In one of the studies, researchers concentrate on developing a model for the information diffusion (Stai et al., 2018). Information diffusion during Natural Disasters in social networks plays a crucial role to minimize the catastrophic effect of the disaster. An approach to manage with earthquake relief and post-earthquake reconstruction using Weibo Information Flow (WIF) model is presented in (Dong et al., 2018). Demonstration of how influence distance approach and MLE approach can be helpful for detection of infection in social networks is provided in (Li et al., 2016). Authors use k-boosting problem to calculate the boosted information spread in (Lin et al., 2017). This method works by finding the k-users to boost so that maximized "boosted" influence spread is started using PRR –Boost. Categorization of sentiments in Twitter for analyzing the health concerns, which helps in solving the problem of spreading public concern about epidemics using two-step sentiment classification approach for Twitter messages, is provided in (Ji et al., 2015). One of the studies focuses on a sarcastic way to detect the fake news (Rubin et al., 2016). Some authors propose the way of assigning a relative influence score and for each user it also focuses on passivity score to identify the users who are very good at spreading information and vice-versa (Romero et al., 2011). In a given social network, the problem of increasing the influence by finding out 'k' nodes (k is a constant) that influence the maximum nodes in a pre-defined network is presented in (Tang et al., 2014). This problem is extensively studied as it has important applications in viral marketing and business developments. Approach for age estimation using first names of people was also done in (Oktay et al., 2014).

Evaluation of the characteristics and relative influence of the Twitter users is done by following the spreading of information that occurred on Twitter in a two-month interval in a given year (Bakshy et a., 2011). The dynamics of Twitter is an interesting study carried out by authors. A model is proposed that measures and analyzes the incidences of the bursts as a function of the information spreading through the network (Myers & Leskover, 2014). Prediction of re-tweets cascade size over time to approximate the logarithmic size of the cascade at any time 'T', and minimization of mean square error of the training data is carried out to find out if a user re-tweets a tweet is the focus of (Kupayskii et al., 2012). Using distant supervision technique, Twitter sentiment classification is performed by using algorithms like Naïve Bayes, Maximum Entropy, and SVM (Go et al., 2009). Training with the emoticon data is done to find the sentiment of the data. SEIZ-enhanced epidemic model is used to characterize eight catastrophic events across the world and a range of event types (Jin et al., 2013). D-Sieve is a tool developed for well-organized management of messages, specifically catastrophic and crisis-related, which is proposed in (Chowdhury et al., 2015). It concentrates on the post-classification processing step that uses two features- stable hashtag association and stable named entity association-to increase the classification precision for a classifier. Prediction analysis of whether a tweet will become viral by finding out what makes Twitter users re-tweet a tweet is done in (Jenders et al., 2013). A method is proposed for influence maximization in social networks. It works by finding the k nodes in an active set B with the maximal influence in a targeted set P (Zhou & Guo, 2014).

To make useful predictions, models and algorithms are proposed to learn the model parameters and test the learned models (Goyal et al., 2010). To boost a particular information spread, a new method is proposed that predicts new social links that can be inserted among existing users of a social network and boost up the spread network (Antaris et al., 2014). It can be extremely valuable to increase the reach of the network. Study of the process of diffusion of information and the impact of external influences in networks is done in (Myers et al., 2012). For expected size maximization of the resulting cascade, a solution is provided in where a solution for the problem of finding a set of k initial seed nodes in a network is provided (Borgs et al., 2014).

One of the identified research gap in existing research is the usage of offline data set instead of dynamic and real time dataset. Another issue which is identified is that for analyzing the information diffusion of news articles only crisis situation is considered. However, no amount of consideration is given to different domains of news like business, entertainment, technology and health etc. Research related to boost news spread networks and takes the advantage of diffusion and infection rate is still limited. Also, less importance is given to preprocessing techniques and cleaning of data prior applying the algorithm which could bring significant results. Therefore, this work will consider dataset comprised of four news articles domains and analyze the results using natural language processing and some preprocessing techniques.

3. PROPOSED METHODOLOGY AND FRAMEWORK

In this section the proposed research framework (figure 1) along with the dataset characteristics is described.

3.1. Dataset

News Aggregator dataset is considered for experimental analysis. Headlines and categories for 400k news items are scraped from the web in 2014 (Gasparetti, 2017). Columns extracted from it are *'TITLE'* which represents the headline of the article; *'PUBLISHER* 'the publisher of the article; *'CATEGORY'* which represents the categories of the news items: '*b'* for business articles, '*t'* for science and technology articles, '*e'* for entertainment articles, '*m'* for medical and health articles; *'TIMESTAMP'* represents the approximate timestamp of the article's publication (seconds are taken since midnight on Jan 1, 1970). Table 1 describes the dataset characteristics. Dataset consists of news articles which are represented using nodes hence forming a news spread network of 422419 nodes and 1056047 edges. The preprocessing and training of dataset is done using NLP techniques.

3.2. Work Done

After the successful extraction of dataset features, it is preprocessed so as to clean the data for further analysis. Preprocessing of data is done as follows: First, dataset is divided into four domains i.e. business, medical, entertainment and technology on the basis of category column in the news aggregator dataset. Subsequently, for each domain the extracted columns (title, publisher and timestamp) are trained by using several functions like .lower() which convert all the capital letters into small letters and .Porter Stemmer algorithm is applied which will remove all the stop-words like is, am, are and only keep the meaningful words. Finally, top 1500 words for each domain that are used maximum number of times to represent every article in the form of vector are selected. This is followed by similarity calculation between articles using TF-IDF and cosine similarity. Cosine similarity between articles is calculated using equation (1).

$$similarity\left(v1,v2\right) = \cos\theta = \frac{v1.v2}{v1v2} \quad (1)$$

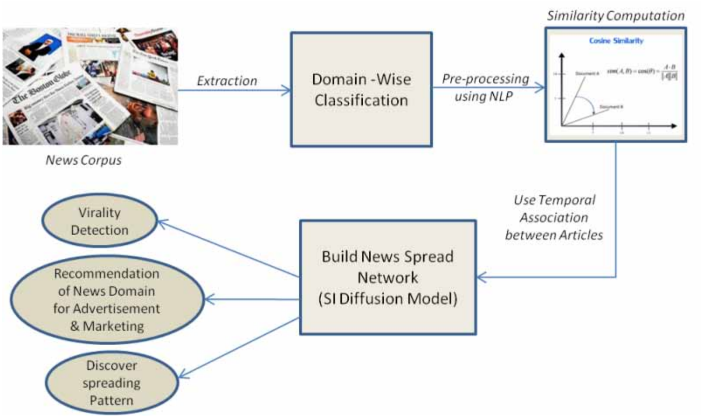
In above equation (1), $v1$ is vector representation of first article and $v2$ is vector representation of second article. Angle between the article vectors is denoted by θ. In figure 2 distribution of similarity among articles for each domain is represented. The x-axis represents the similarity score and the y-axis represents the number of news article pairs. In the graph each bar shows the number of article pairs having the similarity score between the current similarity score and the previous bar similarity score.

Once the similarities of the articles are calculated, most likely infector of an article is identified using average temporal association among articles. Next, the task of

Table 1. Dataset characteristics

Nodes	422419
Edges	1056047
Average similarity coefficient	0.5
Number of news categories	4
Average timestamp	100000 s

Figure 1. Proposed research framework methodology


Similarity Computation
Cosine Similarity
News Corpus
Extraction
Domain -Wise Classification
Pre-processing using NLP
Use Temporal Association between Articles
Build News Spread Network (SI Diffusion Model)
Virality Detection
Recommendation of News Domain for Advertisement & Marketing
Discover spreading Pattern

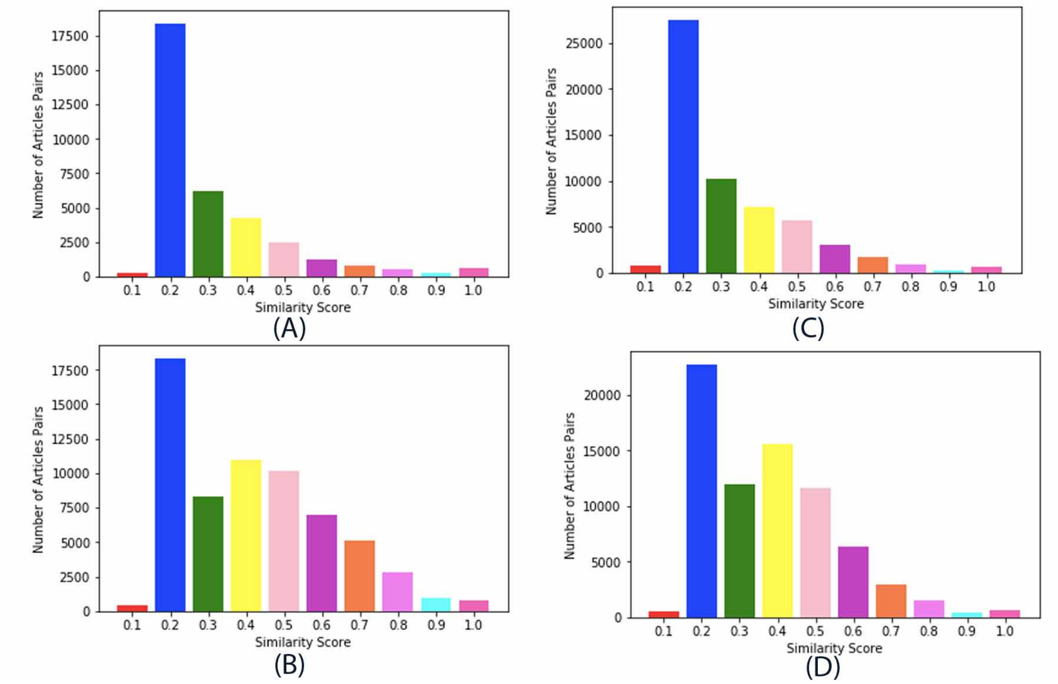
reconstructing the spread network of the news is triggered. This task is accomplished by deploying SI diffusion model. It is the simplest type of all the available epidemic models. It is a Susceptible–Infected model. In this model, initially, all the individuals of the simulation are born into it with no protection (susceptible). Once they get infected by other articles in the simulation and are not treated, individuals remain infected and infectious the whole time and also stay in association with the uninfected population. The SI model is numerically explained using differential equation using equation (2) and equation (3).

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{2}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} = \beta I\left(1 - \frac{I}{N}\right) \tag{3}$$

In above equation (2-3), S is the susceptible population, I is the infectious population, $N = S + I$ is the total population and β shows the infectious rate. Figure 3 represents the empirical SI model curve with constant rate of infection equal to

Figure 2. Similarity distribution curves in distinct networks

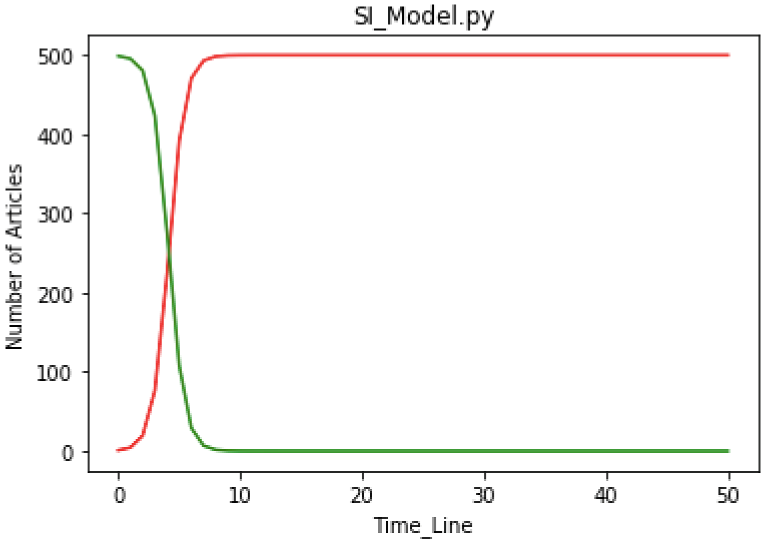


0.003. The x-axis represents the timeline and the y-axis represents the total number of articles which are susceptible and infectious on the scale of 500.

To reconstruct the SI model the nodes of the information spread network are defined as the news articles published from the selected domain and edges of the network are defined as an event when an article infects another article. This implies that all articles published after the first article should have been infected by any previously published article. For the four domains, the following step of algorithm is used separately for the reconstruction of the SI model.

1. Two parameters, temporal time window and the threshold similarity are taken into consideration. Now, for an article to be eligible as an infector, it should come before the infected article within a fixed time window and must have a score of similarity to the infected article greater than the threshold similarity.
2. First, 25 articles from the dataset are taken and the percentage of susceptible articles and the infectious articles on the basis of the time window and the threshold similarity is calculated.
3. Gradually, the articles are increased by the rate of 25 articles and percentage of susceptible articles and the infectious news articles are calculated.
4. The graph is plotted on the x-y plane using the above-calculated percentages to see the curve of change of percentages.

Figure 3. Empirical ideal SI model curve



For calculating the rate of change of infectiousness, percentages of infectious articles for different size of datasets are considered. Consecutive percentages are taken and the difference between the percentages is calculated to find the slope of change of infection in news articles. The rate of change of infection for each domain is calculated and slopes are compared by plotting them on the graph. Similarly, the rate of change of susceptibility is quantified and plotted for each domain for comparison.

4. FINDINGS

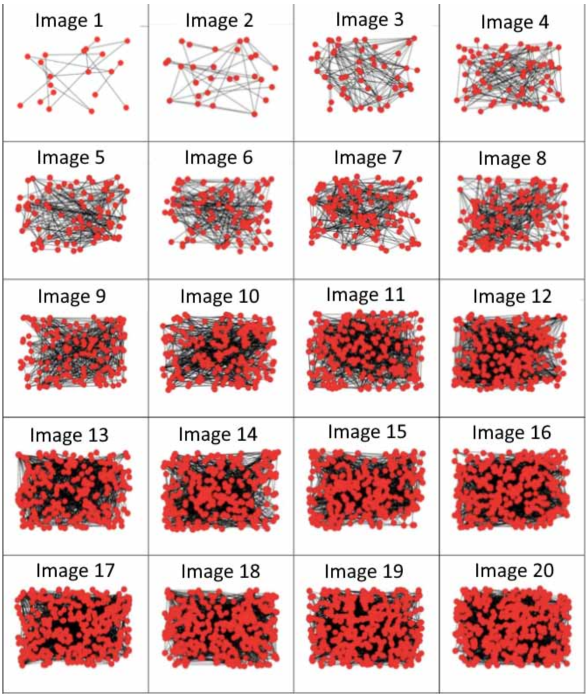
This section will focus on the results of our implementation. Figure 4: (a) depicts spreading of information in business network; (b) depicts spreading of information in medical network; (c) depicts spreading of information in technology network; and (d) depicts spreading of information in entertainment network.

In Figure 4, nodes in the spread networks of infection for each domain represent the articles and the edges show the infections between the articles. The graph for each domain also displays how the network becomes denser as we gradually increase the number of articles. Figure 5 shows the SI model curve for each network. The SI model curve for each domain describes the change of percentages of susceptible and infectious articles in a network with time. The x-axis shows the time-line and the y-axis shows the percentages of susceptible and infectious articles at any point of time.

Figure 6 shows comparison of rate of change of susceptibility and Figure 7 shows comparison of rate of change of infectiousness.

In figure 6, the x-axis shows the time line and the y-axis represents the slope of the change of percentages of susceptible articles. Graph is plotted for comparative

Figure 4a. Spreading of information in distinct networks



analysis with respect to rate of change of susceptibility for all domains. The domain with the highest change in percentages of susceptible articles and the least change in percentages of susceptible articles is identified. Also, it can note that the rate of change of susceptibility is negative and finally terminates to 0 which implies that the percentage of susceptible articles decreases with time and finally almost all articles are infected. Similarly in figure 7, the x-axis show the time line and the y-axis show the slope of the change of percentages of infected articles. Graph is plotted for comparative analysis with respect to rate of change of infectiousness for all domains. The domain with the highest change in percentages of infected articles and the least change in percentages of infected articles is identified. It can be noted that initially the rate of change of infections for business domain is 20%, technology domain is 40%, entertainment domain is 84% and medical domain is 48%. Clearly, the entertainment domain trumps every other domain with the highest rate of change of infection and the business domain loses the battle with the lowest rate of change of infection which shows that the people are most interested in entertainment news and least interested in business news. The order of preferred domain for recommendation is entertainment, medical, technology and business respectively.

5. CONCLUSION AND FUTURE WORK

We have applied SI diffusion model on four domains of news-Business, Medicine, Technology and Entertainment. By analyzing the graph of rate of diffusion of these

Figure 4b. Spreading of information in distinct networks

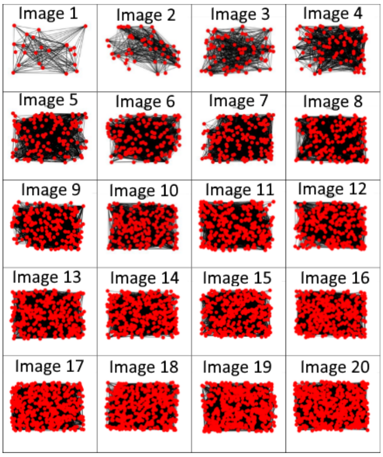
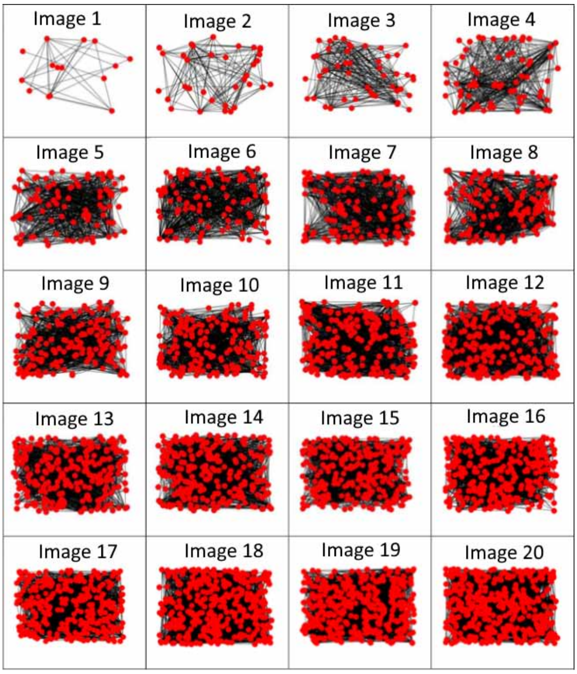
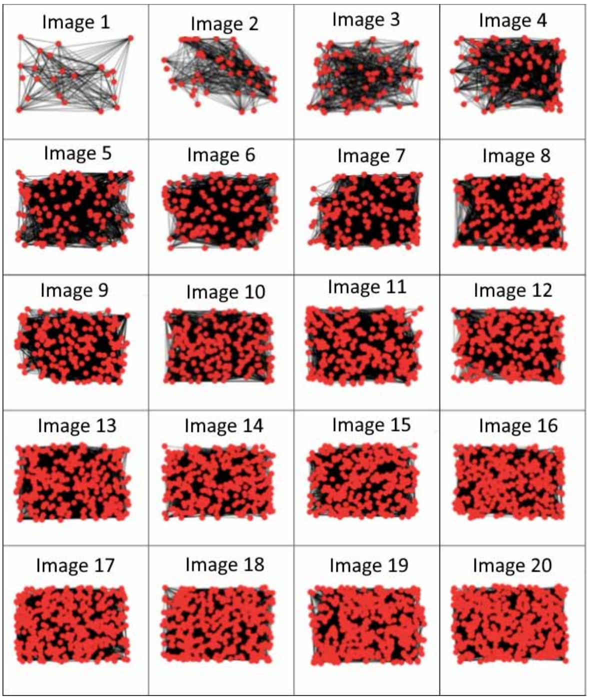


Figure 4c. Spreading of information in distinct networks



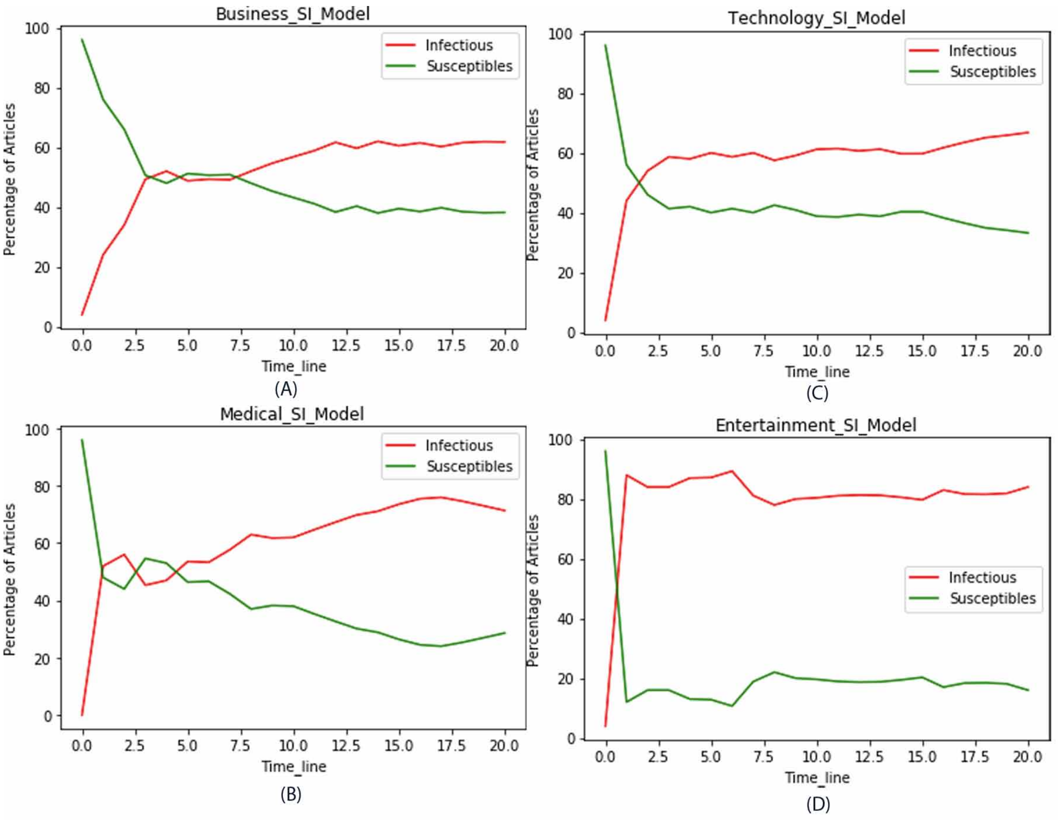
domains we conclude that entertainment has highest rate of infection, followed by technology, then medicine and in the last business. Therefore, entertainment domain has maximum influence and interest generation capability and business domain has minimum capability. An interesting future work would be boosting diffusion rate for influence maximization process by studying other diffusion models for social network analysis. Also, we can analyze authenticity of news to know whether the news which is spread is fake or real news using our reconstructed by using their rate of infection.

Figure 4d. Spreading of information in distinct networks



Also, further investigation is required to distinguish the wide range of diffusion patterns that can be observed in social networks

Figure 5. SI model curve for distinct news spread networks



(A)

(C)

(B)

(D)

Figure 6. Comparison of rate of change of susceptibility

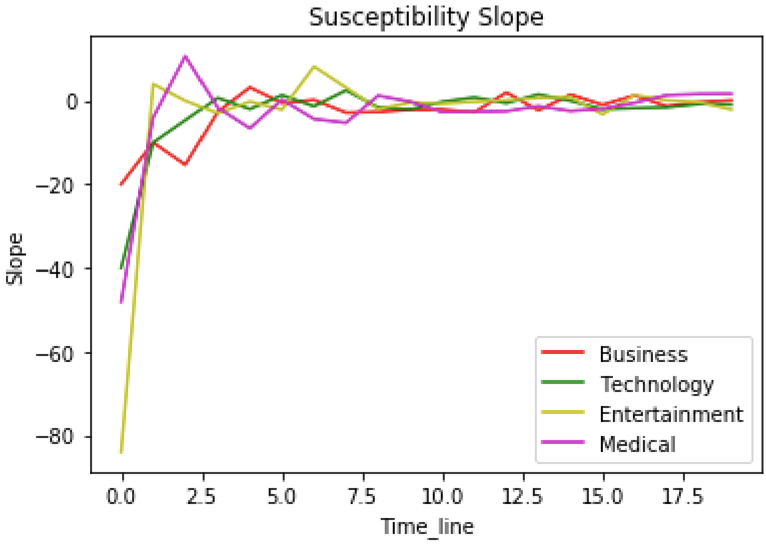
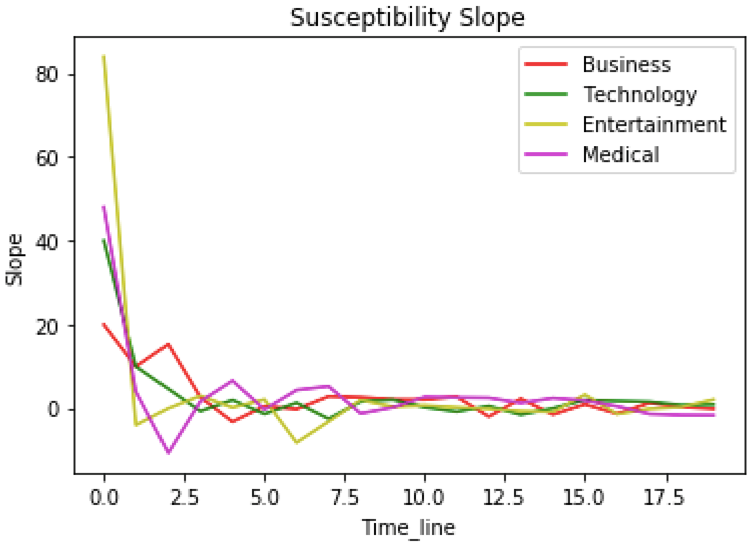


Figure 7. Comparison of rate of change of infectious



REFERENCES

Antaris, S., Rafailidis, D., & Nanopoulos, A. (2014). Link injection for boosting information spread in social networks. *Social Network Analysis and Mining*, *4*(1), 236. doi:10.1007/s13278-014-0236-y

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, February). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65-74). ACM. doi:10.1145/1935826.1935845

Borgs, C., Brautbar, M., Chayes, J., & Lucier, B. (2014, January). Maximizing social influence in nearly optimal time. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms* (pp. 946-957). Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611973402.70

Chowdhury, S. R., Purohit, H., & Imran, M. (2015, May). D-sieve: a novel data processing engine for efficient handling of crises-related social messages. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1227-1232). ACM. doi:10.1145/2740908.2741731

Dong, R., Li, L., Zhang, Q., & Cai, G. (2018). Information Diffusion on Social Media During Natural Disasters. *IEEE Transactions on Computational Social Systems*, *5*(1), 265–276. doi:10.1109/TCSS.2017.2786545

Fang, A., & Ben-Miled, Z. (2017). Does bad news spread faster?.

Gasparetti, F. (2017). *UCI Machine Learning Repository*. Italy: Faculty of Engineering, Roma Tre University. Retrieved from https://archive.ics.uci.edu/ml/datasets/News+Aggregator

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, 1(12).

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2015). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196.

Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2010, February). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 241-250). ACM. doi:10.1145/1718487.1718518

Jenders, M., Kasneci, G., & Naumann, F. (2013, May). Analyzing and predicting viral tweets. In *Proceedings of the 22nd international conference on world wide web* (pp. 657-664). ACM.

Ji, X., Chun, S. A., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, *5*(1), 13. doi:10.1007/s13278-015-0253-5

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013, August). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis* (p. 8). ACM. doi:10.1145/2501025.2501027

Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., & Kustarev, A. (2012, October). Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2335-2338). ACM. doi:10.1145/2396761.2398634

Li, S., Wu, W., & Du, D. Z. (2016). Effector detection in social networks.

Lin, Y., Chen, W., & Lui, J. C. (2017, April). Boosting information spread: An algorithmic approach. In *Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering* (pp. 883-894). IEEE.

Mussumeci, E., & Coelho, F. C. (2018). Reconstructing news spread networks and studying its dynamics. *Social Network Analysis and Mining*, *8*(1), 6. doi:10.1007/s13278-017-0483-9

Myers, S. A., & Leskovec, J. (2014, April). The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web* (pp. 913-924). ACM. doi:10.1145/2566486.2568043

Myers, S. A., Zhu, C., & Leskovec, J. (2012, August). Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 33-41). ACM. doi:10.1145/2339530.2339540

Oktay, H., Firat, A., & Ertem, Z. (2014). *Demographic breakdown of twitter users: An analysis based on names. Academy of Science and Engineering*. ASE.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, September). Influence and passivity in social media. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Berlin, Germany (pp. 18-33). Springer.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection* (pp. 7-17). Academic Press. doi:10.18653/v1/W16-0802

Sarkar, A., Chattopadhyay, S., Dey, P., & Roy, S. (2017, January). The importance of seed nodes in spreading information in social networks: A case study. In *Proceedings of the 2017 9th International Conference on Communication Systems and Networks,* (pp. 395-396). IEEE. doi:10.1109/COMSNETS.2017.7945410

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. doi:10.1109/TCSS.2017.2784184

Tang, Y., Xiao, X., & Shi, Y. (2014, June). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 75-86). ACM. doi:10.1145/2588555.2593670

Tatar, A., Antoniadis, P., De Amorim, M. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*(1), 174. doi:10.1007/s13278-014-0174-8

Wang, Y., Zeng, D., Zheng, X., & Wang, F. (2009, June). Propagation of online news: dynamic patterns. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics 2009* (pp. 257-259). IEEE. doi:10.1109/ISI.2009.5137321

Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds* (pp. 17599-601). Citeseer.

Zhou, C., & Guo, L. (2014). A note on influence maximization in social networks from local to global and beyond. *Procedia Computer Science*, *30*, 81–87. doi:10.1016/j.procs.2014.05.384

Kritika Jain is pursuing a BTech in Computer Science Engineering from Jaypee Institute of Information Technology, Sector -62, Noida, India. She is in her eighth semester. She is interested in the fields of social networks and study of diffusion models.

Somya Jain is working as an Assistant Professor in the Computer Science Engineering Department of Jaypee Institute of Information Technology. She is an IEEE Member and has six years of academic experience. Her interest areas include social networks, software engineering, and data and text mining.