Predictive Analytics for Social Media Virality

Aditi Shah
MSc. Data Analytics
San Jose State University
San Jose, USA
aditirajesh.shah@sjsu.edu

Anshika Khandelwal
MSc. Data Analytics
San Jose State University
San Jose, USA
anshika.khandelwal@sjsu.edu

Parth Marathe
MSc. Data Analytics
San Jose State University
San Jose, USA
parthchinmaya.marathe@sjsu.edu

Tanisha Dhopeshwar
MSc. Data Analytics
San Jose State University
San Jose, USA
tanisha.dhopeshwar@sjsu.edu

Vinay Bhati
MSc. Data Analytics
San Jose State University
San Jose, USA
vinay.bhati@sjsu.edu

***Abstract*—Predictive analytics in social media has emerged as a critical area of research, providing insights into user behavior, content virality, and the dynamic interactions within social networks. This paper synthesizes findings from four seminal studies that employ various machine learning techniques to predict the popularity and virality of social media content. Through a comparative analysis of these methodologies, we identify key predictive factors and model efficacies, ranging from tweet volume and sentiment analysis to community structures and early engagement metrics. Our review highlights the nuanced roles these factors play in influencing content spread, offering robust strategies for content creators and digital marketers to enhance online engagement.**

***Index Terms*—social media, predictive analytics, machine learning, content virality, user engagement**

I. Introduction

The proliferation of social media platforms has provided a rich dataset for analyzing human interaction and content dissemination. Understanding the factors that contribute to the virality of content can aid stakeholders in optimizing strategies for maximum impact. This paper reviews significant contributions in the realm of predictive analytics, focusing on the effectiveness of different machine learning techniques in forecasting the popularity and reach of social media content. We discuss the implications of these predictions in various sectors, including entertainment, marketing, and public information, emphasizing the practical applications of this research in crafting responsive and informed social media strategies.

II. Literature Review

[1] The study represents a seminal exploration into the predictive capabilities of social media, specifically Twitter, in forecasting movie box office revenues. The authors meticulously collected and analyzed tweets in the periods leading up to film releases, applying various machine learning techniques to understand the relationship between public sentiment expressed in tweets and the financial performance of these films. The study found that both the volume of tweets and the nature of sentiments they expressed were highly indicative of the opening weekend's box office performance, outperforming traditional market prediction models based on factors like budget and genre. This research not only highlighted the feasibility of using social media analytics as a predictive tool in entertainment economics but also suggested broader applications in other sectors where public opinion and engagement play critical roles.

[2] This study extended the investigation into the virality of content, particularly news articles across social media platforms. Their study focused on dissecting the attributes of news articles that most significantly impact their likelihood of becoming viral. Using a robust dataset that included a variety of article features such as topics, sentiment expressed, and the presence of influential entities (like celebrities or significant political figures), the researchers applied logistic regression analysis to predict popularity outcomes. They concluded that emotionally charged content, especially those that could incite anger or anxiety, tended to have higher virality. Additionally, articles about divisive and sensational topics were more likely to spread rapidly, illustrating the complex interplay between content characteristics and user engagement behaviors in the digital news cycle.

[3] This paper examines the community structures within social networks and their impact on information spread provided groundbreaking insights into how content virality is facilitated or hindered by the network's inherent structure. By applying graph theory and network partitioning methods, they could identify 'communities' or clusters within larger networks. Their findings suggest that while tightly-knit communities are crucial for the initial propagation of information, 'bridges' or links between disparate communities are essential for achieving widespread dissemination. This study underscores the significance of strategic content placement and targeted marketing within network substructures to maximize outreach and engagement.

[4] This study focuses predicting cascades in social networks addresses the predictability of large-scale sharing of posts or photos within platforms like Facebook. By analyzing various factors such as early user engagement (likes, shares, comments) and intrinsic content features (media type, content

novelty), they developed models to forecast which posts were likely to 'go viral'. Their approach involved complex machine learning algorithms that provided a nuanced understanding of how different elements interact to contribute to the likelihood of a cascade. This research has profound implications for designing algorithmic strategies that optimize for user engagement and can be adapted to predict outcomes in various content-driven online platforms.

III. CRISP-DM Approach

The CRISP-DM methodology offers a structured framework for data mining projects, encompassing six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This approach aids in navigating the complexities of extracting actionable insights from vast datasets. In our project, we applied the CRISP-DM methodology to develop a predictive model for social media post virality, which is critical for optimizing content strategies and enhancing user engagement.

A. Business Understanding

The project begins with a clear definition of objectives from a business perspective. Our aim was to create a predictive model capable of assessing the potential virality of social media posts to assist marketers and content creators in strategizing their engagements more effectively.

B. Data Understanding

This phase involves initial data collection and exploration to gain insights into the data's characteristics. We compiled a comprehensive dataset that includes text content, user engagement metrics, and metadata from social media platforms. Preliminary exploratory data analysis helped us understand the feature distributions and relationships, highlighting potential data challenges.

C. Data Preparation

Data preparation involved rigorous cleaning, transformation, and feature selection processes. We addressed missing values, normalized numerical data, and encoded categorical variables to prepare the dataset for modeling. The data was then split into training and testing sets to ensure robust model training and evaluation.

D. Modeling

We explored various machine learning algorithms to construct models capable of predicting virality. Techniques employed included Bagging and Boosting to enhance model accuracy and robustness. Specific models tested were Ridge Regression (Bagging), Lasso Regression (Bagging), Decision Trees with AdaBoost, Linear Regression with AdaBoost, and Gradient Boosting Regressors. Principal Component Analysis (PCA) was also utilized to reduce the dimensionality of the dataset, which improved computational efficiency and model performance.

E. Evaluation

The evaluation phase was crucial to ascertain the performance of our models. We utilized metrics such as accuracy, precision, recall, and the F1-score to measure the effectiveness of each model. The Lasso Regression model, both with and without PCA, showed excellent performance, offering a good balance between accuracy and computational demands. The Gradient Boosting Regressor also performed well, particularly in configurations without PCA.

F. Deployment

The final step involved deploying the model in a real-world setting, where it can provide actionable insights to users. While the specifics of deployment can vary, integrating our virality prediction model into existing social media management tools or digital marketing platforms would allow for real-time content strategy optimization based on the model's virality predictions.

By following the CRISP-DM framework, our project adhered to a systematic process from the initial business understanding to the final deployment, ensuring the development of a robust and functional predictive model aligned with the strategic goals of enhancing social media content engagement.

IV. System Architecture

The flow of the project was meticulously planned to obtain the best results. It is described in detail in the next section and visually summarized in Figure 1.

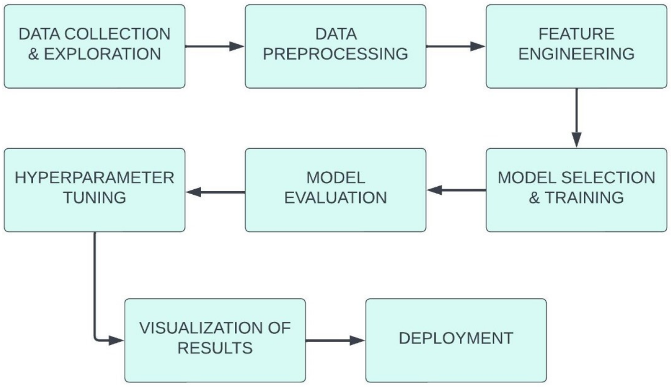

Fig. 1. Project Flow

V. Data Preparation and EDA

We began our data preparation by checking the dataset for any missing values. Fortunately, no missing values were found, allowing us to proceed directly to the exploratory data analysis (EDA) phase. Our EDA aimed to uncover key insights and patterns within the data.

The line plot in figure 2 displays the trend in the number of created discussions over a week (Days 0 to 6). A consistent trend is observed, with visible spikes indicating periods of high activity. These spikes suggest that certain days experience a higher volume of discussions, potentially driven by trending

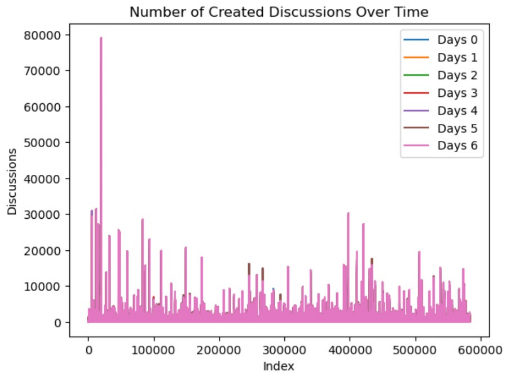

Fig. 2. Number of Created Discussions Over Time

topics or events. Understanding the dynamics of discussions across different days is crucial for strategic planning. This information allows social media managers to time their content releases to coincide with high-activity periods, thereby maximizing user engagement.

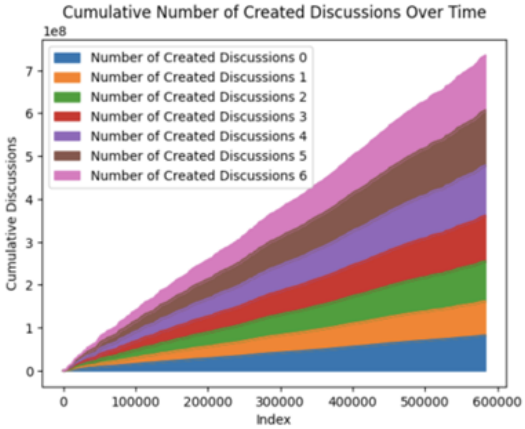

Fig. 3. Cumulative Number of Created Discussions Over Time

The area plot in figure 3 visualizes the cumulative number of created discussions over seven days. A steady increase in cumulative discussions is observed, indicating continuous engagement and interest in the topics discussed. Different colors distinguish the contributions of each day, demonstrating how discussions accumulate over time. This visualization helps in understanding long-term engagement trends on Twitter, identifying periods of high activity and sustained engagement. It provides a clear picture of overall growth and highlights the sustained interest in topics, making it useful for optimizing content strategies to maximize user attention.

The bar plot in figure 4 illustrates the average attention level over a week (Days 0 to 6). An increasing trend is observed, with Day 0 having the lowest attention level and Day 6 the highest. This suggests that discussions or topics gain more attention as the week progresses, potentially due to the accumulation of content and information spread. For social media managers, understanding this pattern is crucial for planning the timing of new topics or content. Introducing significant discussions or promotions earlier in the week can capitalize on the growing attention levels throughout the week, thereby enhancing engagement.

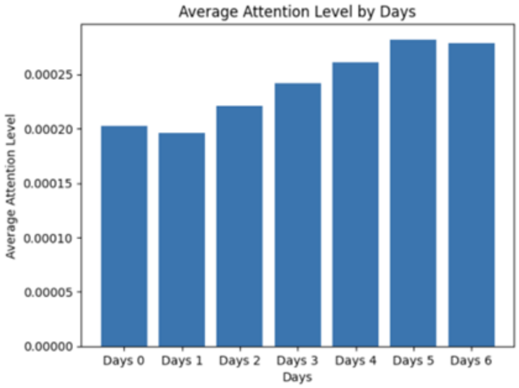

Fig. 4. Average Attention Level by Days

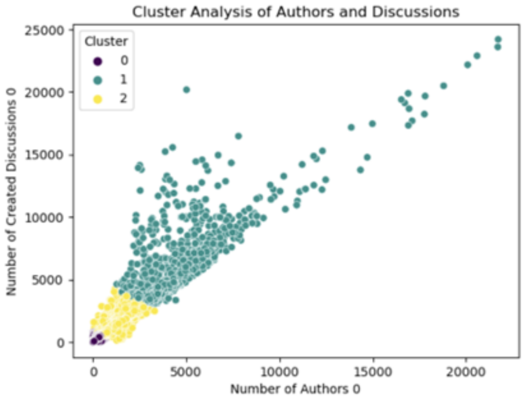

Fig. 5. Cluster Analysis of Authors and Discussions

The scatter plot in figure 5 shows the results of a cluster analysis based on the number of authors and created discussions at Time 0. Three distinct clusters are identified: Cluster 0 (yellow) with low numbers of authors and discussions, Cluster 1 (purple) with moderate numbers, and Cluster 2 (teal) with high numbers. This positive correlation indicates that increased participation leads to more discussions. Understanding these clusters is crucial for strategic planning and engagement optimization. Targeting efforts towards the teal and purple clusters can leverage existing engagement, while different strategies might be necessary to boost participation in the yellow cluster. This cluster analysis provides valuable insights into community dynamics on social media platforms, guiding strategic decision-making.

The pairplot in figure 6 visualizes relationships between key features: Number of Created Discussions 0, Author Increase 0, Attention Level 0, Burstiness Level 0, and Annotation. It shows positive correlations between the Number of Created Discussions 0 and Author Increase 0, Attention Level 0, and Burstiness Level 0, indicating that more authors, higher

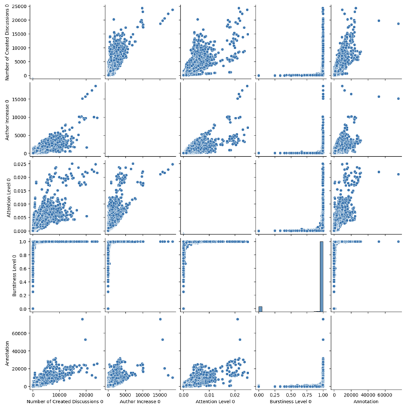

Fig. 6. Pairplot Analysis

attention levels, and increased burstiness correspond to more discussions. Additionally, the Annotation feature, representing overall engagement, also shows positive correlations with these features, highlighting their relevance in predicting user engagement. These findings are crucial for understanding the dynamics of user engagement on social media. The identified relationships between the features and the target variable (Annotation) guide the feature selection process for regression models. The visual patterns observed in the pairplot can assist in feature engineering, such as creating interaction terms or polynomial features, to enhance the model's predictive power. Ensuring a robust regression model involves checking for multicollinearity using the Variance Inflation Factor (VIF) and scaling the features to maintain consistency.

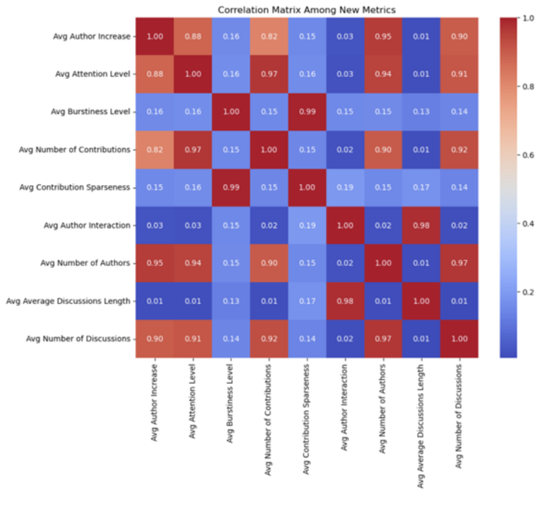

Fig. 7. Correlation Matrix

In figure 7 Metrics such as Avg Author Increase and Avg Attention Level show very high correlations with other metrics, like Avg Number of Contributions (0.97 with Avg Attention Level). This suggests that increases in authors and attention levels are strongly associated with more contributions. Conversely, metrics like Avg Author Interaction have very low correlations with most other metrics, indicating that author interactions might be influenced by different factors not directly related to general activity levels. The high correlation between metrics related to discussion activity (e.g., number of contributions, discussion length) suggests that these factors are often influenced together by the same underlying trends or events.

VI. Feature Engineering

To predict which social media posts would go viral, we started with feature engineering, turning raw data into meaningful variables to boost our model's performance. First, we examined basic engagement metrics like number of created discussions, author increase, and author interactions since these are clear indicators of a post's popularity. But we didn't stop at the basics. We also looked at content features, such as burstiness level, average discussion length, and the number of authors. These elements significantly impact how people interact with a post. By considering all these aspects, we aimed to capture the intricate dynamics of what makes a post go viral.

A. *Model Selection and Training*

Creating our predictive models was both an art and a science. We started by testing various machine learning algorithms to determine which would best predict social media virality. Initially, we used models without applying Principal Component Analysis (PCA). These included K-Nearest Neighbors (KNN), Linear Regression, Stochastic Gradient Descent (SGD) Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Linear Support Vector Machine (SVM), and Kernel SVM.

Each model offered unique insights. KNN helped us understand patterns based on similarity, while Linear Regression revealed linear relationships between features and virality. SGD Regression was efficient with large datasets, allowing rapid iteration. Polynomial Regression captured non-linear relationships, and Ridge and Lasso Regression addressed multicollinearity and feature selection. The SVM models were excellent at finding the best separation between viral and non-viral posts, with Kernel SVM handling non-linear boundaries.

Next, we applied PCA to reduce the dataset's dimensionality, highlighting the most important features. We then retrained the same models—KNN, Linear Regression, SGD Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Linear SVM, and Kernel SVM—using the principal components identified by PCA. This simplification made the models faster and potentially more accurate.

For each model, we fine-tuned hyperparameters to optimize performance. We used cross-validation to train and validate the models, ensuring they didn't overfit and could generalize well to new data.

In summary, we thoroughly explored various machine learning algorithms, both with and without PCA, to find the best method for predicting social media virality. Our goal was to develop robust models capable of making accurate and insightful predictions.

VII. Results and Evaluation

The performance of various regression models was evaluated using three distinct techniques: Bagging, Boosting, and Principal Component Analysis (PCA). The goal was to determine the best-performing model for predicting social media virality. Bagging, or Bootstrap Aggregating, was applied to models like RidgeRegression and LassoRegression to reduce variance and improve robustness. This technique involves training multiple instances of a model on different subsets of the data and averaging their predictions. The results showed that LassoRegression(Bagging) achieved the highest test accuracy both without and with PCA, demonstrating its strong generalization capability. Boosting techniques, including Adaboost and Gradient Boosting, were evaluated for models such as DecisionTreeRegressor and GradientBoostingRegressor. Boosting sequentially trains models to correct the errors of its predecessors, aiming to reduce bias and improve model performance. GradientBoostingRegressor stood out with high training and test accuracies without PCA, indicating its effectiveness in capturing complex patterns in the data.

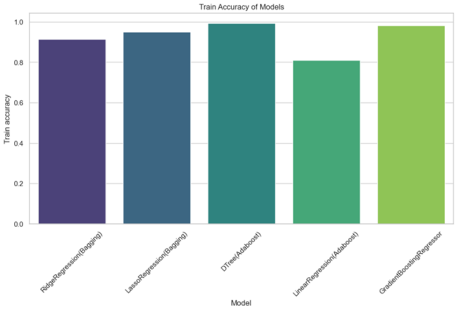

Fig. 8. Train before PCA

The initial comparison of models without PCA as shown in figure 8 shows that GradientBoostingRegressor achieved the highest training accuracy, followed closely by DTree(Adaboost), LassoRegression(Bagging), and RidgeRegression(Bagging). LinearRegression(Adaboost) demonstrated slightly lower training accuracy compared to the other models.

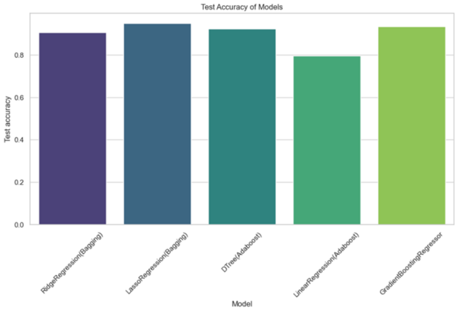

Fig. 9. Test before PCA

Figure 9 shows that in terms of test accuracy, LassoRegression(Bagging) outperformed the other models, indicating its strong ability to generalize effectively to unseen data. Both GradientBoostingRegressor and RidgeRegression(Bagging) maintained high test accuracies, reflecting their robustness. DTree(Adaboost) and LinearRegression(Adaboost) showed comparatively lower test accuracies, suggesting potential overfitting issues or limitations in capturing the data's underlying patterns.

PCA was applied as a dimensionality reduction technique to transform the feature space, making it easier for the models to learn and generalize. The application of PCA improved the performance of most models, particularly PolynomialRegression(PCA) and KNNRegression(PCA), which achieved the highest training accuracies. LassoRegression(PCA) excelled in test accuracy, highlighting PCA's role in enhancing the model's ability to generalize to new data.

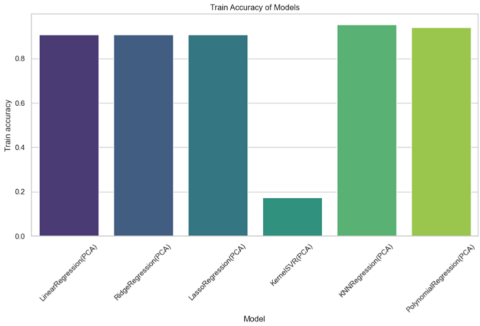

Fig. 10. Train after PCA

When PCA was applied, the training accuracy results were quite consistent as seen in figure 10, with PolynomialRegression(PCA) and KNNRegression(PCA) achieving the highest scores. LinearRegression(PCA), RidgeRegression(PCA), and LassoRegression(PCA) also performed well, whereas KernelSVR(PCA) showed significantly lower training accuracy.

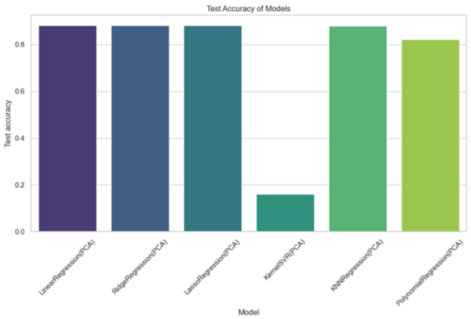

Fig. 11. Test after PCA

Figure 11 highlights that for test accuracy with PCA, LassoRegression(PCA) continued to excel, indicating that PCA helped in enhancing its generalization capabilities. LinearRegression(PCA), RidgeRegression(PCA), and PolynomialRegression(PCA) also exhibited strong test accuracies. KernelSVR(PCA) lagged behind, which may be due to its sensitivity to the transformed feature space or its inherent complexity.

Overall, the application of PCA improved the performance of most models, particularly in terms of test accuracy. LassoRegression, both with and without PCA, emerged as the top performer, demonstrating its effectiveness in predicting social media virality. These findings highlight the importance of dimensionality reduction techniques like PCA in enhancing model performance and generalization. The results underscore the potential of advanced regression techniques in providing accurate and reliable predictions for social media analytics.

VIII. Conclusion

The project sought to utilize advanced machine learning techniques in order to predict the virality of social media posts. The project aimed at deepening our knowledge into engagement dynamics on social media by analyzing vast data and employing different models. Our pioneering method encompassed inventive feature engineering and amalgamation stratagems for multiple models to boost predictions' precision. As a result, we hope to develop a real-time tool that predicts virality while providing insights that can be used to optimize content on social media. Therefore, this project stands out as it fills a gap in literature creating an opportunity for strategic benefits accruing from users who want to analyze their contents with respect to the developments in the field of social media monitoring and analytics.

IX. Future Scope

Refinement of predictive models presents an immediate opportunity, with potential improvements through fine-tuning and optimization or exploration of advanced techniques such as neural networks. Further, the project could benefit from ongoing feature engineering efforts to deepen the understanding of factors influencing social media virality, incorporating additional data sources such as multimedia content and using advanced NLP methods for textual analysis. User personalization offers an exciting avenue for enhancing the relevance of virality predictions, leveraging insights from user behavior and preferences to tailor recommendations. Real-time prediction systems hold promise for applications like immediate feedback on post virality, requiring optimization for streaming data and real-time analytics. At the same time, cross-platform analysis could provide a more comprehensive view of social media dynamics, revealing platform-specific patterns and cross-platform influences. Addressing ethical and privacy considerations is crucial, ensuring user consent, data anonymization, and compliance with regulations such as GDPR. Integration with marketing strategies can optimize ad placements, content scheduling, and influencer collaborations based on predicted virality scores, enhancing overall marketing effectiveness. Conducting longitudinal studies on virality trends over time could reveal shifts in social media engagement patterns, while exploring the interplay between different content genres and cultural influences presents unique challenges and insights into contemporary digital landscapes.

X. Acknowledgement

We express our deepest gratitude to Professor Shih Yu Chang for their invaluable guidance, insightful feedback, and continuous support throughout this project. Their expertise and encouragement were instrumental in the successful completion of our research. We also extend our sincere thanks to the Applied Data Science Department at San Jose State University for providing the necessary resources and facilities. The department's unwavering support and the access to state-of-the-art tools and infrastructure greatly enhanced the quality and scope of our work. We would also like to acknowledge our peers and colleagues for their collaborative spirit and constructive discussions, which contributed to the refinement of our ideas and methodologies. Their feedback and encouragement played a significant role in the advancement of this project.

References

[1] S. Asur and B. A. Huberman, "Predicting the future with social media," *Web Intelligence and Agent Systems*, vol. 10, no. 1, pp. 11–29, 2010.

[2] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[3] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[4] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd International Conference on World Wide Web*, 2014.