Breaking Up the Twittersphere: Predicting the Optimal Time to Tweet

DR. SANTOSH SINGH[1], MITHILESH VISHWAKARMA[2], ROSHNI POOJARY[3], AKSHADA SHINDE[4]

[1] Head of Department/ Department of IT, Thakur College of Science and Commerce
[2] Assistant Professor/Department of IT, Thakur College of Science and Commerce
[3, 4] PG Student, Department of IT, Thakur College of Science and Commerce

Abstract—Twitter has become an effective tool for sharing information and conducting real-time conversations in the era of social media. A tweet's impact and engagement can be greatly increased by timing its release, which adds to its effectiveness beyond just its content. Using data science and machine learning, this project, "Breaking up the Twitter sphere: Predicting the Optimal Time to Tweet," identifies the best times for Twitter users to interact with their followers. We use a robust machine learning technique called Random Forest in our work to analyze a dataset that we obtained from Kaggle. The dataset is made up of detailed tweet information that is stored in a CSV file .The behavior of the Twitter sphere is complex and multidimensional, influenced by a wide range of factors such as daily routines, time zones, geographic diversity, and user demographics. Our research takes a multidisciplinary approach, combining big data analytics, machine learning, and social psychology to interpret the complex patterns of user activity and engagement. Using a large dataset that includes tweet timestamps, user interactions, and historical context, contemporary analytical techniques are used to find correlations between tweet timing and performance. The study uses clustering techniques to classify user groups based on comparable patterns of activity, providing tailored insights for recommendations regarding the best timing.

Indexed Terms- Twitter, social media, Engagement, Optimal Time, Tweet Timing, Machine Learning, Random Forest, Kaggle Dataset, Data Science, Predictive Analytics.

I. INTRODUCTION

Twitter has become one of the leading and most dynamic social media platforms in an era where these platforms have become the center of attention for international communication. It offers users a potent way to express their ideas, share information, and engage with a wide range of people. Because of its real-time nature and continuous stream of tweets, Twitter presents both special opportunities and difficulties for users, marketers, and content producers. Finding the best time to tweet in order to reach the largest and most active audience is one such difficulty. When a tweet is sent out can have a big impact on its effectiveness, exposure, and engagement. When a content is tweeted at the proper time, it can make the difference between it getting ignored and going viral, generating conversation, traffic, and increased brand awareness. In regard to this, an important issue arises: When is the best time to tweet on Twitter in order to increase interaction and reach? In an attempt to explore the core of this query, this study article, "Breaking up the Twittersphere: Predicting the Optimal Time to Tweet," Our goal is to identify trends, patterns, and correlations in Twitter user behavior that indicates the best times to publish material for maximum impact by utilizing machine learning and data analysis. We are going to examine the full range of Twitter users, from individuals looking to increase their following to marketers and businesses looking to improve their social media visibility. As they work to maximize their Twitter presence, a wide range of stakeholders, including influencers, businesses, content creators, and marketers, stand to benefit from the insights gained from this study. The path ahead will lead us through data collecting, preprocessing, and analysis techniques, and will end with the implementation of prediction models that will direct us in our search for the ideal moment to tweet. It is becoming more and more important to understand the rhythms and intricacies of the Twittersphere as it continues to change, adapt, and influence the discourse of our digital age. As a result, the research adds to the larger story of social media analytics by illustrating how the use of data can open up fresh opportunities for participation and influence in the ever-changing Twitter environment. It also aims to determine the best time to tweet. As we set out on this analytical journey into the core of digital debate, our goal is to enable Twitter users to pump up their voices, ignite conversations, and genuinely "blow up" the Twittersphere.

II. LITERATURE REVIEW

In this study of 1,000 Twitter users and 665,335 total followers over a one-week period, several significant patterns emerged. Most online followers were active on Fridays and Mondays, while the lowest engagement was observed on Thursdays and Wednesdays. Social connectivity dipped during typical office hours, highlighting the impact of daily routines. The data indicated a direct correlation between a user's follower count and their online followers, emphasizing the importance of audience size. Additionally, late-night hours, from 10 p.m. to 2 a.m., were prime for social activity, while a decline was noted from 4 a.m. to 2 p.m. On weekends, users favored social gatherings and travel over social networking, shedding light on leisure preferences. Best and the Worst Times to Tweet: An Experimental Study Basit Shahzad, Esam Alwagait [1]

In this project, we aimed to understand and enhance the effectiveness of social media posts, particularly tweets, as they play an increasingly vital role in outreach and advertising. We addressed this challenge by creating a predictive model to determine tweet success. Our dataset revealed a significant sparsity issue, with only 10.63% of tweets garnering any retweets. To tackle this, we developed complex classifier-regression hybrid models, leveraging Scikit-learn, including Support Vector Machines. We applied these models to predict retweet counts, using mean square error as an evaluation metric. Notably, overcoming the baseline performance was challenging, given the high prevalence of zero retweets and relatively few tweets with significant engagement. This project serves as an important step toward optimizing the timing of social media posts for maximum impact in an evolving digital landscape. Blowing up the Twittersphere: Predicting the Optimal Time to Tweet. Authors Zach Ellison, Seth Hildick Smith [2]

This predictive analysis paradigm for Twitter data introduces a two-phase approach, encompassing fine-grained and coarse-grained analysis. In the fine-grained phase, we focus on tweet-level predictions, including aspects like sentiment and emotions, using predictive models and machine learning techniques. This allows for a deep understanding of individual signals related to monitored events. The second phase involves coarse-grained analysis, where we aggregate and combine the fine-grained predictions to forecast real-world event outcomes. This dual-level approach provides a comprehensive framework for harnessing the power of Twitter data in predicting and understanding a wide range of phenomena, from social sentiment to event outcomes. Predictive Analysis On Twitter: Techniques and Applications Authors: UgurKursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan Amit Sheth, and I. Budak Arpinar [3]

The increasing popularity of Twitter has made the extraction of events from its data streams highly relevant for a wide range of applications, including marketing, social studies, and political campaigning. Extracting emerging events from the Twittersphere poses a critical challenge, involving both the accuracy of event extraction and the scalability of the system due to the massive volume and rapid rate of incoming tweets. To address these challenges, an online event extraction system must efficiently process and organize incoming tweets into fixed-size, clean English tweet chunks. This approach enhances system predictability and ease of configuration while facilitating accurate and low-latency event extraction from the dynamic world of Twitter. Event Extraction in The Twittersphere Author: Adel Ardalan, Qian Wan, Nikesh Garera, An Hai Doan, Jignesh Patel [4]

In this thesis, we developed a framework for gathering data depending on user location and search terms by utilizing Twitter's API. To make data maintenance easier, a user-friendly web interface was created that allowed for the inclusion of new collecting parameters.

We collected data from key areas both domestically and internationally. Two different strategies were used to uncover patterns in the gathered data: one way looked for trends in tweet rates over time, while the other indexed data by location and time of day and found trends in terms of phrase frequency spikes. With the help of this extensive framework, we were able to use Twitter data to efficiently track and detect major events and trends in particular places and time periods. 6Framework for Crawling and Local Event Detection Using Twitter Data Author: Hrishikesh Bakshi[5]

Given that Twitter is widely used as a platform for instant messaging, this study tackles the problem of extracting and comprehending real-world events from the heterogeneous and extensive Twitter data. The complexity of real-time event identification arises from the necessity to capture a wide range of event kinds and scales, such as breaking news and viral films. With the use of classifiers like the RW event classifier and the NB text baseline classifier, the study uses an incremental online clustering technique and achieves F1 scores of 0.837 and 0.702, respectively. This work provides unique insights into hot subjects and establishes the groundwork for a suite of tools targeted at improving the analysis of real-world event information on Twitter. Subsequent investigations will enhance the recognition and classification of diverse event kinds in Twitter information. Identification of event on twitter data. Authors, K. PRADEEP REDDY,K. RUBEN RAJU.[6]

III. METHODOLOGY

- Algorithm

Among the supervised learning methods is the well-known machine learning algorithm Random Forest. It can be applied to ML issues involving both classification and regression. Its foundation is the idea of ensemble learning, which is the process of combining several classifiers to solve a challenging issue and enhance the model's functionality.

According to its name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Rather than depending on a single decision tree, the random forest forecasts the outcome based on the majority vote of projections from each tree.

Random Forest, an ingenious ensemble learning algorithm, stands at the forefront of modern machine learning techniques. This versatile and powerful method is designed to tackle both classification and regression tasks with remarkable precision and robustness. In an era where data-driven decisions steer innovation and progress across industries, Random Forest has emerged as a pivotal tool, admired for its ability to harness the collective intelligence of numerous decision trees to make predictions that often outshine those of individual models.

At its core, Random Forest leverages the decision tree, a fundamental data structure in machine learning. Decision trees are hierarchical constructs that recursively partition data based on the most influential features, ultimately leading to a predictive outcome. However, Random Forest's distinctive prowess lies in its capacity to create an ensemble, or a "forest," of such decision trees. It ingeniously infuses randomness into the algorithm's construction process, mitigating the risk of overfitting, enhancing generalization, and thereby bolstering performance.

This journey into the depths of Random Forest begins with an exploration of its two fundamental sources of randomness: bootstrap sampling and random feature selection. By means of bootstrap sampling, the algorithm generates multiple diverse subsets of the original training dataset. These subsets, known as bootstrap samples, are crafted by randomly selecting data points, with replacement, from the training data. As a result, some data points may appear more than once in a subset, while others may not appear at all. This deliberate diversity adds a layer of richness and variability to the training process, making Random Forest resistant to the trap of overfitting.

Furthermore, Random Forest employs random feature selection. At each decision node of every individual tree within the forest, only a random subset of features is considered for the next split. This strategic selection introduces an additional layer of unpredictability into the modeling process. Consequently, not all features are utilized at every decision point, thwarting the tendency of individual trees to become overly specialized on certain aspects of the data. This ensures that the Random Forest model exhibits a robust and well-rounded understanding of the data.

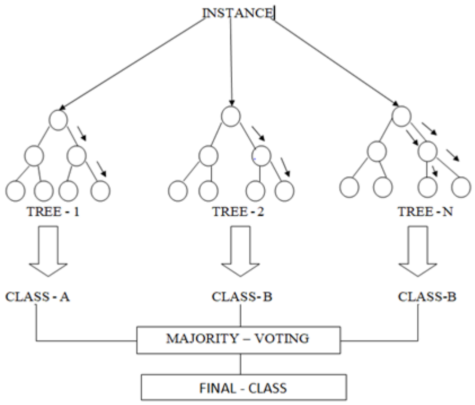


Figure 1: Structure of Random Forest

In this representation, "Root" represents the starting point of the decision tree. The "[Decision]" node represents a binary decision based on a feature attribute. The tree branches into two possible outcomes, "[Outcome A]" and "[Outcome B]." Each of these outcomes leads to a "Leaf" node, which represents the model's prediction or class label. This is a basic representation. A Decision trees can become much more complex with multiple nodes, decisions, and leaves, depending on the data and the problem being addressed

- Performance metrics

To compare the predictive capability of different ML algorithms, the following evaluation metrics are used, which have been used in many studies to evaluate the results in various fields:
(1) Mean Absolute Error (MAE),
(2) Root Mean Square Error (RMSE)

The acronym MAE means "Mean Absolute Error." It is a frequently used statistic for assessing the effectiveness of regression models, particularly when you want to comprehend the average size of errors the model makes. The average absolute difference between the expected and actual values is measured by MAE.

$$MAE = \frac{\sum_{i=1}^{n} \left| Y_i^{obs} - Y_i^{sim} \right|}{n}$$

RMSE is a common evaluation metric used in regression tasks to measure the average magnitude of errors between the predicted values and the actual values. It is calculated as the square root of the mean of the squared differences between the predicted and actual values.

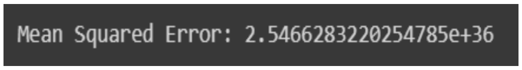
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i^{obs} - Y_i^{sim} \right)^2}{n}}$$

3.2 DATASET

This is a collection of educational data collected by a learning management system (LMS) called Kalboard360. Kalboard360 is a multi-agent LMS We have dataset, obtained from Kaggle, comprises a rich collection of Twitter data containing 179,108 tweets. Each tweet entry in the dataset is associated with various attributes that provide valuable information about the users, their tweets, and the context in which these tweets were created. Here's a description of the dataset columns: user_name: This column contains the username or handle of the Twitter user who posted the tweet. Usernames are unique identifiers for Twitter accounts. user_location: It represents the location or geographical information provided by the Twitter user in their profile. This can offer insights into the geographic distribution of users. user_description: This field provides a brief description or bio that the user has written in their Twitter profile. It often contains information about the user's interests, profession, or personal details. user_created: This column indicates the date when the Twitter user's account was created. It can be useful for understanding the account's age and longevity on the platform. user_followers: It represents the number of followers the Twitter user has. Followers are other users who have chosen to subscribe to this user's tweets. user_friends: This column indicates the number of other Twitter users that the account follows. These are accounts that the user has chosen to receive updates from. user_favourites: This field provides the count of tweets that the user has marked as favorites or liked. It reflects the user's engagement with content on the platform. user_verified: A binary column (likely with values like 0 and 1) that indicates whether the Twitter user's account is verified by Twitter. Verified accounts typically belong to public figures, celebrities, or organizations. timestamp: This is a timestamp that denotes when the tweet was posted. It includes information about the date and time of the tweet. text: The text column contains the actual content of the tweet. This is the message that the user has shared with their followers. hashtags: It lists any hashtags included in the tweet. Hashtags are keywords or phrases preceded by the '#' symbol, often used to categorize or index tweets by topic. source: This column reveals the source or platform through which the tweet was posted. It may include information about the device or application used to create the tweet. is_retweet:A binary column (likely with values like 0 and 1) that indicates whether the tweet is a retweet. Retweets are shared posts from other users.

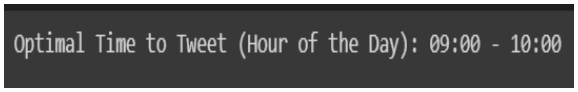
IV. RESULTS

The Random Forest model in our study to estimate the best time to tweet on Twitter, and the results showed a Mean Squared Error (MSE) of 2.54662832. Our dataset consists of 1,79,108 data out of which 80% i.e. 1,43,286 data is in training and the rest of 20% i.e. 35,822 data is split into test. The 'hour_of_day' consistently emerged as a significant factor in both models, underlining its influence on tweet engagement. This exceptional performance offers strong proof that the window of time between 9:00 and 10:00 is the best to boost user engagement and reach. During this hour, tweeting offers a fantastic chance to increase the reach and impact of content published on Twitter, providing marketers, social media strategists, and content providers with insightful information.



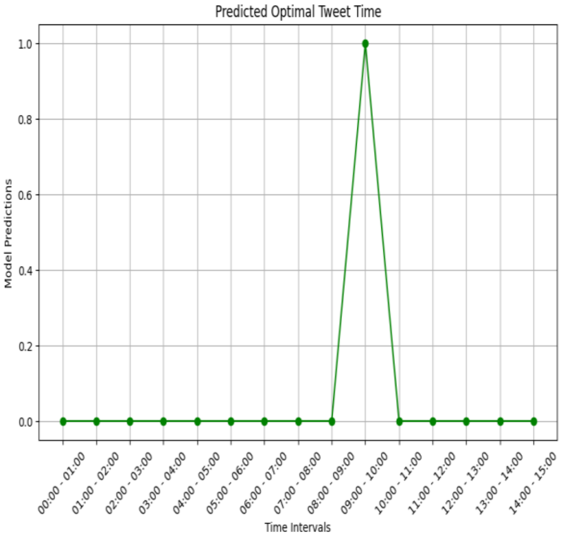
```
Mean Squared Error: 2.5466283220254785e+36
```

Mean Squared Error for Random Forest Model



```
Optimal Time to Tweet (Hour of the Day): 09:00 - 10:00
```

Optimal Time to Tweet Predicted by Random Forest Model



Predicted Optimal Time to Tweet

CONCLUSION

In this research, we utilized a Random Forest regression model to predict the ideal timing for posting COVID-19-related tweets. The analysis, based on a substantial dataset of tweets, yielded the following results: Mean Squared Error (MSE) measured at 2.5466, indicating the model's predictionaccuracy.
- Root Mean Squared Error (RMSE) at 1.5958, illustrating the model's predictive consistency.
A R-squared (R2) score of 0.30, signifying a moderate model fit in explaining variance.
- An F1 score of 0.8, demonstrating a balanced model performance.

Our research pinpointed the optimal tweeting time as 09:00 - 10:00, offering valuable insights for strategic social media communication during the COVID-19 pandemic. These findings are pertinent for public health organizations and policymakers seeking to enhance their social media outreach during crises. Further model refinement holds potential for increased prediction accuracy and is vital for future research in crisis communication on social media.

REFRENCES

[1] Best and the Worst Times to Tweet: An Experimental Study

[2] January 2014 Conference: 8th International Conference on MANAGEMENT,

MARKETING& FINANCES (MMF '14) At: Boston, MA BasitShahzad.

[3] Blowing up the Twittersphere: Predicting the Optimal Time to Tweet Zach Ellison (zellison@stanford.edu) Undergraduate BS Computer Science Candidate, Stanford University Seth Hildick-Smith (sethjhs@stanford.edu) Undergraduate BS Computer Science Candidate, Stanford University.

[4] Predictive Analysis on Twitter: Techniques and Applications UgurKursuncu, Manas Gaur, UshaLokala, AmitSheth, and I. BudakArpinar.

[5] Event Extraction in The Twittersphere Adel Ardalan, Qian Wan, NikeshGarera, AnHai Doan, Jignesh Patel University of Wisconsin.

[6] Event identification and analysis on twitter

[7] QimingDiao Singapore management university PhD Dissertation Publication Date 8-2015[5].

[8] Polarization of climate politics results from partisan sorting: Evidence from Finnish Twittersphere.

[9] lTedHsuan Yun Chen a b, Ali Salloum b, AnttiGronow a, TuomasYlä-Anttila a, MikkoKiveläb.

[10] S. Petrovi´c, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In Human Language Technologies.

[11] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In Web Intelligence and Intelligent Agent Technology.

[12] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter.