Time-Series Topic Analysis of Large-Scale Social Media Data using Two-stage Clustering

Takako Hashimoto
Chiba University of Commerce
The University of Tokyo
Chiba/Tokyo, Japan
takako@cuc.ac.jp

Abstract—**Social media is a highly influential platform for sharing messages, photos, and videos. Understanding public perception through its vast data stream is essential. This study introduces a two-stage clustering method to extract coarse-grained topics from social media text data. First, graph clustering extracts micro-clusters from graphs generated based on the similarity of user posts, with each micro-cluster representing a fine-grained topic. The time series of these micro-clusters are then analyzed in the second stage through time series clustering to reveal coarse-grained topics. In this study, we consider applying this method to Yahoo! Japan News Comments related to the election of two specific candidates in Japan. This is expected to extract people's reactions to the candidates before and after the election.**

Keywords-social media analysis; knowledge discovery; graph mining; two-stage clustering; time series.

I. INTRODUCTION

Social media platforms, such as Yahoo! Japan News, are influential hubs for user interaction through messages and other media, such as photos and videos. Yahoo! Japan News, distributing around 7,500 daily articles and attracting 5 billion monthly visits, features an anonymous comment section where users can post opinions and react to others [1].

With the rapid growth of user-generated content, manually analyzing comments is impractical. To address this, we developed a two-stage clustering method to extract key topics and their temporal patterns from social media data [2]. This approach analyzes public perceptions, tracks opinion shifts, identifies misinformation, and evaluates communication effectiveness. It also bridges public sentiment with policymaking by providing actionable insights.

Unlike traditional topic modeling [3][4], which focuses on classification, our method combines graph clustering [5] of graphs based on the similarity of user posts with temporal analysis, enabling efficient processing of large-scale and sparse data while integrating topic and temporal evaluation.

This paper demonstrates the concept of our two-stage clustering method applying to various topical issues from Yahoo! Japan News Comments. In the first stage, we apply graph clustering to the word co-occurrence graph and extract micro-clusters corresponding to fine-grained topics of comments. We utilize Data Polishing algorithm [5][6] to achieve better scalability for data processing. We extract time series data for each micro cluster by counting the tweets posted within a defined time window. Specifically, we focus on 'burst' events, where a sudden spike in tweet activity corresponds to significant external events (such as election debates or breaking news). By examining these bursts, we can correlate the shifts in public perception with specific political or social occurrences. Finally, we apply time series clustering in the second stage to find the clusters corresponding to coarse-grained topics.

Our method has two key advantages: scalability and the ability to use both textual and temporal information. The method can handle large-scale social media data, making it suitable for long-term analysis. Additionally, by considering temporal patterns, we capture "bursty" activity [7] in the data, reflecting real-world events such as news and natural occurrences, which are essential for understanding underlying topics in social media discussions [8][9].

The rest of the paper is organized as follows. Section II surveys the related work. Our two-stage clustering method for discovering coarse-grained topics is briefly described in Section III. Next, Section IV explains our proposed method with a large-scale Yahoo! Japan News Comments data set regarding the election results of specific candidates in Japan. Finally, we summarize the results in Section VI.

II. RELATED WORK

Recent studies on social media data analysis have used Latent Dirichlet Allocation (LDA) [3] to identify topics in tweets [10][11]. However, LDA has limitations: it assumes documents contain multiple topics and require repeated word instances, making it unsuitable for social media posts that are generally short and low-quality. LDA also needs several hundred iterations, making it inefficient for large datasets, and it ignores temporal information such as timestamps [12][13][14]. Though extensions like X LDA [15] and dynamic LDA [16] address some of these issues, they still face other LDA limitations.

To overcome these problems, we propose a two-stage clustering algorithm using both word and timestamp data. This method applies graph clustering to identify micro-clusters (fine-grained topics) in social media data. Our proposed method, which applies graph clustering to identify fine-grained topics from social media data, extends previous work on topic modeling (e.g., LDA [16]). Unlike LDA, which

 ISBN: 978-1-68558-244-9

requires repeated word occurrences in longer documents, our method is tailored to handle short, sparse data such as tweets. Furthermore, our approach incorporates time series analysis to capture the temporal dynamics often missing in conventional topic modeling techniques. Our clustering algorithm has five key characteristics: quantity, independence, coverage, granularity, and reproducibility. Existing methods like pattern mining [17], community mining [18], and DBscan [19] do not fully meet these criteria, but Data Polishing [5] offers a solution.

Event detection from social media streams is a popular research topic [20][21], with methods focusing on "bursty" events that lead to sharp rises in tweet activity. Our method, however, distinguishes between reactions to breaking news and general social media trends based on the similarity of temporal patterns.

Finally, while many studies focus on predicting social media trends [14][22][23], we develop an automatic method for discovering temporal patterns of collective human attention, helping to distinguish between external shocks and internal effects like word-of-mouth.

III. METHOD FOR DISCOVERING COARSE-GRAINED TOPICS FROM

We introduce a two-stage clustering method [2] to extract coarse-grained topics from social media data (Fig.1). The figure has the following four parts; A: Large-scale social media data (i.e., the tweets and timestamps). B: Graph defined by the similarity between user posts. C: Micro-clusters obtained by graph clustering (first stage clustering). D: Coarse-grained topic obtained by time series clustering (second stage clustering). In the C part, the gray circles represent a micro-cluster. Comments in a micro cluster share a fine-grained topic.

First, we construct a similarity graph of users' posts, where users' posts share similar words. For example, nodes (users' comments) are linked when sharing more than 50% of the words (Fig.1.A–B). Next, using a Data Polishing algorithm, graph clustering is applied to detect micro-clusters, representing fine-grained topics (Fig.1.C). Finally, time-series clustering groups these micro-clusters into coarse-grained topics, revealing how discussions evolve over time (Fig.1.D). This approach enhances scalability and integrates both textual and temporal features, making it suitable for analyzing large-scale comment datasets.

A. Graph Generation

We create the graph from users' posts, an undirected graph where each node represents a tweet/comment and an edge indicates tweet/comment similarity. A pair of tweets are connected if their Jaccard similarity coefficient[24] exceeds the threshold θE.

B. Graph Clustering

We identify fine-grained topics by clustering the users' posts graph to find micro-clusters (i.e., dense subgraphs) (Fig. 2). All the posts in a micro-cluster are expected to have a similar meaning.

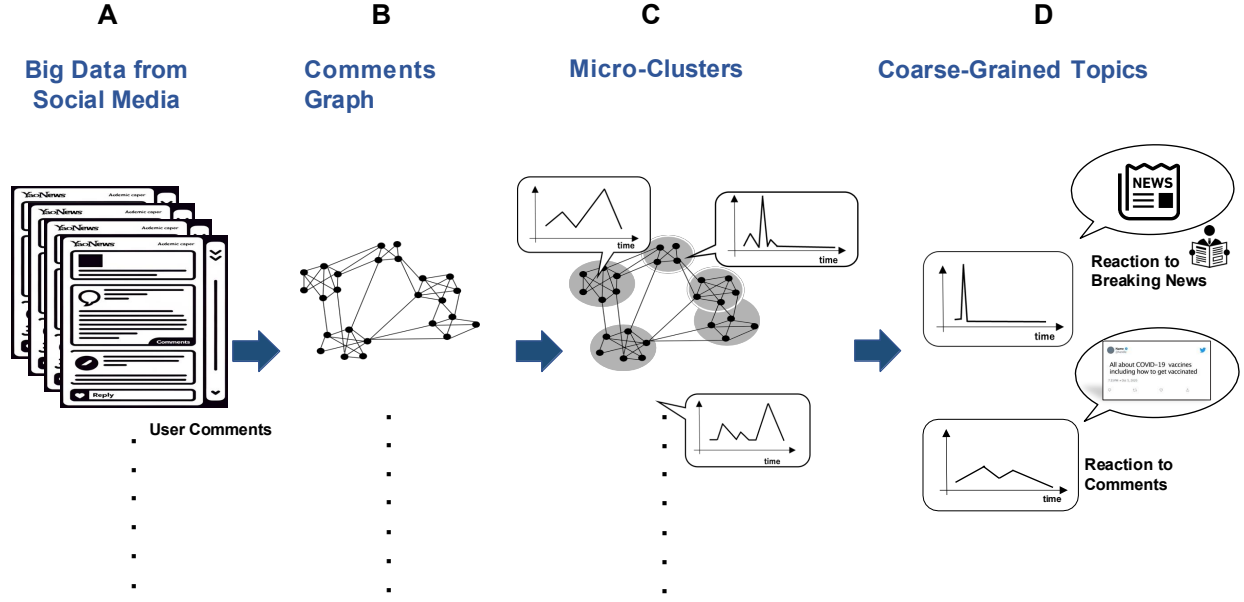


Figure 1. Proposed method (Two-stage clustering method) for discovering coarse-grained topics of public perceptions from social media data.

To identify fine-grained topics, we perform graph clustering to find micro-clusters (dense subgraphs) (Fig. 2). Users' posts within the same micro-cluster are highly similar. An edge is added between nodes (u, v) if the Jaccard similarity of their neighbor sets (N[u], N[v]) exceeds θ_{DP}. Non-satisfying edges are removed. Data Polishing iterates this process until the graph is divided into cliques, which are then identified as topics using maximal clique enumeration with MACE [25].

C. Time Series Clustering

While Data Polishing identifies topics, it often generates too many clusters for existing analysis. To reduce the number of clusters and improve interpretability, we use the users' posts timestamps (Fig.3). We divide time into windows and count the users' posts in each micro-cluster within those windows. Then, we apply K-Spectral Centroid (K-SC) [26] clustering to group micro-clusters with similar temporal patterns. K-SC is chosen for its robustness in capturing clusters using a similarity metric invariant to scaling and shifting, making it efficient for large datasets.

$$F = \sum_{j=1}^{K} \sum_{x_i \in C_j} \hat{d}(x_i, \mu_j)$$

where K is the number of clusters, xi is the *i-th* time series, and Cj is a set that represents the member of the *j-th* cluster. The K-SC's distance metric $\hat{d}(x, y)$ between the two time series (x and y) is defined as follows:

$$\hat{d}(x, y) = \min_{\alpha, q} \frac{\|x - \alpha y_q\|}{\|x\|}$$

where y_q is the result of shifting time series y by q time units, and $\| \ \|$ represent the l_2 norm.

IV. Discovering People's Perceptions about Specific Candidates before- or after- Elections

In this paper, we trying to apply the proposed method (Fig.1) to large-scale Yahoo! Japan News Comments to uncover public perceptions on coarse-grained topics. Yahoo! Japan News Comments, a widely used social media platform in Japan similar to X, often becomes a hot topic influencing public opinion on current issues.

A. Data

The dataset contains comments from Yahoo! Japan News mentioning the names of two candidates in local prefecture governor elections: hereafter CandidateA and CandidateB. For example, we suppose that CandidateA, initially a less known candidate, gained popularity during the election and finished as the runner-up. However, after the election, he faced significant criticism due to harassment allegations.

We can also suppose that CandidateB, the sitting governor of the Prefecture, faced harassment accusations, leading to his resignation. Despite the controversy, he gained support as the campaign progressed.

For these two candidates, Yahoo! Japan News that have one's name in the title will be the target data.

B. Adapting Two-stage Clustering

First, we segment each comment using the Japanese morphological analyzer MeCab [27] and remove stop words such as "kore" (this), "sore" (it), and "suru" (do). Next, we generate a user's comments text graph where each node represents a comment. Comments are connected if the Jaccard similarity coefficient of their word sets exceeds the threshold θE = 0.3, meaning the comments share more than 50% of words in common. In our two-stage clustering method, we set the threshold θ_{DP} to 0.2, and our experiments show that the results are robust to changes in this parameter.

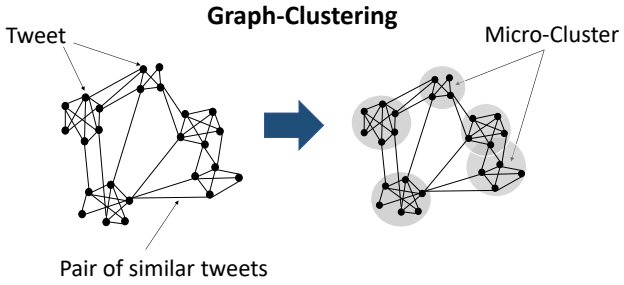


Figure 2. First stage clustering: Graph clustering.

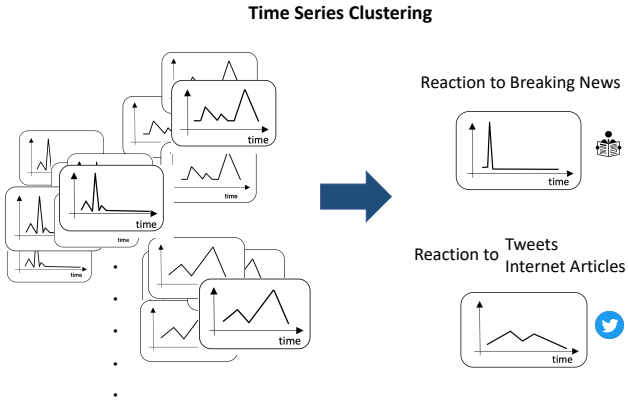


Figure 3. Second stage clustering: Time series clustering.

After identifying micro-clusters from the text graph, we extract coarse-grained topics by clustering the top 1,000 largest micro-clusters using time-series clustering (Sec. III C). The TimeSeriesKMeans algorithm was used, with the number of clusters K is set to 15, utilizing the Python package tslearn for time series analysis. We identified the cluster topics using ChatGPT.

C. Analysis of Two-stage Clustering Result

Fig.4 and Fig.5 present the two-stage clustering result examples (results of 2 clusters out of 15 clusters, respectively, as examples) for comments before and after the election of CandidateA and CandidateB, respectively. Fig.4-a and Fig. 5-a represent data before the election, while Fig.4-b and Fig. 5-b correspond to data after the election:

- Left (Word Cloud): Displays the most frequently used words, representing key themes in discussions
- Center (Time-Series Graph): Shows the temporal distribution of comments within each cluster, capturing shifts in public discourse.
- Right (Contents): Explains the cluster contents briefly.

This layout effectively visualizes how public sentiment and key discussion points changed before and after the election, helping to track emerging concerns and reactions. For example, we may be able to consider the following perspectives from people.

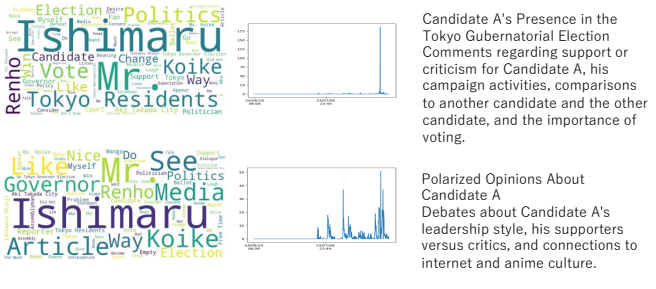
1) Before-election discussion on CandidateA: We could observe that preelection topics predominantly revolved around CandidateA's campaign strategies, political stance, qualifications, and media coverage. There was significant engagement with CandidateA's policies, leadership style, and future expectations, with public discourse centered on comparisons to other candidates and the feasibility of campaign promises. Media influence was also a recurring theme, though largely in the context of how it shaped CandidateA's perception among the public.

2) After-election discussion on CandidateA: We also obsereved that after election, the focus shifted from campaign strategies to critical evaluations of CandidateA's actions and credibility. The discussions became more emotionally charged, reflecting heightened public reactions to controversies, such as harassment allegations. Topics expanded to include broader concerns about political trustworthiness, media fairness, and societal trends. Discussions on gender biases and the humanity of politicians also emerged, reflecting a deeper exploration of societal and political issues. Additionally, the after-election discourse revealed increasing political distrust and dissatisfaction with the system.

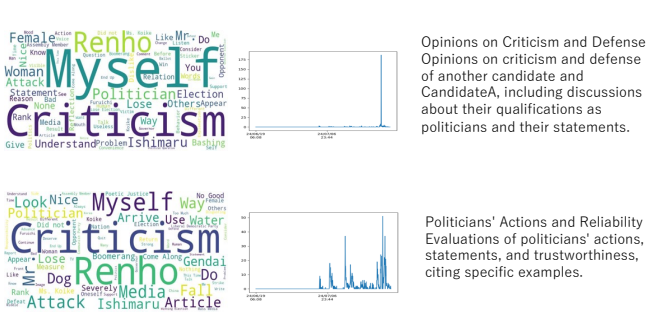
3) Topic change from before-election to after-election for Candidate A: We can consider that the transition from before-election to after-election discourse reflects a shift from forward-looking campaign analysis to retrospective evaluations of political credibility, public trust, and societal implications, with a heightened emphasis on emotional and systemic critiques.

4) Before-election discussion on CandidateB: Prior to the election, the discourse primarily centered on CandidateBs campaign activities, leadership, political stance, and media coverage. Public opinion was shaped by discussions on CandidateB's qualifications as a governor, his policies, and his comparison to other political figures.

Media influence played a significant role in forming public perceptions, with many comments reflecting concerns

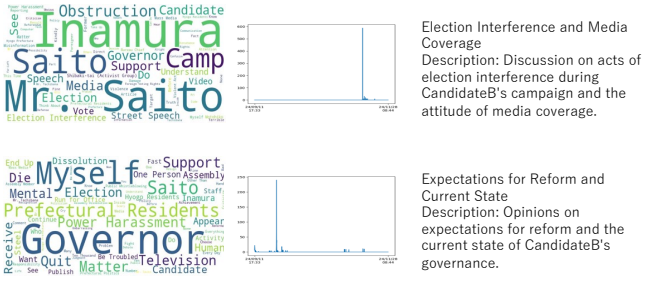


b. Cluster Examples of CandidateA - Before election

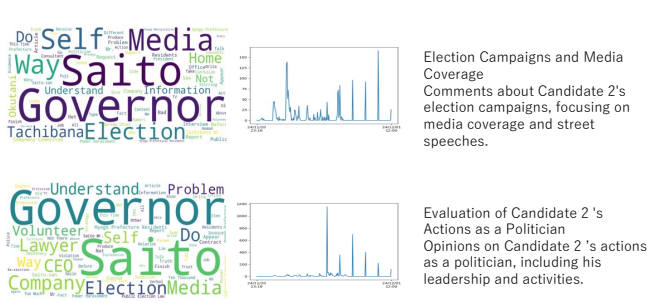


b. Cluster Examples of CandidateA - After election

Figure 4. CandidateA's Coarse-grained topic examples obtained by two-stage clustering.



a. Cluster Examples of CandidateB - Before election



b. Cluster Examples of CandidateB - After election

Figure 5. CandidateB's Coarse-grained topic examples obtained by two-stage clustering.

about the bias in reporting and CandidateBs reception in the media.

5) After-election discussion on CandidateB: After the election, the focus transitioned from before-election campaign strategies to more critical evaluations of CandidateB's actions and trustworthiness as a politician. The discourse became more polarized, with discussions increasingly reflecting public reactions to election results, controversies such as harassment allegations, and concern over political integrity. The after-election discussions also introduced themes of media fairness, misinformation, and political expectations, highlighting a growing dissatisfaction with politicians and the political system. There was a deeper exploration of topics such as trust in politicians, gender biases, and the social impact of political decisions.

6) Topic change from before-election to after-election for Candidate B: The shift from before-election to after-election discourse can be considered to illustrate a movement from campaign-focused discussions to more intense evaluations of political credibility, media influence, and societal concerns. The after-election phase may be marked by a heightened emphasis on trust issues, emotional responses, and systemic critiques.

This analysis is intended to indicate a direction, and the current data does not go far enough to capture this trend accurately. However, it is thought that it will become possible to grasp such trends by refining the data.

V. Conclusion

This study demonstrated the effectiveness of our two-stage clustering method in analyzing large-scale social media data, particularly in tracking public perceptions before and after elections. By applying the method to Yahoo! Japan News Comments, we identified key topics and their temporal evolution, uncovering shifts in political sentiment, media influence, and public trust.

Our findings highlight a notable shift in discussions moving from campaign strategies and candidate qualifications (before the election) to critical evaluations, controversies, and trust issues (after the election). Furthermore, increasing political polarization and concerns over media bias were observed, reflecting broader societal trends. Despite its effectiveness, limitations remain. Noisy data and evolving linguistic patterns could affect the method's performance. Future research will focus on enhancing robustness against noise and improving temporal clustering techniques. Beyond political analysis, this approach can also be adapted to analyze social media discussions around other societal issues, such as public health, environmental concerns, and economic policies. By applying this method to a variety of topics, we can gain deeper insights into how public opinion evolves in response to diverse challenges, enabling better-informed decision-making for policymakers, media analysts, and researchers.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23K21728.

References

[1] Media personel blog, meditsubu (in japanese), https://media-radar.jp/contents/meditsubu/columns5-yahoo-news-pv/, Jan 2024, 2025.03.03.

[2] T. Hashimoto, T. Uno, Y. Takedomi, D. Shepard, M. Toyoda, N. Yoshinaga, M. Kitsuregawa, and R. Kobayashi, "Two-stage clustering method for discovering people's perceptions: A case study of the covid-19 vaccine from twitter," 2021 IEEE International Conference on Big Data (Big Data), pp. 614–621, 2021.

[3] D. M. Blei, A. Y Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, 3: pp. 993–1022, 2003.

[4] D. M. Blei, "Probabilistic topic models," Communications of the ACM, 55(4): pp. 77–84, 2012.

[5] T. Uno, H. Maegawa, T. Nakahara, Y. Hamuro, R. Yoshinaka, and M. Tatsuta, "Micro-clustering by data polishing," 2017 IEEE International Conference on Big Data (Big Data), pp. 1012–1018. IEEE, 2017.

[6] T. Hashimoto, D. L. Shepard, T. Kuboyama, K. Shin, R. Kobayashi, and T. Uno, "Analyzing temporal patterns of topic diversity using graph clustering," The Journal of Supercomputing, 77(5): pp. 4375–4388, 2021.

[7] J. Kleinberg, "Bursty and hierarchical structure in streams," Data mining and knowledge discovery, 7(4): pp. 373–397, 2003.

[8] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in twitter," Proceedings of the 21st international conference on World Wide Web, pp. 251–260, 2012.

[9] R. Kobayashi, P. Gildersleve, T. Uno, and R. Lambiotte, "Modeling collective anticipation and response on wikipedia," Proceedings of the International AAAI Conference on Web and Social Media, 15(1): pp. 315–326, 2021.

[10] R. J. Medford, S. N Saleh, A. Sumarsono, T. M Perl, and C. U. Lehmann, "An "infodemic": leveraging high-volume twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," Open Forum Infectious Diseases, 7(7):ofaa258, 2020.

[11] J. C. Lyu, E. L. Han, and G. K Luli, "Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis," Journal of medical Internet research, 23(6):e24435, 2021.

[12] J. Hurlock and M. L Wilson, "Searching twitter: Separating the tweet from the chaff," Fifth International AAAI Conference on Weblogs and Social Media, pp. 161–168, 2011.

[13] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," Proceedings of the 20th international conference on World wide web, pp. 675–684, 2011.

[14] T. Murayama, S. Wakamiya, E. Aramaki, and R. Kobayashi, "Modeling the spread of fake news on twitter," Plos one, 16(4):e0250419, 2021.

[15] W. Xin Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," European conference on information retrieval, pp. 338–349. Springer, 2011.

[16] D. M. Blei and J. D. Lafferty, "Dynamic topic models," Proceedings of the 23rd international conference on Machine learning, pp. 113–120, 2006.

[17] T. Uno, et al., "Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets," Fimi, volume 126, 2004.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cybercommunities. Computer networks, 31(11-16): pp. 1481–1493, 1999.

[19] M. Ester, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," kdd, volume 96, pp. 226–231, 1996.

[20] M. Hasan, M. A. Orgun, and R. Schwitter, “A survey on real-time event detection from the twitter data stream,” Journal of Information Science, 44(4): pp. 443–463, 2018.

[21] C. Comito, A. Forestiero, and C. Pizzuti, “Bursty event detection in twitter streams,” ACM Transactions on Knowledge Discovery from Data (TKDD), 13(4): pp. 1–28, 2019.

[22] R. Kobayashi and R. Lambiottem “Tideh: Time-dependent Hawkes process for predicting retweet dynamics,” Tenth International AAAI Conference on Web and Social Media, 2016.

[23] J. Proskurnia, P. Grabowicz, R. Kobayashi, C. Castillo, P. C. Mauroux, and K. Aberer, “Predicting the success of online petitions leveraging multidimensional time-series,” Proceedings of the 26th International Conference on World Wide Web, pp. 755–764, 2017.

[24] P. Jaccard, “The distribution of the flora in the alpine zone,” 1. New phytologist, 11(2): pp. 37–50, 1912.

[25] K. Makino and T. Uno, “New algorithms for enumerating all maximal cliques,” Scandinavian workshop on algorithm theory, pp. 260–272. Springer, 2004.

[26] J. Yang and J. Leskovec, “Patterns of temporal variation in online media,” Proceedings of the fourth ACM international conference on Web Search and Data Mining, pp. 177–186, 2011.

[27] T. Kudo, “Mecab: Yet another part-of-speech and morphological analyzer,” https://sourceforge.net/projects/mecab/, 2013, 2025.03.03