Predicting Spatial Spread on Social Media

Rimjhim
Indian Institute of Technology Patna
India
rimjhim.pcs16@iitp.ac.in

Sourav Dandapat
Indian Institute of Technology Patna
India
sourav@iitp.ac.in

ABSTRACT

Understanding and prediction of spreading phenomena are vital for numerous applications. The massive availability of social network data provides a platform for studying spreading phenomena. Past works studying and predicting spreading phenomena have explored the spread in dimensions of time and volume, such as predicting total infected users, predicting popularity, predicting the time when content receives a threshold number of infected users. However, as the information spreads from user to user, it also spreads from location to location. In this paper, we attempt to predict the spread in the dimension of geographic space. In accordance with the past spreading prediction problems, we design our problem to predict the spatial spread at an early stage. For this, we utilized spatial features, social features, and emotion features. We feed these features into existing classification algorithms and evaluate on three datasets from Twitter.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Web mining**;
• **Social and professional topics** → *Geographic characteristics.*

KEYWORDS

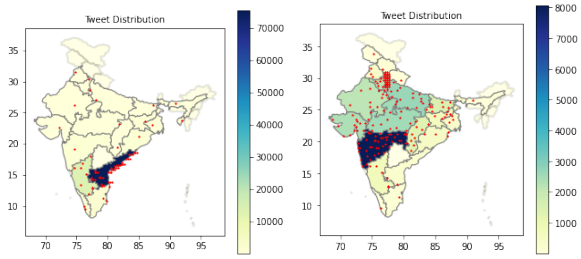
Social Networks, Spreading Phenomena, Location, Geographic Information Retrieval

ACM Reference Format:
Rimjhim and Sourav Dandapat. 2022. Predicting Spatial Spread on Social Media. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, Lyon, France, 4 pages. https://doi.org/10.1145/3487553.3524646

1 INTRODUCTION

The study of spreading phenomena is vital for many applications such as epidemic outbreak understanding, targeted marketing, virality prediction and many more. The study of spreading phenomena has also gained much importance for social media where information can spread like a forest fire. The phenomena of social contagion [15] in social media can result in creating viral content and this can also inhibit viral content. Online social media is now the voice of news, politics, jobs, awareness, and campaigns. Thus, understanding and predicting the spread on social media is of the utmost importance.

Studies related to spreading in online social media [11] have yielded useful insights about the spreading patterns. Another branch of related studies focuses on predicting parameters of spread such as popularity [20], virality [17], volume [10]. Prior studies have focused on predicting spread in terms of number of infected users, number of retweets/likes/shares, peak-time prediction. As content spreads from user to user, it also spreads from location to location [4, 6]. Past studies have characterized spatial spreading patterns but the prediction of spatial spread has still not received much attention. Models predicting spread in terms of volume may not be sufficient to predict spread in terms of geographical space. An online information with a larger volume doesn't necessarily have a larger spatial spread. Figure 1(a) shows the spatial spread of the hashtag $50DaysToAlluArjunBirthday$, which has 98,629 tweets (volume) and Figure 1(b) shows the spatial spread of the hashtag $14Feb\_ParentsWorshipDay$, which has 31,388 tweets (volume). Interestingly, we can observe in Figure 1 that hashtags with higher volume do not necessarily imply a wider geographical spread. It is noteworthy that, although the hashtag *50DaysToAlluArjunBirthday* has a larger volume, it is spread only in the southern part of the country, while *14Feb_ParentsWorshipDay*, which has a lesser volume, has spatial distribution throughout the country. In addition to providing broader insights into spreading phenomena, spatial spreading links spreading behavior to a real-world situation. This motivates us to design experiments that predict spatial spread.



(a) 50DaysToAlluArjunBirthday (b) $14Feb\_ParentsWorshipDay$

Figure 1: Spatial distribution of tweets from two hashtags $50DaysToAlluArjunBirthday$ and $14Feb\_ParentsWorshipDay$.

We define our problem in a similar manner to other spreading prediction tasks such as popularity prediction, which predicts the volume of content/information at an early stage. In this study, we predict the spatial distribution of information across geographical locations rather than its volume. We adopt the metric *focus* [2, 4] which quantifies the maximum intensity of geographical spread of

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524646

an information. A higher focus value indicates a local spread and a lower focus value indicates a global spread [6]. We predict spatial spread of hashtags with initial 50 tweets. In order to predict the spatial spread of tweets, we extract relevant features from the first 50 tweets and feed these features to a series of machine learning classifiers. We run our experiments on three twitter datasets and also compare the results with state-of-the-art popularity prediction methodologies. The results from this preliminary experiment shows that our feature can give accurate enough predictions of spatial spreading across varied datasets. There is a significant increase of around $11\% - 21\%$ in accuracy from state-of-the-art.

Rest of the paper is organized as follows. Section 2 presents the relevant literature. Section 3 describes the data on which experiments are performed and section 4 elaborates the adopted methodology for our experiments. Section 5 discusses the obtained results and section 6 concludes the study with future scope.

2 LITERATURE

In this section we detail past works from the areas of i. Spreading in social networks, ii. Spreading prediction, and iii. Spatial spreading.

Spreading in social networks The availability of social media data has given a platform for understanding the spreading phenomena. There have been discoveries of interesting phenomena and factors leading to spread. The user network [18], content features [9], environment [11] are major role players for spreading while other factors like geography [4] and homophily [7] have also been found to play a role. Multiple factors acting simultaneously on the content executes the phenomena of social contagion [15] which results in content spread. However, in this study we focus on spreading across different locations (states) in a country.

Prediction of Spread Past literature has attempted to predict multiple dimensions of spread and advances in machine learning has helped in obtaining outstanding results. In terms of volume, researches have predicted user volume [10], and event volume [20]. Other parameters of spread such as user engagement [5], peak time [19], virality [17] have also been successfully predicted on social media. When content spreads, it creates a cascade of engagement and a branch of literature focus on understanding and prediction on several dimensions of cascade. The past works [9] have predicted different dimensions of cascade. Spreading across space is also one important parameter of spreading phenomena. In this study, we predict spatial spreading of online content.

Spatial Spread As the availability of social media data has accelerated research around spreading behaviour. The availability of location field or the geotagged social media has also triggered researches around spatial spreading. [6] has analysed huge social media data to understand spatial spread across different countries of the world. Other studies [4, 14] have also analysed the phenomena of spatial spreading among different countries of the world. Visualization of spatial spreading also adds to the understanding of the phenomena as done by [13]. Unfurling of events across space has also been studied [12]. In this study, we use insights from these studies to derive useful features for prediction of spatial spreading.

3 DATASETS

For this we use three different location-tagged Twitter data described as following.

	Datasets		
	Geotagged India	COVID-India	COVID-USA
Total tweets	3,140,235	522,719	2,931,186
Total Hashtags	93,718	464,968	1,976,449
Total Location tweets	1,409,747	493,202	2,894,163
Final Hashtags(tweet count >50)	585	787	3257

Table 1: Descriptive Statistics of Datasets.

Geotagged India We collect geo-tagged tweets from Twitter Location Filtering Streaming API with 'India' as the required location for a period of around two months(6^{th} Feb - 30^{th} March 2018) which consists around $3M$ tweets, $1M$ unique users and a total of $93,718$ unique hashtags. .

COVID-India This dataset is a publicly available location tagged tweet dataset[1] related to COVID-19. This is global dataset, so we extract tweets belonging to the country India from this dataset resulting in a total of $0.5M$ tweets.

COVID-USA This dataset is also created from the publicly available global dataset used in *COVID-India* dataset. Here we select tweet from the country USA, resulting in a total of $2.9M$ tweets.

For all the three datasets, we tag each tweet to a state-level location within the relevant country from the given location field. For the country India (Geotagged India and COVID-India), we use the dictionary made available by [2] for tagging state-wise locations. For USA, the location tags already contain the state name, we use the same for tagging state-level location to a tweet. For all the datasets, we discard tweets which cannot be tagged to a state-level location. In this study, we predict spatial spread of hashtags. For each dataset, we collect all hashtags and tweets containing them. We keep hashtags with a minimum of 50 location-tagged tweets in the final dataset. Table 1 shows data statistics for the three datasets.

4 METHODOLOGY

In this section, we describe the problem statement followed by the methodology of feature construction and classification.

Problem Statement:

In this study, we perform an early stage prediction of spatial spread of hashtags. Borrowing the problem of early stage prediction of [17], we predict final spread (over all occurrences in the dataset) of a hashtag with initial 50 tweets only. For calculating the spatial spread we deploy the metric *focus*, used frequently in the past literature [6] which represents the maximum intensity of occurrence of a hashtag at any location. If P_i represents the probability of occurrence of a hashtag at any location i and L is the set of locations where the hashtag has occurred, then the focus score, F, of the hashtag is given as

$$Focus(F) = max_{i \in L}[P_i] \tag{1}$$

Focus is a continuous variable with range between 0 and 1. We convert this into a classification problem by dividing into two classes called *local* and *global*. We use the threshold of 0.4 for dividing into the two classes according to the study of [6]. A hashtag having overall focus value > 0.4 is a local hashtag and rest all is global hashtag. This converts our problem to a binary classification problem.

[1] https://crisisnlp.qcri.org/covid19

Feature Extraction For each hashtag, we first sort the initial 50 tweets according to its arrival time. For each feature, we create a vector of size 50 such as $x_0, x_1, x_2....x_{50}$ where each element of the vector is a feature value corresponding to tweets $t_0, t_1, t_2....t_{50}$. For this study, we have used spatial features, social features, and emotion features. Next, we explain details of each feature and its calculation on a set of tweets.

Spatial Features : In spatial features, we use 4 types of features explained as following:

Focus: This metric is the same as defined in equation 1 and measures the maximum intensity of a hashtag at any location. Note that we are predicting the same metric but the final prediction values are calculated over all the tweets but in feature calculation we are limited to initial 50 tweets. For creating the feature vector for *focus*, we calculate *focus* values in a cumulative way i.e. feature vector $f_1, f_2, f_3..f_{50}$ is calculated on tweets $t_0, [t_0, t_1], [t_0, t_1, t_2], ..[t_0, t_1..t_{50}]$.

Entropy: The metric *Entropy* [6] captures the randomness of occurrence of a hashtag and represents the number of bits required to capture the spatial uncertainty of a hashtag. Entropy can be calculated using the formula $Entropy(E) = \sum_{i \in l_e} -P_i \log_2 P_i$, where all variables are maintained as in the metric *focus*. If a hashtag is present only at a single location, the *entropy* value is 0, and this increases as the uncertainty of location increases.

Spread: Spread [6] measures the geographical extent of hashtag distribution and is calculated as the average *geographical distance* between a central location and all other locations where the hashtag has appeared. All the distances calculated in this study are geodesic distance [16]. For the central location, we consider the location which corresponds to the *focus* value and is termed as the *focus location*. We can formulate the *spread* as $Spread(S) = \frac{1}{|L|} \sum_{l_i \in L} dist(l_i - l\_focus)$, where *dist* is the *geographical distance*, l_focus is the *focus location*, and L is the set of locations except l_focus where the hashtag has appeared. We calculate *entropy* and *spread* in a cumulative way as done for *focus*.

Geographic Distance: This feature is the geographical distance between locations of two consecutive tweets. We have input tweets arranged in order of their arrival time. For the first tweet, it is 0 and for the rest it is geographic distance between locations of tweet t_n and t_{n-1}.

Social Features : Social media is characterized by many contextual information. We use contextual information from twitter to derive following social features.

Adoption Lag: This is the time lag between two consecutive tweets. We have 50 tweets, for the first tweet it is 0 and for all rest it is time difference between tweet t_n and t_{n-1}.

Follower Count: This is the follower count of the user who has tweeted the tweet. Since there are 50 tweets, there are 50 follower count also.

Retweet Status : Retweets do construct an important aspect of information spread. Here, for distinguishing an original tweet from retweet, we choose a binary value. For the created feature vector $x_0, x_1...x_{50}$, x_i is 1 for an original tweets and 0 for retweets.

Emotion Features : Past work finds that emotion plays a vital role in information diffusion [1]. In order to include this feature, we calculate the value of 5 types of emotion for each tweet. For each emotion, we opt for binary values of emotion. For finding out if an emotion is present in the tweet or not, we did majority voting on output of three popular approaches of calculating emotion. The three approaches are NRClex[2], Empath[3] and text2emotion[4]. We calculate emotions sad, anger, surprise, fear, happy. For each feature, if a tweet has a sad emotion, then it is given a value 1 else it is given the value 0.

Finally, the vectors of each feature are concatenated to create the final feature vector of each hashtag. The created vectors are then fed into existing machine learning classifiers to evaluate the predictions of spatial spreading.

Comparative Approaches: For a comparative study, we compare results of spatial spreading with state-of-the-art classification algorithms. Popularity prediction is one of the interesting studies around spreading phenomena. We also compare our proposed methodology with state-of-the-art popularity prediction methodologies. Comparative approaches are described next.

Base 1 [8]: This baseline used category booster classifier on a number of derived features for predicting social media popularity. Using the available features for our dataset, we model it for the task of spatial spread prediction.

Base 2 [3]: This is a neural network based approach which fuses multiple features for social media popularity prediction. We implement this using features available in our dataset and perform spatial spread classification in place of popularity classification. *Popular machine learning classification algorithms:* We also feed our derived features to a few popular machine learning algorithms which are known to give high performance for the task of binary classification. We use linear regression (LR), k-nearest neighbour (KNN), naive bayes (NB) and support vector machines (SVM) for analyzing the results of spatial spread prediction.

Tree based classification algorithms: Tree-based classifiers gives higher accuracy of predictions for hand-crafted features. We use three types of tree-based classification algorithms i.e. decision trees (DT), random forests (RF), and extra trees (ET). Our extracted features are used to train these algorithms for spatial spread classification.

Experimental settings For all the classification algorithms, we use the best parameters empirically. We use stratified 5-fold cross validation where we train for 4-folds of all hashtag data and test on the rest for each considered dataset. We use standard metrics *accuracy*, *macro average F1-score*, and *weighted average F1-score* which captures the overall performance of the classification. We run all the classification algorithms five times and report the average results.

5 RESULTS AND DISCUSSION

In this section, we present location-tagging evaluation, performance analysis of spatial spread classification and insights from the results. *Location-tagging Evaluation:* We first evaluate our location tagging approach used for creating the datasets. For evaluating location tagging, we deployed three independent graduate annotators. For

[2]https://pypi.org/project/NRCLex/
[3]https://github.com/Ejhfast/empath-client
[4]https://pypi.org/project/text2emotion/

Table 2: Prediction Results

Algorithms	Geotagged India			COVID-India			COVID-USA		
	Accuracy	Macro Avg F1-score	Weighted Avg F1-score	Accuracy	Macro Avg F1-score	Weighted Avg F1-score	Accuracy	Macro Avg F1-score	Weighted Avg F1-score
Base1 [8]	0.72	0.72	0.72	0.59	0.45	0.50	0.65	0.39	0.51
Base2 [3]	0.51	0.34	0.35	0.71	0.41	0.59	0.78	0.73	0.79
LR	0.50	0.43	0.58	0.75	0.60	0.80	0.69	0.53	0.77
KNN	0.70	0.70	0.70	0.63	0.59	0.62	0.66	0.61	0.66
NB	0.50	0.34	0.66	0.40	0.39	0.44	0.40	0.37	0.48
SVM	0.51	0.34	0.68	0.71	0.41	0.83	0.65	0.40	0.78
DT	0.51	0.34	0.68	0.84	0.80	0.84	0.65	0.40	0.79
RF	0.79	0.79	0.80	0.84	0.81	0.83	0.78	0.75	0.78
ET	**0.83**	**0.83**	**0.83**	**0.86**	**0.83**	**0.87**	**0.80**	**0.75**	**0.81**

each dataset, we randomly sampled 5000 tweets. Annotators were provided with the sampled tweets and their corresponding location meta data. Annotators were advised to tag a state-level location to each tweet from internet searches and gazetteers. The final location of a tweet is the location given by at least two annotators; the rest are discarded. Then we measure the precision score of the location-tagging from our methodology with respect to the tagged location by the annotators. This delivered an overall precision score of 0.95, 0.96, and 0.98 for the *Geotagged India*, *COVID-India*, and *COVID-USA* datasets respectively.

Performance Analysis Table 2 shows the performance evaluation of spatial spread classification on three twitter datasets. From the table, we can summarize the following conclusions.

We observe that Extra Tree (ET) classifiers applied on our features are able to obtain high values of prediction accuracy i.e. 0.83, 0.86, and 0.80 for *geotagged India*, *COVID India*, and *COVID-USA* datasets respectively. Also, the predictions are significantly higher (11%-21% by accuracy) than state-of-the-art popularity prediction algorithms. This observation is consistent with all the metrics. This signifies that our features are meaningful and contain sufficient information for predictions. Additionally, other spreading prediction methodologies such as popularity prediction do not perform well for the task of spatial spread prediction.

Tree-based classification algorithms (DT, RF and ET) are performing better than LR, KNN, NB, and SVM. Additionally, we note that the ratio of classes may be imbalanced in a scenario of spatial spreading prediction. We observe that other classification algorithms like Base2, NB, SVM are giving a lower macro average F1-score. However, RF and ET classifiers give higher macro average F1-score, indicating accurate prediction of both the classes(local and global in our case).

6 CONCLUSION AND FUTURE WORK

In this study, we predict spatial spread in online social media. Experimentation on three datasets gives high accuracy of predictions and this is 11% – 21% higher than state-of-the-art spread (popularity) prediction methodologies. This study sheds light on the geographical aspect of spreading phenomena and invites exploration of the spatial dimension of spread. As a future work, we intend to add more features and develop more robust classification algorithms for predicting spatial spread.

REFERENCES

[1] Roshni Chakraborty. 2018. Characterizing User Reactions Towards Twitter's 280 Character Limit. In *Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation*. 48–51.

[2] Nikhil Cheke, Joydeep Chandra, and Sourav Kumar Dandapat. 2020. Understanding the Impact of Geographical Distance on Online Discussions. *IEEE Transactions on Computational Social Systems* 7, 4 (2020), 858–872.

[3] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2682–2686.

[4] Didier Henry, Erick Stattner, and Martine Collard. 2018. Information Propagation Routes between Countries in Social Media. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, ACM, 1295–1298.

[5] Yuheng Hu, Shelly Farnham, and Kartik Talamadupula. 2015. Predicting User Engagement on Twitter with Real-World Events. *ICWSM* 15 (2015), 168–177.

[6] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 667–678.

[7] Márton Karsai, Gerardo Iñiguez, Riivo Kikas, Kimmo Kaski, and János Kertész. 2016. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific reports* 6 (2016), 27178.

[8] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4565–4569.

[9] Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei. 2018. Joint modeling of text and networks for cascade prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[10] Zhiwei Liu, Yang Yang, Zi Huang, Fumin Shen, Dongxiang Zhang, and Heng Tao Shen. 2017. Event early embedding: Predicting event volume dynamics at early stage. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 997–1000.

[11] Byungjoon Min and Maxi San Miguel. 2018. Competing contagion processes: Complex contagion triggered by simple contagion. *Scientific reports* 8, 1 (2018), 10422.

[12] Paul Mousset, Yoann Pitarch, and Lynda Tamine. 2018. Studying the Spatio-Temporal Dynamics of Small-Scale Events in Twitter. In *Proceedings of the 29th on Hypertext and Social Media*. ACM, 73–81.

[13] Vanessa Peña-Araya, Anastasia Bezerianos, and Emmanuel Pietriga. 2020. A comparison of geographical propagation visualizations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[14] Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld, and Niklas Henckell. 2019. Hashtag usage in a geographically-local microblogging app. In *Companion Proceedings of The 2019 World Wide Web Conference*. 919–927.

[15] Daniel A Sprague and Thomas House. 2017. Evidence for complex contagion models of social contagion from observational data. *PloS one* 12, 7 (2017).

[16] Kendall Taylor, Kwan Hui Lim, and Jeffrey Chan. 2018. Travel itinerary recommendations with must-see points-of-interest. In *Companion Proceedings of the The Web Conference 2018*. 1198–1205.

[17] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific reports* 3, 1 (2013), 1–6.

[18] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting Successful Memes Using Network and Community Structure.. In *ICWSM*.

[19] Hai Yu, Ying Hu, and Peng Shi. 2020. A prediction method of peak time popularity based on twitter hashtags. *IEEE Access* 8 (2020), 61453–61461.

[20] Shuo Zhang and Qin Lv. 2017. Event organization 101: Understanding latent factors of event popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.