Ignorance isn't bliss: an empirical analysis of attention patterns in online communities

Ignorance isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities

Claudia Wagner*, Matthew Rowe†, Markus Strohmaier‡, and Harith Alani†
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†Knowledge Media Institute, The Open University, Milton Keynes, UK
Email: m.c.rowe@open.ac.uk, halani@open.ac.uk
‡ Knowledge Management Institute and Know-Center, Graz University of Technology,Graz, Austria
Email: markus.strohmaier@tugraz.at

***Abstract*—Online community managers work towards building and managing communities around a given brand or topic. A risk imposed on such managers is that their community may die out and its utility diminish to users. Understanding what drives attention to content and the dynamics of discussions in a given community informs the community manager and/or host with the factors that are associated with attention, allowing them to detect a reduction in such factors. In this paper we gain insights into the idiosyncrasies that individual community forums exhibit in their attention patterns and how the factors that impact activity differ. We glean such insights through a two-stage approach that functions by (i) differentiating between seed posts - i.e. posts that solicit a reply - and non-seed posts - i.e. posts that did not get any replies, and (ii) predicting the level of attention that seed posts will generate. We explore the effectiveness of a range of features for predicting discussions and analyse their potential impact on discussion initiation and progress.**

Our findings show that the discussion behaviour of different communities exhibit interesting differences in terms of how attention is generated. Our results show amongst others that the purpose of a community as well as the specificity of the topic of a community impact which factors drive the reply behaviour of a community. For example, communities around very specific topics require posts to fit to the topical focus of the community in order to attract attention while communities around more general topics do not have this requirement. We also found that the factors which impact the start of discussions in communities often differ from the factors which impact the length of discussions.

***Index Terms*—attention, online communities, discussion, popularity, user generated content**

I. Introduction

Social media applications such as blogs, video sharing sites or message boards allow users to share various types of content with a community of users. For the managers of such communities, the investment of time and money means that community utility is paramount. A reduction in activity could be detrimental to the appearance of the community to outside users, conveying an impression of a community that is no longer active and therefore of little utility. The different nature and intentions of online communities means that what drives attention to content in one community may differ from another. For example, what catches the attention of users in a question-answering or a support-oriented community may not have the same effect in conversation-driven or event-driven communities. In this paper we use the number of replies that a given post on a community message board yields as a measure of its attention.

To explore these and related questions, our paper sets out to study the following two research questions:

1) *Which factors impact the attention level a post gets in certain community forums?*
2) *How do these factors differ between individual community forums?*

Understanding what factors are associated with attention in different communities could inform managers and hosts of community forums with the know-how of what drives attention and what catches the attention of users in their community. Empowered with such information, managers could then detect changes in such factors that could potentially impact community activity and cause the utility of the community to alter.

We approach our research questions through an empirical study of attention patterns in 20 randomly selected forums on the Irish community message board Boards.ie.[1] Our study was facilitated through a two-stage approach that (i) differentiates between seed posts - i.e. thread starters on a community message board that got at least one reply - and non-seed posts - i.e. thread starters which did not get a single reply, and (ii) predicts the level of attention that seed posts will generate - i.e. the number of replies. Through the use of five distinct feature sets, containing a total of 28 features and including *user*, *focus*, *content*, *community* and *post title* features, we analysed how attention is generated in different community forums. We find interesting differences between these communities in terms of what drives users to reply to thread starters initially (through our *seed post identification experiment*) and what factors are associated with the length of discussions (through our *seed post activity level prediction experiment*). Our work

[1] http://www.boards.ie

is relevant for researchers interested in behavioural analysis of communities and analysts and community managers who aim to understand the factors that are associated with attention within a community.

The paper is structured as follows: section 2 describes related work within the fields of attention prediction on different social web platforms. Section 3 describes the dataset and Section 4 describes the features used in our analysis. Section 5 presents our experiments on identifying seed posts and anticipating their attention level in different communities. Section 6 discusses our findings and relates them to previous research. Section 7 concludes the paper with a summary of the key findings gleaned from our experiments and plans for future work.

II. Related Work

Attention on social media platforms can be gauged through assessing the number of replies that a piece of content or user receives. Within this context [1] consider the problem of reciprocity prediction and study this problem in a communication network extracted from Twitter. They essentially aim to predict whether a user A will reply to a message of user B by exploring various features which characterise user pairs and show that features that approximate the relative status of two nodes are good indicators of reciprocity. Our work differs from [1], since we do not aim to predict who will reply to a message, but consider the problem of identifying posts which will start a discussion and predicting the length of discussions. Further, we focus on exploring idiosyncrasies in the reply behaviour of different communities, while the above work studies communication networks on Twitter without differentiating between individual sub-communities which may use Twitter as a communication medium.

The work presented in [2] investigates factors that impact whether Twitter users reply to messages and explores if Twitter users selectively choose whom to reply to based on the topic or, otherwise, if they reply to anyone about anything. Their results suggest that the social aspect predominantly conditions users' interactions on Twitter. Work described in [3] considers the task of predicting discussions on Twitter, and found that certain features were associated with increased discussion activity - i.e. the greater the broadcast spectrum of the user, characterised by in-degree and list-degree levels, the greater the discussion activity. Further, in our previous work [4] we explored factors which may impact discussions on message boards and showed, amongst others, that content features are better indicators of seed posts than user features. Similar to our previous work [4] we also aim to predict discussions on message boards, but unlike past work, which aimed to identify global attention patterns, we focus on exploring and contrasting the discussion behaviour of individual communities.

Closely related to the problem of anticipating the reply-behaviour of social media users is the problem of predicting the popularity and virality of content. For example, the work described in [5] consider the task of predicting the rank of stories on Digg and found that the number of early comments and their quality and characteristics are useful indicators. Hong et al. [6] investigated the problem of predicting the popularity of messages on Twitter measured by the number of future retweets. One of their findings was that the likelihood that a portion of a user's followers will retweet a new message depends on how many followers the user has and that messages which only attract a small audience might be very different from the messages which receive huge numbers of retweets. Similar work by [7] explored the relation between the content properties of tweets and the likelihood of the tweets being retweeted. By analysing a logistic regression model's coefficients, Naveed et al. [7] found that the inclusion of a hyperlink and using terms of a negative valence increased the likelihood of the tweet being retweeted. The work of [8] explores the retweet behaviour of Twitter users by modeling individual micro-cosm behaviour rather than general macro-level processes. They present four retweeting models (general model, content model, homophily model, and recency model) and found that content based propagation models were better at explaining the majority of retweet behaviours in their data. Szabo et al. [9] studied content popularity on Digg and YouTube. They demonstrated that early access patterns of users can be used to forecast the popularity of content and showed that different platforms reveal different attention patterns. For example, while Digg stories saturate fairly quickly (about a day) to their respective reference popularities, YouTube videos keep attracting views throughout their lifetimes. In [10] the authors present a mutual dependency model to study the virality of hashtags in Twitter.

Although its is well-known that sub-communities of users can be identified on most social media applications, previous research did not explore differences in the attention patterns of such sub-communities. To the best of our knowledge, our work is the first to focus on exploring idiosyncrasies of communities' attention patterns by comparing the reply behaviour of different community forums. We also provide an extended set of features to assess the effects that community and focus features have on reply behaviour, something which has not been explored previously.

III. Dataset: Boards.ie

In this work, we analysed data from an Irish community message board, Boards.ie, which consists of 725 community forums ranging from communities around specific computer games or spiritual groups to communities around general topics such as films or music. Since our goal is to uncover the idiosyncrasies that individual community forums exhibit and the deltas between them, we selected 20 forums at random.

- *Forum 374 - Weather*: Community of users who have special interest in weather. This forum allows users to talk about the current, future and past weather all over the world and share information - e.g. weather pictures.
- *Forum 10 - Work & Jobs*: The community around this forum consists of users who are looking for jobs, offering jobs and/or are seeking advice in work-related things. This means that the community has, on the one hand,

a support and advice offering purpose and, on the other hand, is a marketplace for users who are in similar situations.
- *Forum 221 - Spanish*: Community of practice where users share a common long-term goal - namely to learn, improve or practice their Spanish.
- *Forum 343 - Golf*: Community of users who are interested in the sport Golf. In this forum users can discuss anything related with golf.
- *Forum 646 - adverts.ie Support*: A support oriented forum for adverts.ie, which is a community based marketplace where individuals can buy or sell items online.
- *Forum 235 - Rip Off Ireland*: Support-oriented forum which aims to help consumers in Ireland avoid being ripped off with the current spate of Euro price hikes.
- *Forum 865 - Home Entertainment (HE) Video Players & Recorders*: Community of users formed around a specific group of products namely HE Video Players and Recorders. In this forum users are seek advice and discuss issues related these products.
- *Forum 544 - Banking & Insurance & Pensions*: Support and advice oriented community of users who seek or provide advice about banking, insurance and pensions.
- *Forum 876 - Construction & Planning*: Forum where users can discuss topics related to construction and planning.
- *Forum 267 - Astronomy & Space*: Information and content-sharing community of users who are interested in astronomy and space.
- *Forum 669 - Google Earth* : Forum where users talk about Google Earth.
- *Forum 55 - Satellite*: Information and content-sharing community where users who are interested in satellite television can discuss this topic.
- *Forum 858 - Economics*: Community of users who have a special interest or expertise in economics.
- *Forum 44 - CTYI*: Community of users around the Centre for the Talented Youth of Ireland (CTYI) which is a youth programme for students between the ages of six and sixteen of high academic ability in Ireland.
- *Forum 538 - Japanese RPG*: Community of users playing Japanese role games.
- *Forum 227 - Television*: Discussion about television related topics such as TV series.
- *Forum 607 - Music Production*: Community of music producers and/or people interested in music and music production in general.
- *Forum 630 - Real-World Tournaments & Events*: Forum where users talk about events and tournaments - i.e. competitions involving a relatively large number of competitors, all participating in a sport, game or event.
- *Forum 190 - North West*: Forum around the North West of Ireland, where users who live in the North West or plan to visit the North West can discuss related questions.
- *Forum 625 - Greystones & Charlesland*: Forum where users talk about everything related with Charlesland and Greystones which are both located about 25 kilometres from Dublin city centre.

For our analysis we use all data published in one of these 20 forums in the year 2006. We use this year to enable comparisons of attention patterns with our previous work [4] over the same time period. Table I describes the properties of the dataset.

TABLE I
DESCRIPTION OF THE BOARDS.IE DATASET.

Forum	ID	Users	Posts	Threadstarter	Seeds
Work & Jobs	10	2371	13964	1741	1435
Music Production	607	308	2018	295	265
Golf	343	394	3361	415	364
Astronomy & Space	267	247	782	141	97
Weather	374	439	7598	233	209
HE Video Players & Recorders	865	134	294	61	52
Banking & Insurance & Pensions	544	956	3514	531	459
Google Earth	669	117	584	37	32
Satellite	55	1516	14704	1714	1620
Economics	858	73	260	28	26
Espanol (Spanish)	221	21	86	31	21
Rip Off Ireland	235	28	329	34	28
Construction & Planning	876	34	202	35	
CTYI	44	39	1505	42	39
Japanese RPG	538	71	1157	75	71
adverts.ie Support	646	304	1227	216	172
Television	227	2086	17442	1238	1139
North West	190	376	4866	291	271
Greystones & Charlesland	625	396	4930	418	382
Real-World Tournaments & Events	630	640	18551	1475	1172

IV. FEATURE ENGINEERING

Understanding what factors drive reply behaviour in online communities involves defining a collection of features and then assessing which are important and which are not. Within our approach setting we can identify the features that impact upon seeding a discussion - through our seed post identification experiments - and how features are associated with seed posts that generate the most attention.

For each thread starter post we computed the features by taking a 6-month window, based on work by [4], [11], prior to when the post was made. That means, we used all the author's past posts within that window to construct the necessary features - i.e. constructing a social network for the user features, assessing the forums in which the posts were made for the focus features and inferring topic distributions per user based the content of posts he/she authored within the previous 6 month. For the features that relied on topic models, we first fit a Latent Dirichlet Allocation [12] model which we use later for inferring users' topic distributions. For training the LDA model we aggregated all posts authored by one user in 2005 into an artificial user document and chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and $T = 50$) which we optimised during training by using Wallach's fixed point iteration method [13]. Based on the empirical findings of [14], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. We used the trained model to infer the average topic

distributions (averaged over 10 independent runs of a Markov chain) of a user at a certain point in time by using all posts he/she authored within the last 6 months.

We define five feature sets: user features, focus features, content features, community features and title features, as follows.

A. User Features

User features describe the author of a post via his/her past behaviour, seeking to identify key behavioural attributes that are associated with seed and non-seed posts. For example, a post may only start a lengthy discussion if published by a rather active user.

- *User Account Age:* Measures the length of time (measured in days) that the user has been a member of the community;
- *Post Count:* Measures the number of posts that the user has made.
- *Post Rate:* Measures the number of posts made by the user per day.
- *In-degree:* For the author of each post, this feature measures the number of incoming communication connections to the user.
- *Out-degree:* This feature measures the number of outgoing communication connections from the user.

B. Focus Features

Focus features measure the topical concentration an author. Our intuition is that by gauging the topical focus of a user we will be able to capture his/her areas of interest or expertise. For the first two features, we use the frequency distribution of forums a user has published posts in to approximate his/her interests or expertise, while for the last three features we learn topics from a collection of posts and annotate users with topics by using LDA.

- *Forum Entropy:* Measures the forum focus of a user via the entropy of a user's forum distribution. Low forum entropy would indicate high focus.
- *Forum Likelihood:* Measures the likelihood that the user will publish a post within a forum given the past forum distribution of the user.
- *Topic Entropy:* Measures the topical focus of a user via the entropy of a user's topic distributions inferred via the posts he/she authored. Low topic entropy would indicate high focus.
- *Topic Likelihood:* Measures the likelihood that the user will publish a post about certain topics given the past topic distribution of the user's posts. Therefore, we measure how well the user's language model can explain a given post by using the likelihood measures:

$$likelihood(p) = \sum_{i=0}^{N_p} \ln P(w_i|\hat{\phi}, \hat{\theta}) \tag{1}$$

  N_p refers to the total number of words in the post, $\hat{\phi}$ refers to the word-topic matrix and $\hat{\theta}$ refers to the average topic distribution of a user's past posts. The higher the likelihood for a given post, the greater the post fits to the topics the user has previously written about.
- *Topic Distance:* Measures the distance between the topics of a post and the topics the user wrote about in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between the user's past topic distribution and the post's topic distribution. The JS divergence is defined as follows:

$$D_{JS} = \frac{1}{2} D_{KL}(P||A) + \frac{1}{2} D_{KL}(A||P) \tag{2}$$

  where $D_{KL}(P||A)$ represents the Kullback Leibler divergence between a random variable P and A. The KL divergence is calculated as follows:

$$D_{KL}(P||A) = \sum_i P(i) \log \frac{P(i)}{A(i)} \tag{3}$$

  The lower the JS divergence, the greater the post fits the topics the user has previously written about.

C. Post Features

Post features describe the post itself and identify attributes that the content of a post should contain in order to start a discussion. For example, a post may only start a lengthy discussion if its content is informative or if it was published at a certain time in the day.

- *Post Length:* Number of words in the post.
- *Complexity:* Measures the cumulative entropy of terms within the post, using the word-frequency distribution, to gauge the concentration of language and its dispersion across different terms.
- *Readability:* Gunning fog index using average sentence length (ASL) [15] and the percentage of complex words (PCW): $0.4 * (ASL + PCW)$ This feature gauges how hard the post is to parse by humans.
- *Referral Count:* Count of the number of hyperlinks within the post.
- *Time in day:* The number of minutes through the day from midnight that the post was made. This feature is used to identify key points within the day that are associated with seed or non-seed posts.
- *Informativeness:* The novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure.
- *Polarity:* Assesses the average polarity of the post using Sentiwordnet.[2] Let n denote the number of unique terms in post p, the function $pos(t_.)$ returns the positive weight of the term $t_.$ from the lexicon and $neg(t_.)$ returns the negative weight of the term. We therefore define the polarity of p as:

$$\frac{1}{n} \sum_{i=1}^{n} pos(t_i) - neg(t_i) \tag{4}$$

[2]http://sentiwordnet.isti.cnr.it/

D. Community Features

Community features describe relations between a post or its author and the community with which the post is shared. For example, members of a community might be more likely to reply to a post which fits their areas of interest or they might be likely to reply to someone who contributed a lot to discussions in the past.

- *Topical Community Fit:* Measures how well a post fits the topical interests of a community by estimating how well the post fits into the forum. We measure how well the community's language model can explain the post by using the likelihood measure which is defined in equation 1, where $\hat{\theta}$ refers to the average topic distribution of posts that were previously published in that forum. The higher the likelihood of the post, the better the post fits to the topics of this community forum.
- *Topical Community Distance:* Measures the distance between the topics of a post and the topics the community discussed in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between a community's past topic distribution and a post's topic distribution. The JS divergence is defined in equation 2. The lower the JS divergence, the greater the post fits the topical interests of the community.
- *Evolution score:* Measures how many users of a given community have replied to a user in the past, differing from *in-degree* by being conditioned on the forum. Theories of evolution [16] suggest a positive tendency for user A replying to user B if A previously replied to B. Therefore, we define the evolution score of a given user u_j as follows:

$$evolution(u_j) = \sum_{i}^{U} \frac{U(u_{j,i}) + 1}{U} \tag{5}$$

  where U refers to the total number of users in a given forum and $U(u_{j,i})$ refers to the number of users who replied to user u_j in the past.
- *Inequity score:* Measures how many users of a given community a user has replied to in the past, differing from *out-degree* by being conditioned on the forum. Equity Theory [17] suggests a positive tendency for user A replying to user B if B previously replied more often to A than A to B. Therefore, we define the inequity score of a user u_j as follows:

$$inequity(u_j) = \sum_{i}^{U} \frac{|P(u_{i,j})_{reply}|}{|P(u_{j,i})_{reply} + 1|} \tag{6}$$

  where U refers to the total number of users in a given forum, $P(u_{i,j})_{reply}$ refers to the probability that user u_i replies to user u_j and $P(u_{j,i})_{reply}$ refers to the probability that user u_j replies to user u_i

E. Title Features

Title features describe the title of a post itself and identify attributes that the title should contain in order to start a discussion. We decided to separate title features from post features in order to be able to capture potential affects of the user interface since the current Boards.ie user interface encourages users to decide which post to read based on the title. Therefore, our intuition is that in some community forums, title features may have a greater influence on the start of discussions as well as on the development of lengthy discussions.

- *Title Length:* Number of words in the title of the post.
- *Title Question-mark:* Measures the absence or presence of a question-mark in the title.

V. Experiments

Understanding what drives attention in different forums and their implicit communities enables us to reveal key differences between those forums. To detect such deltas we apply our two-stage prediction approach to (i) detect seed posts within each forum and (ii) predict the level of activity that such seed posts will generate. We begin by explaining our experimental setup before going on to discussing our findings and observing how the communities differ from one another in their discussion dynamics.

A. Experimental Setup

For our experiments we took all the thread starter posts - i.e. that were both seeds and non-seeds - published in each of the 20 forums throughout the year 2006. For each thread starter we constructed the features as described in the previous section. We performed two experiments using our generated datasets, each intended to explore the research questions: (i) *Which factors may impact the attention level a post gets in certain community forums?* and (ii) *How do these factors differ between individual community forums?*

1) Seed Post Identification: The first experiment sought to identify the factors that help differentiating between posts that initiate discussions and posts that do not get any attention in different communities. To this end, we performed *seed post identification* through a binary classification task using a logistic regression model. For each forum, we divided the forum's dataset into a training/testing split using an 80/20% split, trained the logistic regression model using the former split and applied it to the latter. We tested each of the five feature sets in isolation - i.e. user, focus, post, community and title - such that the model was trained using only those features, and then tested all the features combined together. To assess how well each model performed, we measured the F1 score, which is the harmonic mean of precision and recall, and the Matthews correlation coefficient (MCC), which is a balanced measure of the quality of binary classification and can be used even if the classes are of very different sizes. The MCC measure returns a value between -1 and $+1$: a coefficient of $+1$ represents a perfect prediction, 0 is no better than random prediction and -1 indicates total disagreement between prediction and observation. The F1 score is frequently used by the Information Retrieval community, while the MCC

is widely used by the Machine Learning community and in statistics where it is known as phi (ϕ) coefficient.

The best performing model was then chosen based on the F1 score and MCC value and the coefficients of the logistic regression model were inspected to detect how the features were associated with seed posts, thereby identifying the factors which impact reply behaviour. To gain further insights into which features contribute most to the classification model, we also ranked the features of the best performing model by using the Information Gain Ratio (IGR) as a ranking criterion.

2) Activity Level Prediction: For our second experiment, we sought to identify the factors that were correlated with lengthy discussions and how they differed between communities. To do this we performed *seed post activity level prediction* through a linear regression model. We maintained the same splits as in our previous experiment and filtered through the seed posts in the 20% test split using the best performing model in each community. We then trained a linear regression model using the seed posts in the training split and predicted a ranking for the identified seed posts in the test split based on expected discussion volume. This allowed us to pick out the key factors that were associated with generating the most activity by concentrating our rank assessments on the top portion of the posts. We trained the linear regression model using each of the five feature sets in isolation and then used all the features combined together. We chose the best performing model based on its rank prediction accuracy and assessed the statistically significant coefficients of the regression model for the relation between increased attention and its features.

To evaluate our predicted rank, we used the Normalised Discounted Cumulative Gain (nDCG) at varying rank positions, looking at the performance of our predictions over the top-k documents where $k = \{1, 5, 10, 20, 50, 100\}$, and then averaging these values. nDCG is derived by dividing the Discounted Cumulative Gain (DCG) of the predicted ranking by the actual rank defined by (iDCG). DCG is well suited to our setting, given that we wish to predict the most popular posts and then expand that selection to assess growing ranks, as the measure penalises elements in the ranking that appear lower down when in fact they should be higher up. Let $rank_i$ be the actual position in the ranking that seed post i should appear and N be the number of items in the total set of seed posts that are to be predicted, we then define $rel_i = N - rank_i + 1$ and DCG based on the definition from [18] as:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(1+i)} \tag{7}$$

B. Results: Seed Post Identification

Comparing the F1 score and MCC values of different forums in Table II reveals interesting differences between communities and corroborates our hypothesis that the reply behaviour of users in different communities is impacted by different factors. Table II shows the 9 forums for which a classifier trained with our features outperformed the baseline classifier. We decided not to analyse the results from the other 11 forums, since our classifier did not outperform (but only matched) the performance of the baseline. We assume that this happens because most of these 11 forums are rather inactive forums such as forum 44 or 858 (i.e. only a few messages have been published in 2006 and therefore our classifier had not enough examples of seed and/or non-seed posts to learn general attention patterns). Another potential explanation is that the discussion behaviour of these communities is in part rather random and/or driven by other, external factors which we could not take into account in our study. For example the discussion behaviour of the communities around specific locations or regions (such as community 190 and 625) might for example be impacted by spatial properties of users while the discussion behaviour of the community around forum 227 (Television) seems to be mainly driven by external events (e.g. start of a new series).

Our results from the seed post identification experiment show that for most of the 9 forums a classifier trained with a combination of all features achieves the highest performance boost. Only for the community around forum 267 (Astronomy and Space) a classifier trained with content features alone performs best. This example nicely shows that this community seems to be mainly content driven since its main purpose is to share information and content. Another exception is the community of practice around forum 221 (Spanish) for which a classifier trained with title features alone and a classifier trained with user features alone outperforms a classifier trained with all feature groups. This indicates that the features of those two groups best capture the characteristics of seed and non-seed posts in this community.

To gain further insights into the factors that impact attention in different communities we inspected the statistically significant coefficients of the best performing feature group learned by the logistic regression model. The coefficients can be interpreted as the log-odds for the features. Therefore, a positive coefficient denotes a higher probability of getting replies for posts having this feature. In addition to interpreting the statistically significant coefficients we also ranked the features of the best performing feature group by using the Information Gain Ratio (IGR) as a ranking criterion. The higher the information gain of a feature the higher the average purity of the subsets that it produces. A feature with a maximum information gain ratio of 1 would enable perfect separation between seed and non seed posts. Due to space constraints we only discuss features with an $IGR >= 0.1$.

Our results suggest that in the community around forum 10 (Work & Jobs) which has a support and marketplace function, longer posts (content length's $coef = 0.063$ and $p < 0.001$) which do not really contain new information (informativeness $coef = -0.028$ and $p < 0.001$) and/or links ($coef = -0.592$ and $p < 0.01$) are far more likely to get replies. Further, posts which contain question marks ($coef = 0.454$ and $p < 0.01$) in their title are more likely to attract the attention of this support-oriented community. Finally, since the topic of this community is quite general, posts are not required to be topically similar to other posts in the forum (community fit's $coef = -221.844$

TABLE II
F1 SCORE AND MATTHEWS CORRELATION COEFFICIENT (MCC) FOR DIFFERENT FORUMS WHEN PERFORMING SEED POST IDENTIFICATION. THE BEST PERFORMING MODEL FOR EACH FORUM IS MARKED IN BOLD.

forumid	User		Focus		Content		Community		Title		All	
	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
10	0.0	0.75	0.0	0.75	0.071	0.76	0.0	0.75	0.0	0.75	**0.1**	**0.766**
607	0.332	0.839	0.0	0.802	0.0	0.802	0.0	0.802	0.0	0.802	**0.359**	**0.857**
343	0.0	0.769	0.0	0.769	0.093	0.782	0.0	0.769	0.0	0.769	**0.148**	**0.789**
267	0.078	0.609	-0.132	0.531	**0.242**	**0.673**	0.078	0.609	0.0	0.549	0.181	0.643
865	0.0	0.533	0.0	0.533	0.0	0.533	0.0	0.533	0.0	0.533	**0.632**	**0.815**
544	0.0	0.818	0.0	0.818	-0.052	0.809	0.0	0.818	0.0	0.818	**0.109**	**0.828**
55	0.0	0.913	0.0	0.913	0.0	0.913	0.0	0.913	0.0	0.913	**0.144**	**0.918**
221	0.447	0.625	-0.447	0.25	0.0	0.486	0.0	0.333	**0.707**	**0.829**	0.0	0.333
630	0.0	0.678	0.0	0.678	-0.044	0.675	0.0	0.678	0.0	0.678	**0.109**	**0.686**

and $p < 0.01$) in order to attract attention.

Another support and advise oriented community is the community around forum 343 (Golf). The topic of this community is a more specific than the topic of the previous community. In this community the content of a post needs to be rather complex (*coef* $= 2.261$ and $p < 0.01$) and should also not contain links (*coef* $= -0.586$ and $p < 0.05$) in order to attract attention. Further posts which are topically distinct from what the Golf community usually talks about (community distance *coef* $= -4.528$ and $p < 0.05$) are less likely to get replies. This indicates that within the community specialist terminology is used and the divergence away from such vocabularies reduces the likelihood of generating attention to a new post.

The community around forum 865 (HE Video Players & Recorders) has an advice seeking and experience sharing purpose but only for one specific group of products. For this community forum all features' coefficients are not significant. However, a classification model trained with all features outperformed a random baseline classification model with a MCC value of 0.632. By looking at the feature list ranked by the *IGR*, we note that only one feature contributed to this performance boost, namely the inequity score (*IGR* $= 0.7$). The coefficient of the inequity score in the regression model is negative (*coef* $= -5.025$) which indicates that a post is less likely to get replies if it is authored by a user who replied to many posts in this forum in the past but hasn't got many replies himself in this forum. One possible explanation is that in support oriented communities users who reply to many posts are more likely to be experts. It is not surprising that posts of such expert users are less likely to get replies since less users have enough expertise to answer or comment on the post of an expert.

The main purpose of the community around forum 544 (Banking & Insurance & Pensions) is also for seeking advice and sharing experiences and information. In this community shorter posts (content length *coef* $= -0.017$ and $p < 0.05$) authored by users who are new to the topic - or have not published anything about the topic before (topic distance *coef*= 2.890 and $p < 0.01$) - are more likely to get replies. When inspecting the *IGR* based feature ranking of the content group, we find that only the complexity of content is a useful feature for informing a classifier which has to differentiate between seed and non seeds (*IGR* $= 0.354$). This indicates that short, but complex posts which have been authored by newbies are most likely to catch the attention of this community.

The main purpose of the community around forum 267 (Astronomy & Space) is to share information and content and to engage in discussions. Long posts (*coef* $= 0.083$ and $p < 0.05$) which do not contain many novel terms (informativeness *coef* $= -0.029$ and $p < 0.05$) but are positive in their sentiment (polarity's *coef* $= 4.556$ and $p < 0.05$) are very likely to attract the attention of this community. The content feature with the highest *IGR* is the number of links per post (*IGR* $= 0.1$). Since the coefficient of the number of links is positive in our regression model we can conclude that a higher number of links indicates that the post is more likely to get replies (*coef* $= 0.157$) in this forum. This suggests that in this forum posts which are long, informative and re-use the vocabulary of the community are more likely to attract attention.

Also for the topical community around forum 55 (Satellite) the main purpose is to share information and content and to engage in discussions. In this community posts authored by users who have a high forum likelihood are less likely to get replies (*coef* $= -5.891$ and $p < 0.01$). This suggests that users who stimulate discussions in this community have to focus their activity away from this forum. Further posts which are topically distant from the topics the community usually talks about are again less likely to get replies (*coef* $= -2.944$ and $p < 0.01$). This pattern indicates that users who focus their activity away from this community and then post a new thread that is about topics which seem to be in the topical interest area of the community are more likely to get replies.

The community around forum 221 (Spanish) is a community of practice which means that the community members have a common interest in a particular domain or area, and learn from each other. This community is mainly impacted by user and title factors, however all features' coefficients are not significant. Ranking the features by their *IGR* shows that the most important feature for discriminating between posts getting replies and posts not getting replies is the title length (*IGR* $= 0.558$). Interestingly in this forum, posts with short titles are more likely to get replies. The longer the title the less likely a post gets replies (title length's *coef* $= -0.326$). The second most important feature is the user account age (*IGR* $= 0.381$). Users who have owned an account for

longer are more likely to get replies in this forum than users who recently created their account. This suggests that in communities where the members share a common long-term goal and/or have a shared interest which is rather stable over time, the duration of users' community membership is a good feature to predict if a post will become a seed post or not.

The community around forum 630 is a rather open and diverse community of users who are interested in all kind of events and/or want to promote events. For forum 630 (Real-World Tournaments & Events) a classifier trained with all features performed best. The only significant feature for this forum is the community distance (*coef* $= -1.185$ and $p < 0.05$). This indicates that posts which do not fit the topical interests of this community are less likely to get replies.

C. Results: Activity Level Prediction

To explore which factors may affect the number of replies a post gets, we first identified the feature groups which lead to the best model for each community forum (see Table III) and then analysed the statistically significant coefficients of the best performing model from each community.

Interestingly, our results suggest that the factors that impact whether a discussion starts around a post tend to differ from the factors that impact the length of a discussion. For example for the support and advise oriented community around forum 343 (Golf) content and community features contribute most to the identification of seed posts, but focus features are most important for predicting the activity level of discussions around seed posts. This indicates that it is important that a post's content has certain characteristics (e.g. contains only few links) and fits the topical interests of the community in order to start a discussion, but afterwards it is important that the author of a post has certain topical and/or forum focus in order to stimulate a lengthy discussion in this forum. In forum 865 (HE Video Players & Recorders) the seed post identification works best when using features from all feature groups, but for predicting the activity level a post will produce a linear regression model trained with content features works best. This indicates that posts which manage to stimulate lengthy discussions in this forum share some content characteristics. Also for the community around forum 544 (Banking & Insurance & Pensions) which also has an advice seeking purpose a model using all feature groups performs best in the seed post identification task. However, for predicting the length of discussions which a seed post will generate a model trained with community features only ranked the posts most accurately according to their discussion length. This suggest that in this forum it makes a difference who authored a post and how this person relates to the community when predicting the discussion length around a post. For the topical community around forum 55 (Satellite) the main purpose is to share information and content and to discuss satellite television. Also in this community a model trained with all feature groups performs best in the seed post identification task. However for predicting the discussion length of seed posts a regression model trained with title features only works best. This indicates that in this community title features impact if a post will stimulate a long discussion. Our results show that seed posts with longer titles (*coef*=0.03003 and $p < 0.05$) are more likely to stimulate lengthy discussions.

For certain communities, such as the community around forum 267 (Astronomy & Space) whose main purpose is to share information and content, the same group of features, namely content features, works best for identifying posts around which a discussion will start and for predicting the length of a discussion. This indicates that in this community users' discussion behaviour is mainly impacted by characteristics of posts' content and therefore content features alone are sufficient to predict users' reply behaviour. Other factors play a minor role in this community.

For the community around forum 630 (Real-World Tournaments & Events) and the community around forum 10 (Work & Jobs) a classification model using all features performs best in both tasks, the seed post identification and the activity level prediction tasks. For the community around forum 10 (Work & Jobs) our results show that posts authored by users who replied to many other users in the past (*coef* of users' out-degree is 0.005 and $p < 0.01$) and have longer titles (*coef*=0.034 and $p < 0.01$) are more likely to stimulate lengthy discussions than other posts. One potential explanation is that posts with longer titles are more likely to attract the attention of this community and that users in this community are more likely to be involved in lengthy discussions with users who have replied to them before. For the community around forum 630 our results suggest that posts authored by users with a high inequity score are more likely to lead to lengthy discussions (*coef*=0.0015 and $p < 0.05$). This suggests that in this community rather active users who frequently reply to other community members' posts but do not get many replies themselves are most likely to stimulate lengthy discussions. It seems that users in this community are more likely to reply to posts of other users who replied to their own posts in the past. Also in this community posts with longer titles are slightly more likely to stimulate lengthy discussions (*coef*=0.04145 and $p < 0.001$). One potential explanation for that is that posts with longer titles tend to catch the attention of more users who then read the post and reply to it. However, one needs to note that although the effect is statistically significant the effect size is very small which indicates that the dependent variable (discussion length) is expected to only increase slightly when that independent variable (title length) increases by one, holding all the other independent variables constant.

Finally, in the community of practice around forum 221 (Spanish) no lengthy discussions happened within the selected time period and therefore we could not analyse factors that impact lengthy discussions.

VI. Discussion of Results

Our findings from the seed post identification experiment demonstrate that different community forums exhibit interesting differences in terms of how attention is generated and that

TABLE III
AVERAGED NORMALISED DISCOUNTED CUMULATIVE GAIN $nDCG@k$ VALUES USING A LINEAR REGRESSION MODEL WITH DIFFERENT FEATURE SETS. A $nDCG@k$ OF 1 INDICATES THAT THE PREDICTED RANKING OF POSTS PERFECTLY MATCHES THEIR REAL RANKING. POSTS ARE RANKED BY THE NUMBER OF REPLIES THEY GOT.

Forum	User	Focus	Content	Commun'	Title	All
10	0.599	0.561	0.452	0.516	0.418	**0.616**
221	0.887	0.954	0.863	0.954	0.88	**0.985**
267	0.63	0.703	**0.773**	0.6	0.75	0.685
343	0.558	**0.727**	0.612	0.634	0.572	0.636
544	0.5	0.514	0.607	**0.684**	0.461	0.574
55	0.574	0.42	0.655	0.671	**0.73**	0.692
607	0.77	0.632	0.814	0.48	0.686	**0.842**
630	0.707	0.459	0.635	0.547	0.485	**0.762**
865	0.673	0.612	**0.85**	0.643	0.771	0.796

the same features which have a positive impact on the start of discussions in one community can have a negative impact in another community. For example, our results from the seed post identification experiment suggest that a high number of links in a post has a negative impact on the post getting replies especially in communities having a supportive purpose (such as community 343 and 10). However, in the community around forum 267, which mainly has an information and content sharing purpose, the contrary is the case. Posts which tend to have many links are more likely to get replies in this community forum. This example nicely shows that the purpose of a community may influence how individual factors impact the start of discussions in a community forum.

It is also interesting to note that for support oriented forums (such as forum 865 and 544) users which seem to be rather new to a topic (i.e. have not published posts before which are topically similar to the content produced by this community) are more likely to get replies. Further, we notice that the importance of whether a post fits the topical focus of a community or not is largely dependent on the subject specificity of the community. In other words communities around very specific topics (such as the community around the sport Golf) require posts to match the topical focus of the community in order to attract attention, while communities around more general topics (such as the community around topic Work and Jobs) do not have this requirement.

In our previous work [4] we learnt a general pattern for generating attention on Boards.ie by performing seed post identification using all data from 2006, not just a selection of forums. The best performing model contained all features (user, content and focus), and indicated that the inclusion of hyperlinks was correlated with non-seed posts, while seed posts were those that had a high forum likelihood - i.e. the user had posted in the forum before and was therefore familiar with the forum. The results from our current work have identified the key differences between this general attention pattern and the patterns that each community exhibits. For instance for the 9 analysed forums, 7 perform best when using all features - similar to our previous work - while for the 2 remaining forums, one forum performs best when using content features and another when using title features. Additionally we find differences in the patterns: for forum 55 we find that the lower the forum likelihood the greater the likelihood that the user will generate attention, this being the converse of the general pattern learnt previously [4]. For forums 10 and 343 we find that an increased number of hyperlinks reduces the likelihood of the post generating attention, agreeing with the general attention pattern, while for forum 267 a greater number of hyperlinks increases the likelihood of generating attention.

Our results from the activity level prediction experiment show that the factors that impact whether a discussion starts around a post tend to differ from the factors that impact the length of this discussion. For example, in the community around forum 10 (Work & Jobs) a posts which has question marks in the title is more likely to get a reply but in order to stimulate lengthy discussions it is more important that the title of a post has a certain length rather than that it contains question marks.

It is also interesting to note that the title length is the only feature which has a significant positive impact across several communities on the number of replies a post gets. This suggests that in some communities posts with longer titles are more likely to stimulate lengthy discussions. We assume that this happens because long titles may on the one hand attract more users to read the posts and on the other hand long titles may be correlated with high quality or substantivity of posts's content. It is also likely to be an effect caused by the platform's interface, as users are presented with a list of threads in a given community each of which is listed by its title. The first piece of information, along with the username of the author, that community members see is the title of the post.

We also found a shared attention pattern between the Golf and Real-World Tournaments and Events communities, since in these communities posts which are topically distant from what these communities usually talk about are less likely to stimulate lengthy discussions. Therefore we can conclude that although most attention patterns which we identified in our work are local and community-specific, cross-community patterns also exist and can be identified with our approach.

Comparing these findings to our previously work [4] once again reveals interesting differences between the general pattern learnt across the entirety of Boards.ie for activity level prediction and the per-forum patterns that we have found in this paper. For instance in [4] the general pattern indicated that lower forum entropy and informativeness together with increased forum likelihood lead to lengthier discussions, while for forum 343 we found an increase in forum entropy to be associated with an increase in activity. For the other features none were found to be significant.

VII. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

In this paper, we have presented work that identifies attention patterns in community forums and shows how such patterns differ between communities. Our exploration was facilitated through a two-stage approach that provided novel features able to capture the community and focus information pertaining to the creators of community content.

Our results show that the attention patterns of different communities are impacted by different factors and therefore suggest that these patterns may only be valid in a certain context and that the existence of global, context-free attention patterns is highly questionable. In our previous work [4] we focussed on identifying global attention patterns and found amongst others that posts including links are less like stimulate discussions. In this work we show by analysing attention patterns of individual communities that this global attention pattern is only valid for certain forums. The global attention patterns one learns heavily depend on the mixture and constitution of the sample of communities which one analyses. Therefore, we can conclude that *ignorance isn't a bliss* since understanding the idiosyncrasies of individual communities seem to be crucial for predicting which post will catch the attention of a community and manages to stimulate lengthy discussions in a forum.

We found for example that in support-oriented or advice seeking communities posts which contain many links in their content are less likely to get replies, while in information and content sharing oriented communities a high number of links may even have a positive impact and make posts more likely to attract the attention of such a community. Further we observed that in support-oriented communities especially posts authored by newbies tend to be more likely to get replies. This suggests that the purpose of a community impacts which factors drive the reply behaviour of this community. Beside the purpose of a community we also found that the specificity of the subject of a community may impact which factors explain the discussion behaviour of a community. Communities around very specific topics require posts to fit to the topical focus of the community in order to attract attention while communities around more general topics do not have this requirement. Finally we also found that the factors which impact the start of discussions in communities often differ from the factors which impact the length of discussions.

Although our work is limited to a small number of communities on one message board platform, Boards.ie, it uncovers an interesting problem: the problem of identifying the context in which attention patterns may occur. In our work we use the number of replies a post gets to assess how much attention it attracts. However, we want to point out that the number of replies is just a proxy metric and other metrics such as the number of views could be used as well. Since these metrics tend to be correlated we believe that using other proxy metrics would lead to similar results.

Community managers and hosts invest time, effort and money into providing a community which is useful and attractive to its users. By understanding what factors influence community attention patterns, we can provide actionable information to community managers who are in desperate need for systematic support in decision making and community development. We hope that our research is a first step towards analysing the context in which certain types of behavioural patterns hold. Our future work will further investigate the context of attention patterns in different communities by clustering communities according to the factors which are best for predicting which post will get the attention of a community.

ACKNOWLEDGMENT

Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. The work of Matthew Rowe and Harith Alani was supported by the EU-FP7 project Robust (grant no. 257859).

REFERENCES

[1] J. Cheng, D. Romero, B. Meeder, and J. Kleinberg, "Predicting reciprocity in social networks," in *he Third IEEE International Conference on Social Computing (SocialCom2011)*, 2011.

[2] D. Sousa, L. Sarmento, and E. Mendes Rodrigues, "Characterization of the twitter @replies network: are user ties social or topical?" in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 63–70. [Online]. Available: http://doi.acm.org/10.1145/1871985.1871996

[3] M. Rowe, S. Angeletou, and H. Alani, "Predicting discussions on the social semantic web," in *Extended Semantic Web Conference*, Heraklion, Crete, 2011.

[4] ——, "Anticipating discussion activity on community forums," in *The Third IEEE International Conference on Social Computing*, 2011.

[5] H. Rangwala and S. Jamali, "Defining a Coparticipation Network Using Comments on Digg," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 36–45, 2010. [Online]. Available: http://dx.doi.org/http://dx.doi.org/10.1109/MIS.2010.98

[6] 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*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 57–58.

[7] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," in *WebSci '11: Proceedings of the 3rd International Conference on Web Science*, 2011.

[8] S. A. Macskassy and M. Michelson, "Why do People Retweet? Anti-Homophily Wins the Day!" in *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Menlo Park, CA, USA: AAAI, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2790

[9] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[10] T.-A. Hoang and E.-P. Lim, "Virality and susceptibility in information diffusions," in *ICWSM*, 2012.

[11] J. Chan, C. Hayes, and E. Daly, "Decomposing Discussion Forums using Common User Roles," in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Apr. 2010.

[12] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[13] H. M. Wallach, "Structured topic models for language," Ph.D. dissertation, University of Cambridge, 2008.

[14] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proceedings of NIPS*, 2009. [Online]. Available: http://books.nips.cc/papers/files/nips22/NIPS2009_0929.pdf

[15] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill, 1952.

[16] B. McKelvey, "Quasi-natural organization science," *Organization Science*, vol. 8(4), 1997.

[17] J. Adams, "Inequity in social exchange," *Adv. Exp. Soc. Psychol.*, vol. 62, pp. 335–343, 1965.

[18] C.-F. Hsu, E. Khabiri, and J. Caverlee, "Ranking Comments on the Social Web," in *Computational Science and Engineering, 2009. CSE '09. International Conference*, vol. 4, August 2009.