Proc. 2019 IEEE International Conference on Computer Communication and Neworks,
Valencia, Spain, July 29 - August 1, pp. 1-9.

A Reaction-Based Approach to Information Cascade Analysis

James Flamino
Department of Physics, Applied Physics, and Astrophysics
Rensselaer Polytechnic Institute
Troy NY, USA
flamij@rpi.edu

Boleslaw K. Szymanski
Department of Computer Science
Rensselaer Polytechnic Institute
Troy NY, USA
szymab@rpi.edu

***Abstract*—Online social media provides massive open-ended platforms for users of a wide variety of backgrounds, interests, and beliefs to interact and debate, facilitating countless information cascades across many subjects. With numerous unique voices being lent to the ever-growing information stream, it is essential to consider the question: how do the many types of conversations within an information cascade characterize the process as a whole? In this paper we analyze the underlying features of the dynamics of communication, and use those features to explain the inherent properties of the encompassing information cascade. Utilizing "microscopic" trends to describe "macroscopic" phenomena, we set a paradigm for analyzing information dissemination through the individual user interactions that sprout from a source topic, instead of trying to interpret the emergent patterns themselves. This paradigm yields a set of unique tools for a myriad of application in the field of information cascade analysis: from topic classification of sources to time-series forecasting. We use these tools in a 88-million-row dataset for Reddit to show their conceptual effectiveness and accuracy when compared to the ground truth.**

***Index Terms*—information cascade, response features, topic classification, time-series forecasting, conversational dynamics, semantic analysis**

I. Introduction

How can we understand the nature of a topic? To answer this query, let us consider a simpler question: How can we guess the genre of an unknown movie while actively watching it in a theatre? One way would be to examine the title and character dialogue. Another way would be to analyze the visuals and special effects. But a more unconventional approach might be to turn around and look to the audience for the answer. If we do not want to use dialogue or visuals, we can use the audience's reactions. For example, if the audience is laughing frequently, its likely to be a comedy. But if they are mostly crying out in fear, the film is probably a horror movie. And while there are outliers, the process described is reliable enough given that an audience's emotional reaction is inherently tied to the nature of a movie's genre. After all, the director of a horror film needs the audience to react in fear, otherwise their film will most likely fail. And just as these emotional reactions from an audience fundamentally describe the associated movie, topics in online social media have their own set of unique "user reactions" that can be extracted and used to characterize the unique properties of the discussed subject matter without any help from text analysis. And while natural language processing (NLP) in itself has set an unprecedented paradigm for understanding human interactions and the nature of information cascades, in this paper we will show that the alternative approach of using non-NLP, reaction-based analysis can provide significant insight into the understanding of topics in online social media.

In this paper we will start our description of this novel approach in Section 3 by describing the response features we choose to use, their mathematical quantification, the methods we use to employ them, and a visualization of their patterns. We then validate these features in order to show their consistency and resilience in real-world datasets in Section 4. In Section 5 we introduce time-dependency to further describe the robustness of this approach. Then in Section 6 we utilize the response features in all forms for a couple of applications in order to demonstrate their capabilities, while testing accuracy and effectiveness. We conclude our work in Section 8, where we review the implications of a response-based approach to information cascade analysis.

II. Datasets

Given our focus on conversational dynamics, we decided to extract a dataset from an online social media platform that encourages in-depth discussions ofna wide variety of topics and subjects. The format that we found to best fit this description was the group of online platforms called *forums*. A forum is a network of registered users in which any user can freely submit posts about certain topics (generally under some related category). This post triggers responses to the post material. In turn these responses trigger more responses, resulting in a cascade of information passed between unique users. A majority of these cascades will be short-lived, and are quickly superseded by more recent topics, but what conversations do occur due to that post follow the theme established by the source post, characterizing a majority of the interactions that occur within a cascade.

One of the most well-known forum-like social media platforms is Reddit. In Reddit the posts are clustered by Subreddit, which generally encompasses a defining theme (e.g. the Subreddit $r/politics$ is comprised of discussions about U.S. politics). Within a Subreddit, a user can create a submission

978-1-7281-1856-7/19/$31.00 ©2019 IEEE

pertaining to the Subreddit's genre. The submission then becomes available to all other users. Users can vote on the quality of the submission and start discussions in the comment section of the submission. The more provocative the subject of a submission, the greater the response, ultimately increasing the activity of the submission and the vote count, which inevitably increases the exposure of the submission, evoking additional responses. In essence, Reddit is a platform that rewards posts that elicit conversational cascades; exactly what we are looking for.

We use the data provided from [1], which offers up a Reddit dataset that covers $5,692$ Subreddits, $88M$ submissions, and $887.5M$ comments over a time range of 2006 to 2014. The comments in this dataset are formatted as a comment tree extension, accentuating the natural branching conversations that sprout from the source submission. We reformatted this into event sequence style, where the identifiers id, $root\_id$, and $parent\_id$ illustrates the event's position in a cascade. To clarify, the $root\_id$ indicates the id of the source submission, while the $parent\_id$ indicates the id of the event that the posting user is responding directly to. This event sequence was uploaded into a MySQL database, with the $88M$ submissions being supplemented with their respective titles, text bodies, and scraped headlines from any linked URLs using Reddit's Official API.

III. Characterizing conversational dynamics

Our objective is to use the dynamics of conversations to characterize and understand the overarching topic. With respect to Reddit, this means we want to use the comment tree to identify the theme of the submission without any text mining from the comment tree and even from the submission itself. But what kind of features can we extract that are indicative enough? Our answer is this: we will capture features using the innate bias that a majority of users will unconsciously exhibit given a specific topic. If a user is biased in a certain way, they will react differently in more than just text. A good example of the application of this idea can be found in [2], where the authors show that users have stable, consistent reactions associated with a given topic, called a "Social Genotype Model" (specifically within a Twitter dataset). While proving the stability of this model, the authors quantified non-semantic features to train them in a hashtag classifier. These features pertained to an individual user's use of a hashtag through retweets, and included measures such as time between a user's exposure to a tweet given a hasthag and their first retweet of said hashtag. The results for the paper demonstrated consistent topic-dependent behavior and proved the underlying point previously mentioned: users as a whole will always respond differently depending on the subject matter. Given this understanding, we can begin to confidently consider measures that quantify user interactions.

A. Definitions

Given Reddit's emphasis on forum-like interactions, the emergent online social media network $G(U, E)$ is not formed in a typical fashion, where edges connect unique users like $e = (u_i, u_j), e \in E$. And while user's can "follow" other users in Reddit, the more apparent link to content is through the subscription to Subreddits. Once subscribed, a user becomes part of a collective of fellow subscribers that are all updated when any other subscribed user post a submission to that Subreddit. This leads us to assumption 1.

Symbol	Definition
R	Some Subreddit within Reddit
U	All users subscribed to R
s	A submission within R, represented as a set of users that responded to the submission, $s = \{u\}, s \subseteq U$
$T(N, E)$	A tree network representing the structure of hierarchically linked comments within a submission
n	A user-generated comment within T, $n \in N$
n_0	The head node. Technically, this is the submission text submitted to the subreddit that triggers the comment cascade.
e	A directed edge within T, representing the direction of information flowing between comments (from respondee to responder), $e \in E$
B	The set of branches found in T. $B_k = \{n\}, B_k \subseteq N$ where the k^{th} branch contains some subset of linked comments (including n_0) generated for a submission.

Table 1: Symbols and definitions

***Assumption** 1:* Reddit can be considered a set of Subreddit tags R. Each Subreddit has a unique set of associated users U. All users in set U are connected with an edge representing information flow. This system forms an isolated, *undirected complete graph*

Naturally the network of Reddit becomes more complicated once you consider users subscribe to multiple Subreddits and follow other users, but when considering the graph topology within the scope of a single Subreddit, the assumption makes sense. And given this approach, we can intuitively state that submissions within a Subreddit are also generally detached from each other, with each information cascade eliciting responses from different subsets of users at different rates. The graphical topologies of these submissions are fundamentally different, however, as responses to a source node tend to stack recursively. This leads us to assumption 2.

***Assumption** 2:* Each submission that occurs within a Subreddit is an independent information cascade, representing a unique set of users $s = \{u\}, s \subseteq U$ interacting for a short period of time, with the resultant event sequence following the graph topology of a *directed tree network*.

We now represent the event sequence generated by the users of some submission s as $T(N, E)$ where each $n \in N$ represents a message somewhere in the event sequence triggered by said submission. A connecting edge is defined as

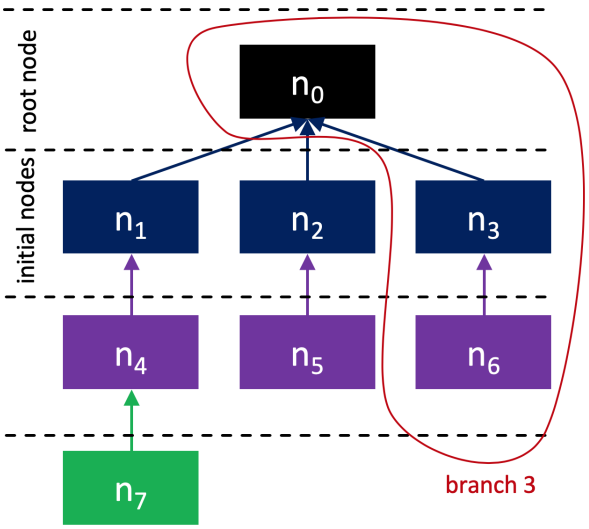

Figure 1: Representation of a directed tree network T.

$e = (n_i, n_j), e \in E$ where n_j is the responder's message and n_i is the respondee's message. It is important to note that each message n does have a uniquely associated user u. While each node n is unique, a user has free reign to generate as many nodes as they please, resulting in user degeneracy in T. We illustrate this terminology and the architecture of T in Figure 1. As seen in Figure 1, we refer to n_0 as the *root node*, and all nodes that directly link to the root node are referred to as *initial nodes*. All other nodes are simply referred to as *response nodes*. We emphasize that we can extract a subset of directly correlated messages called branches. Branches are represented by $B = \{B_1, B_2, ..., B_m\}$ where m is equal to the total number of initial nodes. We can use some k^{th} branch B_k where $B_k \subseteq N$ to evaluate the miniature response cascade that is triggered by the k^{th} initial node. We also show that for every branch, we include the root node as an element in the set. This is because for all intents and purposes, the root node is the submission's contents itself, thus we must include this node to ensure our branch-specific features capture that initial reaction from users that directly respond to the submission. Table 1 summarizes the mentioned symbols. Now that we have established our definitions, we can begin constructing our response features from the empirical data.

B. Approach

For analysis, we split our features into two sets: individual features and aggregate features. For the individual features, the real values must describe the nature of user interactions within a single branch $B_i \in B$. To generate these values, we design features that capture innate behavior exhibited in both scale and time without recursively analyzing each involved user or message. We shall refer to these as our *individual features*.

1) *Depth*: The path length between the root node and the farthest response node in a branch. Using Dijkstra's Algorithm (DA)

$$\text{DEP}(B_i) = \max_{n \in B_i, n \neq n_0} (\text{DA}(n_0, n))$$

2) *Magnitude*: The maximum in-degree centrality in a branch. Given the adjacency matrix A for T then

$$\text{MAG}(B_i) = \max_{n \in B_i} (\sum_k a_{k,n}), a_{k,n} \in A$$

3) *Engagement*: The total number of users involved in a branch where n_u is the set of messages generated by user u.

$$\text{ENG}(B_i) = |\{u \mid |n_u \cap B_i| > 0\}|$$

4) *Longevity*: The absolute amount of time between the creation of an initial node and the latest response node.

$$\text{LNG}(B_i) = \max_{n \in B_i}(t(n)) - \min_{n \in B_i, n \neq n_0}(t(n))$$

In addition to these 4 features, we introduce two additional comprehensive features for a total of 6. We use these comprehensive features to evaluate a submission's aggregate conversational dynamic within T independent of B. The methods are extracted from [3], in which the authors evaluate human communication in temporal networks using entropy and introduce entropy-based measures for human communication. In particular, we chose to consider both the first and second order entropy measures that were introduced. These two methods produce distinct numerical representations that will portray aggregate comment interactions succinctly. Thus we refer to these measures as our *aggregate features*.

5) *First order entropy*: The probability $p_1(u)$ of some user u within s generating a message n for T.

$$\text{S}_1 = -\sum_{u \in s} p_1(u) ln(p_1(u))$$

6) *Second order entropy*: The probability $p_2(e_{ij})$ of a unique edge $e_{ij} = (n_i, n_j)$ being formed between two messages by two specific users u_i and u_j within T.

$$\text{S}_2 = -\sum_{e_{ij} \in E} p_2(e_{ij}) ln(p_2(e_{ij}))$$

We now have an appropriate set of features to describe the response dynamic within an information cascade. Employing the individual measures for some submission s yields a matrix M_s of dimensions $m \times l$ where $m = |B|$ and $l = 4$ for the 4 individual features defined above.

$$M_s^T = \begin{bmatrix} \text{DEP}(B_1) & \text{DEP}(B_2) & \dots & \text{DEP}(B_m) \\ \text{MAG}(B_1) & \text{MAG}(B_2) & \dots & \text{MAG}(B_m) \\ \text{ENG}(B_1) & \text{ENG}(B_2) & \dots & \text{ENG}(B_m) \\ \text{LNG}(B_1) & \text{LNG}(B_2) & \dots & \text{LNG}(B_m) \end{bmatrix}$$

Matrix M_s describes our individual features for all branches of s. Combined with our aggregate features, we have a compact yet detailed portrayal of a submission's entire conversational dynamic.

IV. Validating response features

With the methods for conversational characterization defined, a necessary next step is to validate the actual effectiveness of the selected features in the scope of our dataset. How accurately do these features depict a Reddit submission? Are the features capable of automatically extracting and representing underlying characteristics of submissions that can be empirically confirmed?

A. Genre classification

The first question can be answered through Subreddit classification. Given a few distinct Subreddits to act as labels, can we use the extracted response features from a subset of submissions to train a classifier to distinguish between these labels? To run this test we consider the Subreddits $r/politics$, $r/gaming$, $r/soccer$, and $r/atheism$. We curated a set of 1000 submissions from each Subreddit, only selecting submissions whose total comment counts ranged from $1000 - 3000$ comments. Then for each submission we calculate M_s and the aggregate features. We condense M_s by finding the maximum depth (the height of the tree), average breadth, average engagement, and average longevity for each submission. We map the submission's measures into $\mathbb{R}^6$ feature space by taking these values and appending the aggregate features to the set. We can represent this process as

$$\begin{aligned} f &= < \max_B(DEP), \operatorname{avg}_B(MAG), \operatorname{avg}_B(ENG), \\ &\qquad \operatorname{avg}_B(LNG), S_1, S_2 > \\ &= < f_1, f_2, f_3, f_4, f_5, f_6 > \end{aligned}$$

Where vector f is our feature vector. Pairing f with its associated label R_i where $R_i \in R$, we then split the assembled dataset as $70\%/30\%$ training/testing subsets and train a Support Vector Machine Classifier (SVM) [4]. For each element in the test subset the SVM attempts to correctly label R given f for s. A positive match results in a score of 1 for s, and a negative match results in a 0. We present the average score for the entire test subset as the classifier score for the SVM. Table 2 shows the classifier scores between several label sets.

R	Classifier Score
politics,gaming,soccer	0.86
politics,gaming	0.99
politics,soccer	0.94
gaming,soccer	0.89
politics,atheism	0.78

Table 2: Classifier scores for combinations of Subreddit labels

Note that the average score is 0.89. This indicates that the patterns of conversation are distinctly different between Subreddits due to their themes. In general people will not talk about games in the same ways they will talk about religion. This innate difference in conversation style contributes enough to the overall structure that an SVM can accurate classify a submission using only the six response features. In addition to this observation, the scores in Table 2 can also be seen as a high-level quantification of the similarities and differences between themes. For example, while the classification score between $r/politics$ and $r/gaming$ is 0.99, the score between $r/politics$ and $r/atheism$ is 0.78. Intuitively, the score difference makes sense. While intense debates are not unknown to the world of gaming, the depth and intensity of debates that sprout up from every aspect of politics provides a clear dichotomy in conversational structure. However, the Atheism Subreddit also provides an environment for fierce debate due to its religious connotations which might be consistent enough to be more similar to politics.

B. Topic Clustering

Although classification is a fairly robust means to test the effectiveness of a set of feature labels, there are many other ways to evaluate the features. In particular, these features are meant to represent the patterns of user response within a submission. Knowing this, another effective form of testing might be to use the features to attempt to find underlying patterns that differentiate sub-clusters of submissions. If this is possible, it would indicate that the response features themselves are rich and informative enough to extract previously undetected trends.

To test for potential sub-clusters, we evaluate a set of submissions under a single R_i. Using K-means clustering paired with the Silhouette Score [5] to determine the K, we can find distinct clusters among the submissions of the same label. We define these clusters as $C = \{C_1, C_2, ..., C_K\}$ where K is the cluster number given the highest average Silhouette Score. Figure 2 presents the results of this exercise for 1000 submissions within the $r/hockey$ Subreddit, with the clustered submission projected into 2D Euclidean space using Principle Component Analysis (PCA). We reduce dimensions after K-means for the sake of visualization of the clusters to illustrate the distinct differences between groups due to response features.

Keyword	Cluster 1 Freq.	Cluster 2 Freq.
game thread	204	1
playoff	76	1
series	35	22
friday	1	31
trash talk	1	32

Table 3: Comparison of keyword frequency between the two identified topic clusters in $r/hockey$

As seen in Figure 2, there are two apparent clusters, delineating a difference in conversational structure within the Subreddit itself. And as seen in Table 3, keyword extraction of the submissions involved in each cluster reveals that there is in fact an underlying theme that separates the two clusters. Cluster 1 contains submissions pertaining to "Game Thread" discussions, while Cluster 2 is primarily related to trash talking and memes. Intuitively we know there is explicit difference

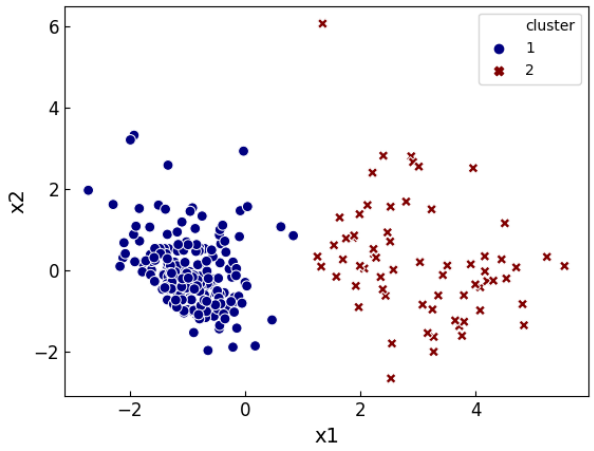

Figure 2: PCA of response features for 1000 submissions from $r/hockey$, clustered by K-means (K=2).

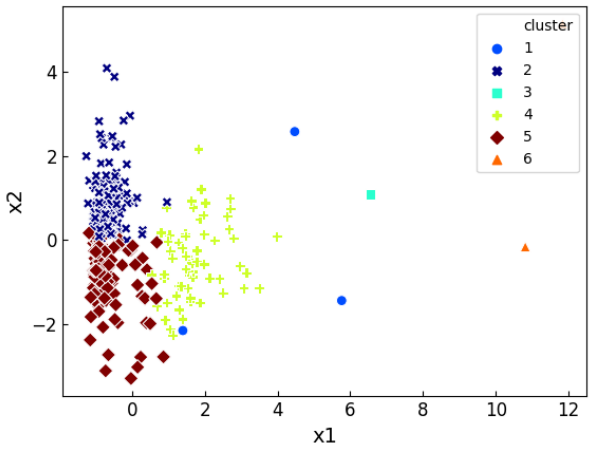

Figure 3: PCA of response features for 1000 submissions from $r/soccer$, clustered by K-means (K=6).

between the two in how they are carried out in conversation, and this understanding is captured by K-means. By extension, this system of submission clustering can be considered a form of non-semantic topic modeling, as it identifies stable topics exclusively through conversational dynamics.

In addition to this, clustering response features can also highlight underlying differences between similar topics. We will consider an example by clustering 1000 submissions from $r/soccer$. As seen Figure 3 there are 6 identified clusters, with 3 of them containing the majority of the submissions. By analyzing the text of the involved submissions, we can identify the textual differences in topic between these primary clusters. While C_4 contains only general discussions, C_2 and C_5 consist exclusively of "Match Thread" submissions. And aside from the fact that C_2 and C_5 do contain entirely different sets of involved soccer teams in their submission text, there is no other clearly apparent textual disparity between these "Match Thread" clusters. But there is an definite underlying difference. Supplemental analysis of the comment structure reveal that submissions in C_2 are often subject to more extensive debate, as opposed to submissions in C_5, which tend to only have surface-level comments. Thus while there was little text-based dissimilarities between the two clusters, the response features were still able to identify additional latent differences in the similar topics due to user reactions. And although this auxiliary information is not necessarily useful for a purely keyword-based approach, it could be quite beneficial for more in-depth analyses of events by accentuating contrasts in audeince feedback within even single topics.

V. Time-dependent response features

As seen in the previous section, response features describe not only overall response trends between themes, but also model the more subtle differences within themes. But before we consider applications using these static features, we can explore an even more dynamic representation of user interactions: time-dependent response features. While measuring the static network of a complete information cascade provides a rich set of features, evaluating the information as it evolves allows us to capture these same features as they change with the cascade. This approach of extracting time-dependent response features can potentially yield an even more in-depth understanding of the cascade.

A. Definition

Instead of analyzing a static graph tree network for a submission, we now look at the network as a temporal graph $T_t(N_t, E_t)$ where T_t evolves with time t. t is defined within the time range $t_0 < t < t_{max}$ where t_0 is the time at which the submission was created and t_{max} is the time of the last recorded response. A temporal graph means that $N_t \subseteq N$ and $E_t \subseteq E$ as there are messages and edges still to be created. There is also only a subset of existing branches at t represented by $B(t) \subseteq B(t_{max})$ with a current subset of involved nodes. Now taking the available branches and the temporal network itself, we can take a static snapshot at time t to extract a row of response features.

B. Tests

Given this approach, we now have the ability to obtain a time-dependent response feature matrix F of dimensions $t \times l$ where t is $t \leq t_{max}$ and is used as a hyperparameter for our tests, and where $l = 6$, representing the size of our static feature space at time t.

$$F = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,6} \\ f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,6} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{t,1} & f_{t,2} & f_{t,3} & \cdots & f_{t,6} \end{bmatrix}$$

Timesteps	Gaming Score	Soccer Score
1	0.55	0.55
2	0.65	0.68
3	0.66	0.80
4	0.70	0.85
5	0.75	0.90
6	0.90	0.96

Table 4: Scores of K-means cluster classification for Subreddits Gaming (K=7) and Soccer (K=6)

Timesteps	Gaming, Politics	Soccer, Gaming
1	0.83	0.77
2	0.66	0.77
3	0.82	0.87
4	0.80	0.87
5	0.84	0.90
6	0.85	0.92

Table 5: Scores of Subreddit label classification for label sets, $|R| = 2$.

To evaluate the effectiveness of F, we consider two tests: classification by Subreddit, and classification by topic cluster. Though we are testing Subreddit classification again, our methodology for this test is different. We gather the same curated list of 1000 labeled submissions for $r/politics$, $r/gaming$, $r/soccer$, and $r/atheism$, and randomly choose two of those Subreddits to actively use. We then truncate their submission lists to a size of 100. Finding F for each submission for some t, we consider each submission using a Leave-One-Out approach. Thus given a test submission F_{test} with a hidden label R_{test}, we find the Euclidean Distance between the test submissions feature matrix and some other feature matrix. This measure is expressed by

$$d(F^X, F^Y) = \frac{1}{t} \sum_{t_{step}=0}^{t} \sum_{j=1}^{6} |f'^X_{t_{step},j} - f'^Y_{t_{step},j}|$$

Where F^X and F^Y represent some arbitrary features matrices, and f' indicates that the feature row at t_{step} has been standardized. So with a set of feature matrices and F_{test} we find

$$F_{target} = \arg\min_{F_i \in F} d(F_{test}, F_i)$$

Essentially, we are performing feature matrix matching. With F_{target} being the closest to F_{test}, we take F_{target}'s label R_{target} and assign it as R_{test}. Finally, we compare the real and predicted R_{test}. We repeat this process for topic clusters, but instead of looking for a Subreddit label, we try to predict to which cluster $C_{test} \in C$ our current F_{test} belongs to. The results for both tests given some set of C labels and some combination of R labels, each for a time range t, are presented in Table 4 and Table 5, respectively. For both tables the results hold that even for $t = 3$ the classification score is relatively high. This demonstrates that even within the first few hours the innate behaviors exhibited by conversing users are characteristic enough to be separable by both overarching theme and even by topic.

VI. Applications

Now that we have shown the accuracy and effectiveness of response features in both static and time-dependent settings, we can begin considering the applications within the field of information cascade analysis. Given the robustness of the response features and the breadth of the field of analysis, there are a wide variety of viable applications available. In this section we consider two of the more straightforward applications: tail forecasting and outlier detection. For simplicity, we will refer to the models developed in the following subsections as *resF* models, as they incorporate response features as an integral component.

A. Tail forecasting

In the previous section we illustrated how powerful the time-dependent response features are, even for a small t. We can take advantage of feature matrix matching to help us predict the tail of a cascade and complete its time-series. To utilize features for such a task we consider the tests we ran in the previous section. Given some initial time $t = t_{limit}$ we find the closest feature matrix F_{target} to our test submission. At this point, we take the response features F_{target} for the time range $t_{limit} < t \leq t_{max}$ to predict the remaining points in the test submission's time-series. Now the question arises: how do we reverse-engineer a time-series from response features?

Considering we know the time-series closest in similarity to our test submission, we can now use its response features to outline the constraints the must be followed by the time-series of the test submission as it evolves. We take the f rows of F_{target} for $t > t_{limit}$ and add them to the bottom of the incomplete matrix F_{test} whose original dimensions were $t_{limit} \times 6$. We then analyze the actual cascade that derives F_{target} to find the current number of branches at time t: $\beta(t) = |B(t)|$. With this and the expanded feature matrix F_{test}, we can outline our *resF* model for forecasting the incomplete tail of a test submission's time-series: at timestep t, iterate through $\beta(t)$ number of branches. For each branch, we sample depth, magnitude, engagement, and longevity from the closest submission's M_s. If these values are smaller than their previous ones, nothing changes for the branch. If the values are greater, then the branch is updated by adding nodes and edges to match the sampled values. This process is repeated for every branch at each timestep until t_{max}.

As a baseline, we also train an auto-regressive-integrated-moving-average (ARIMA) model [6] to predict the tail-end of test submission time-series. We compare the ARIMA model and the *resF* model to a selection of randomly selected time-series from $r/gaming$ and $r/politics$ where $t_{limit} = 3$. The predicted curves are evaluated against the ground truth using the Kolmogorov-Smirnov (KS) test [7] and Root-Mean-Squared-Error (RMSE). Figure 4 and Figure 5 show the results

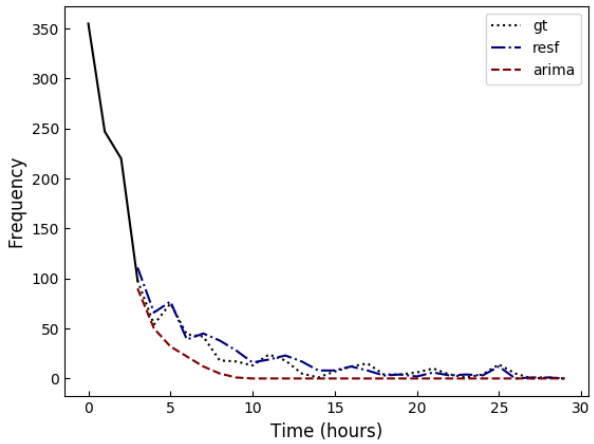

Figure 4: Forecast comparison for a $r/politics$ submission where $t_{limit} = 3$

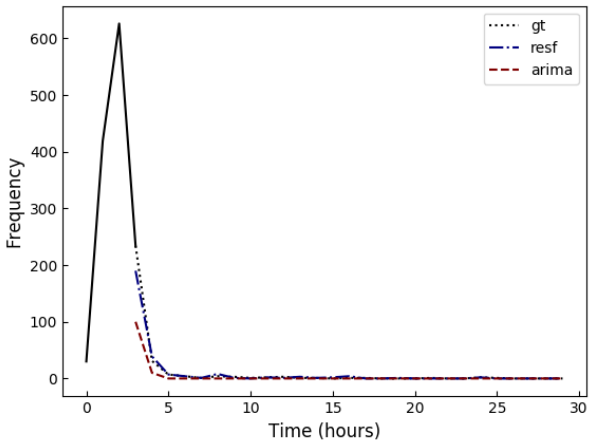

Figure 5: Forecast comparison for a $r/gaming$ submission where $t_{limit} = 3$

of the *resF* model compared to the predictions made by the ARIMA model for a submission from $r/politics$ and a submission from $r/soccer$, respectively. In both comparisons, the *resF* model forecasts an accurate tail pattern, matching the tail's local maxima with significant consistency. Conversely, the ARIMA model's predicted tails tend to drop off quickly, missing a good portion of the natural taper. In term of measurements, *resF* outperforms ARIMA as well, with KS D-values of $0.1166, 0.1143$, KS p-values of $0.7837, 0.9672$, and RMSE values of $4.972, 11.57$ in contrast to ARIMA's KS D-values of $0.5667, 0.4571$, KS p-values of $0.0001, 0.0007$, and RMSE values of $19.31, 36.76$. In a KS test, the null hypothesis indicates the two comparative samples are drawn from the same distribution. Thus, the high p-values and low D-values of the *resF* predictions imply the resultant empirical distribution functions are quite similar to the reference distribution functions derived from the ground truths. This conclusion, paired with the low RMSE values, demonstrates the predictive power of the *resF* model. It is important to note that tail forecasting can also be paired with the label classification method discussed in the previous section. Given a topic of the cascade, the set of potential F_{target} matrices can be filtered, allowing for more efficient and accurate predictions.

B. Outlier detection

One of the simplest yet most compelling applications is the use of the projection of the static response features for outlier detection. As illustrated in Section 4, we can identify topic clusters within a Subreddit using submission response features. Projecting these submissions into 2D Euclidean space like in Figure 2 or Figure 3 can visualize feature groupings that gives us some insight into how similar submissions are in feature space. In addition to analyzing groupings, we can also pinpoint isolated submissions that lie apart from the clusters. The significant spatial distance indicates a distinct difference in the response features associated with that isolated submission. We can use this identifying factor to indicate submissions of potential interest in a large set of general submissions. We test our method for outlier detection by analyzing 1000 submissions for $r/gaming$, again, with a comment range of $1000 - 3000$. Evaluating these submissions' response features and reducing dimensions using PCA produces Figure 6. While a majority of submissions can be found in C_3 and C_7, there are a few outliers. In terms of comparative distance, C_5 is the most remote.

If we analyze the text of the greatest group of submissions, we find that C_7 primarily contains video game giveaways, while C_3 consists of more general discussions. C_5, however, has a single submission whose text contains a story about a person who beat cancer and stayed emotionally strong thanks to video games. In the context of $r/gaming$ this can be seen as a very appealing story, and the influx of supportive responses confirms this. This analysis emphasizes the indicative power of response features when it comes to capturing unique conversational cascades. Applied to a much larger, more general dataset for some arbitrary online social media would allow a researcher to identify a set of outliers. With these outliers pinpointed, they could then begin to find additional patterns that could inform them of the nature of virality for that particular platform.

VII. Related Work

This paper began with a discussion of the paradigm of topic and sub-topic detection through user reactions. And while the idea of characterizing user reactions and biases for the analysis of cascades is not new, its potential is primarily unrealized. As mentioned in Section 3, [2] introduces novel research that illustrates the stability of user topic bias in Twitter. Unfortunately, this stable representation of bias is only sparesely applied, being primarily used for the prediction of topic-specific influencers and the high-level analysis of Twitter's follower structure. Extracting user reactions has been

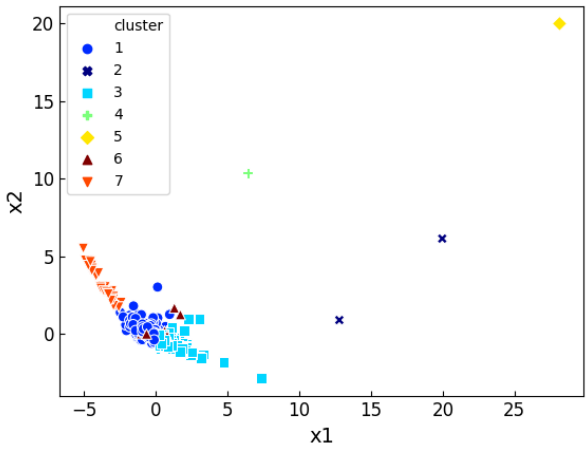

Figure 6: PCA of 1000 submission response features for the Subreddit $r/gaming$.

seen as an essential tool for social media analysis for a while now as well. However, the use of reactions tends to be limited to NLP and the quantification of baseline interest (like number of comments per minute) [8]–[10]. In our paper, the concept of innate topic bias being exhibited through response behavior is expanded and used as the cornerstone for response feature design, allowing these features to avoid any semantic dependencies while completely overhauling the idea of how comments reflect interest.

The first two tests of these response features in this paper were on label classification and topic clustering, both of which can be considered a form of topic classification. Previously, topic classification model architecture was primarily built around the extraction of text [11], [12], though network structure has been used as auxiliary data to enhance the model [13]. A non-semantic approach to topic classification is novel, however, as the resultant model does not rely on any kind of text extraction methods to define topic groups. Instead, we harness user reactions for automatic classification of topic clusters, without a priori assumptions about topic modeling methods.

The response features outlined in this paper were also used in tail prediction of information cascades and the detection of submission outliers as application examples. Concerning the first of the two applications, the prediction of bursty information cascades is one of the central focuses of online social media analysis [14], and a necessary component of cascade prediction can be found in tail forecasting. Tail forecasting tests the ability to extract the latent trends of a bursty event and extrapolate the remainder of its lifetime, from predicting the magnitude of the initial peak to calculating the effect of additional shocks past the first spike. Most models tend to use a Hawkes process [15], [16] or follow some multi-feature architecture [17], [18]. Often the former outperforms the latter given that the multi-feature approach often requires extensive feature engineering and ad-hoc decisions on the learning hyperparameters. This approach also tends to lock the model into a certain platform or problem domain due to the potentially unique features that might end up being introduced into the model's training process. And while the *resF* model could be considered a multi-feature model in terms of cascade prediction, it does not suffer the common shortcomings mentioned above. The *resF* model has a small set of robust features that are independent of hyperparameters. And given *resF*'s focus on simply leveraging a user's innate bias of topics to predict cascades, this process can be generalized to any platform that has posts and comments that topologically represent isolated tree graphs.

Outlier detection in the broader scope of understanding viral or unique posts on online social media has also been a major focal point of research. This has usually been done through identifying influential users [19] or by analyzing the behavior of past trends [20], [21]. But these approaches can be refined by considering the response features of involved groups (i.e. followers or subscribers). Given that response features will automatically isolate unique posts, this phenomena can be used to enhance methods of detection of both influential users and trending topics.

VIII. Conclusion

In this paper we introduced a novel approach to analyzing social media information cascades using only the audience reactions triggered by a source post. Following this approach, we designed unique features to represent different aspects of conversational dynamics. Using these measures as response features to characterize an information cascade, we validated the effectiveness of these representative features through label classification and topic clustering. After validation, we enhanced the static response features by introducing a temporal dependency, allowing us to analyze response features for a submission as it evolves with time. This time-dependent method was validated as well using label classification. Given these tools, we finally introduced some direct applications: tail forecasting for evolving time-series, and outlier detection. Both of these applications produced accurate and valid results, further indicating the effectiveness of harnessing response features. Overall, a reaction-based approach to information cascade analysis is an effective tool for non-semantic topic classification, time-series prediction, and much more.

IX. Acknowledgements

This work was sponsored in part by DARPA under contract W911NF-17-C-0099, the Army Research Office (ARO) under contract W911NF-17-C-0099, and the Office of Naval Research (ONR) under grant N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the U.S. Government.

REFERENCES

[1] J. Hessel, C. Tan, and L. Lee, "Science, askscience, and badscience: On the coexistence of highly related communities," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[2] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 236–242, ACM, 2013.

[3] M. Kulisiewicz, P. Kazienko, B. K. Szymanski, and R. Michalski, "Entropy measures of human communication dynamics," *Scientific reports*, vol. 8, no. 1, p. 15697, 2018.

[4] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[5] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[6] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to time series and forecasting*, vol. 2. Springer, 2002.

[7] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[8] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1195–1198, ACM, 2010.

[9] B. D. Horne, S. Adali, and S. Sikdar, "Identifying the social signals that drive online discussions: A case study of reddit communities," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9, IEEE, 2017.

[10] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the life cycle of online news stories using social media reactions," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 211–223, ACM, 2014.

[11] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pp. 90–94, Association for Computational Linguistics, 2012.

[12] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Topic modeling for short texts with auxiliary word embeddings," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 165–174, ACM, 2016.

[13] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 251–258, IEEE, 2011.

[14] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM Sigmod Record*, vol. 42, no. 2, pp. 17–28, 2013.

[15] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 6–14, ACM, 2012.

[16] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522, ACM, 2015.

[17] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proceedings of the 21st international conference on World Wide Web*, pp. 1145–1152, ACM, 2012.

[18] 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*, pp. 925–936, ACM, 2014.

[19] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65–74, ACM, 2011.

[20] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," in *Fifth international AAAI conference on weblogs and social media*, 2011.

[21] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*, pp. 57–58, ACM, 2011.