Predicting News Popularity by Mining Online Discussions

Georgios Rizos
CERTH-ITI
Thessaloniki, Greece
georgerizos@iti.gr

Symeon Papadopoulos
CERTH-ITI
Thessaloniki, Greece
papadop@iti.gr

Yiannis Kompatsiaris
CERTH-ITI
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

The paper presents a framework for the prediction of several news story popularity indicators, such as comment count, number of users, vote score and a measure of controversiality. The framework employs a feature engineering approach, focusing on features from two sources of social interactions inherent in online discussions: the comment tree and the user graph. We show that the proposed graph-based features capture the complexities of both these social interaction graphs and lead to improvements on the prediction of all popularity indicators in three online news post datasets and to significant improvement on the task of identifying controversial stories. Specifically, we noted a 5% relative improvement in mean square error for controversiality prediction on a news-focused Reddit dataset compared to a method employing only rudimentary comment tree features that were used by past studies.

Keywords

news popularity prediction; online discussion analysis

1. INTRODUCTION

The automatic identification of popular online news posts has been shown to be a good proxy to newsworthiness, when compared to content annotated by experts in Twitter [9] and Reddit [10]. News or content aggregation websites use various scores, such as upvote, view and favorite counts, to quantify different aspects of popularity for an online item. This is important for search and recommendation, as well as for providing trending analytics and improving ranking and categorization of content in order to increase visibility and promote discussions on the website. From a news gathering standpoint, it is useful to identify highly scoring, i.e., trending or viral multimedia items, as well as controversial and discussion raising content, especially early in the lifetime of a post. In this study, we focus on the prediction of online post popularity on news-related subreddits and the technological news aggregators SlashDot and BarraPunto by mining online discussions around an online story.

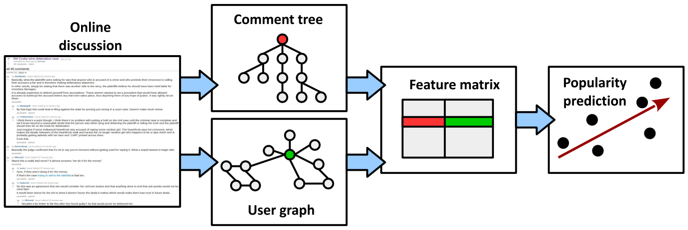


Figure 1: Feature extraction overview

Past attempts for news post popularity prediction or online discussion size prediction focused on modelling the time series of the same observed value to be predicted. However, this approach disregards a powerful source of information: social interactions that originate from the initial post, e.g., share trees of users that are formed on Facebook or Twitter when an online item is reposted. This study focuses on mining and analyzing social graphs formed by the comments that contribute to online discussions. Such analysis offers additional insights into other attributes characterizing news stories, such as their *controversiality*.

There have been past efforts focusing on the prediction of hashtag popularity on Twitter by analyzing adoption graphs [3, 16, 1]; yet, their adaptation to the problem of news post popularity prediction has been limited. These studies focused only on the prediction of a partially observable variable (i.e., adoption graph size), whereas the identification of sensational, newsworthy content may be related to additional characteristics, such as upvote score or controversiality. Additionally, no past approach has sufficiently leveraged the structure of social interactions as an additional signal for the prediction task. Given the comments in a discussion thread, we consider *two* social graphs: the *radial comment tree* formed with respect to the initial post and the *user reply graph* formed around the original poster.

We address online news popularity prediction by extracting features from both the comment tree and the user graph (see Figure 1) that represent an online discussion. We consider graph- and information-theoretic features, new to the popularity prediction problem, that characterize several aspects of the discussions, such as: tree complexity and user recurrence. We validate our approach on three news-oriented datasets from Reddit, SlashDot and BarraPunto, and produce models that predict various popularity indicators such

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890096 .

as counts of comments and users, as well as score and controversiality. Our purely graph structural approach is meant to learn which graph features are reliable indices of popularity within a news site. In particular, this study makes the following contributions:

- Improved news popularity prediction by combining information from the comment tree and user graph, and by considering graph invariant features for representing these graphs.
- Extensive experimental comparisons between the proposed method and past approaches demonstrating that the proposed graph-based features significantly improve the prediction of controversiality.

2. BACKGROUND

2.1 Learning to predict news post popularity

As a discussion about a news post grows, it attracts several actions by readers exposed to it. Apart from the comments and the readers that contribute to the discussion, the post gathers upvotes, downvotes, and views, depending on the Online Social Network (OSN) under examination. We call these social signals the *prediction targets* and we denote them by y. Each time we collect them using the API provided by an OSN, we get the value at the time of the request and not the history of values. After a certain saturation point after the item is posted, we assume that the target values are stable. We call the elapsed time since an item's posting, the *lifetime* t of the item and we denote by t_∞ the lifetime at the late, stabilized stage.

In this study, we assume that there is a *training* set of news posts, for which we have collected the comment interactions at lifetime t_∞, along with all the prediction targets $\mathbf{y}_{train}$. Suppose now that we collect the discussion content for a new *test* news post at an early point $t_* << t_\infty$ in its lifetime and we want to make some sort of inference on the future values of y_{test} at t_∞. We learn a hypothesis h based on the feature matrix extracted from the social interactions of the training set $\mathbf{F}_{train}$ and infer y_{test} by evaluating the hypothesis on the test post feature vector $\mathbf{f}_{test}^{(t_*)}$.

Since we can only extract the test features at t_*, it makes sense to compare them to training features $\mathbf{F}_{train}^{(t_*)}$ at a point in their lifetime equal to t_*. Since each comment collected is accompanied by a timestamp, we are in a position to extract features $\mathbf{F}_{train}^{(t_*)}$ at any lifetime $t_* < t_\infty$. We form $\mathbf{F}^{(t)}$ by extracting engineered features from social graphs as will be further described in Section 3.

2.2 Social interaction graphs

Let c_{t_0} be an online post, u_0 the user who posted it and $t_0 = 0$ the lifetime of the post at the time of its creation. Since we intend to work with online social interactions around the post, we call c_{t_0} the *cascade root* and u_0 the *original poster*. Let t_k be the lifetime of the post when the k-th comment is added to the discussion. At this point we also define the function $\text{user}(c_{t_k})$ that returns the user who posted the k-th comment, e.g. $\text{user}(c_{t_0}) = u_0$.

Comment tree: Let us define the *comment tree* $T(C, R)$ representing the structure of the comments associated with the cascade root. Let C be the set that contains the cascade root and all associated comments c_{t_k} (either directly responding to the root or indirectly associated with it through a number of intermediate replies). Following that, R is the set of directed edges that denote the reply direction, i.e. from the new comment, to an older one. The comment tree is thus a directed radial tree of size $|C|$, with the cascade root as its initiation point. We denote each distinct depth level in the tree as d and the set of total distinct depth levels as D. Let $\text{depth}(c_{t_k})$ be the function that returns the depth of comment c_{t_k} or a set of same-depth comments, where we define that $\text{depth}(c_{t_0}) = 0$. We finally denote by $\text{dist}(c_{t_i}, c_{t_j})$ the distance between two vertices on a tree/graph.

User graph: We denote by $G(U, E)$ the *user graph* that is formed by the users' interactions via their replies. We denote by U the set of users that are implicated in the discussion and by E the set of directed edges that represent one user's reply to another within the scope of the discussion. $U(i)$ is the total degree of vertex i in the user graph G. We further denote by $\text{deg}_{in}, \text{deg}_{out}$ the vectors that contain for all users their in- and out-degrees respectively. We do not include users that are anonymous or have deleted accounts.

As far as we know, this is the first time that a comment tree and the underlying user graph are *both* leveraged for online content popularity prediction (see Figure 1).

2.3 Related work

We describe three areas of related work, while highlighting the novel aspects of our own study.

News story popularity prediction: There have been time-series based approaches for news post discussion size prediction, such as the recent study performed in [14]. The approach described in [15] utilizes a hand-crafted feature set for predicting discussion size. Both aforementioned approaches do not take into account any structural properties of the discussion and adopt rudimentary notions of popularity, e.g., the discussion size. Finally, the study by Lee et al. [11] is the most similar to our own, as it employs feature engineering in order to quantify several online discussion characteristics, albeit not focusing on news post popularity prediction and including only rudimentary attributes of a comment tree. None of these approaches make any attempt to capture the complexities of discussion-based user graphs and comment trees for analysis.

Popularity prediction in general: Related studies that address online popularity prediction focus on classifying Twitter hashtags [16, 1] and Facebook posts [3] as popular/not-popular according to hashtag adoption counts and share counts respectively. In both cases graph-based features are used to capture qualities of an underlying, global user graph based on the users present *only* in the adoption process, as well as additional signals, e.g. the geolocation of the pertinent tweets [1], and hence are not directly applicable and comparable to the online discussion setting that we consider.

Online interaction graph mining: The authors in [6] assume that the comment tree Hirsch-index is a proxy for discussion controversiality in SlashDot. Similarly, the study in [8] utilizes a variation of the Hirsch-index, this time as a proxy for user recurrence in discussions. Finally, the authors of [5] perform a simulation study that indicates that the share graph Wiener index is a proxy of the inherent quality of a post. These studies however, do not address popularity prediction and do not experimentally show these indices to be successful popularity predictors. We include them in our methodology, along with additional graph invariants.

Table 1: Comment Tree Features

Feature	Definition
comment_count [11]	$f_{d_0} = \lvert C\rvert$
max_width [11]	$f_{d_1} = \max\limits_{\forall d \in D}(\lvert d\rvert)$
ave_width [11]	$f_{d_2} = \frac{1}{\lvert D\rvert} \sum_{\forall d \in D} \lvert d\rvert$
max_depth [11]	$f_{d_3} = \max\limits_{\forall c_{t_k} \in C}(\text{depth}(c_{t_k}))$
ave_depth	$f_{d_4} = \frac{1}{\lvert C\rvert} \sum\limits_{c_{t_k} \in C} \text{depth}(c_{t_k})$
depth_width_ratio_max	$f_{d_5} = \frac{f_{d_1}}{f_{d_3}}$
depth_width_ratio_ave	$f_{d_6} = \frac{1}{\lvert D\rvert} \sum\limits_{\forall d \in D} \frac{\text{depth}(d)}{\lvert d\rvert}$
hirsch_index	$f_{d_7} = \max(\text{depth}(d)) : \lvert d\rvert \geq \text{depth}(d)$
wiener_index	$f_{d_8} = \frac{1}{\lvert C\rvert(1-\lvert C\rvert)} \sum\limits_{i \in C} \sum\limits_{j \in C} \text{dist}(i, j)$
randić_index	$f_{d_9} = \sum\limits_{i,j \in R} \frac{1}{\sqrt{C(i)C(j)}}$

3. FEATURE ENGINEERING

This section presents the proposed feature engineering approach to form the set of features $\mathbf{F}^{(t)}$ used for prediction. The fully *invariant* graph representation is still an open problem; however, there do exist sets of indices that attempt to approximate such a description. As such, we will use a number of such indices that we specifically propose for addressing this problem. We divide the features in three groups: comment tree, user graph and temporal features.

3.1 Comment tree features

The features used for representing a comment tree are presented in Table 1. The ones that do not have a citation next to them, are proposed here for the problem of popularity prediction. We expect that a *discussion in which each comment contributes instigates further discussion will produce a highly branched comment tree*. Hence, we propose the use of the Randić index [12] due to its superior properties as shown in a recent survey of complexity indices [13]. We further include the Hirsch index that was assumed to be a good proxy for controversiality in [6] and similarly the Wiener index for virality in [5]. Certain features that describe basic attributes of a comment tree have been used before in a discussion quality prediction study [11].

3.2 User graph features

The features by which we represent user graphs are enumerated in Table 2. Apart from the user count, the rest of the features have not been used for popularity prediction before. We use four information-theoretic features based on the user graph, with the purpose of *quantifying whether the replies are made by or are targeted to a certain number of dominant users or are more uniformly distributed*. These are based on the statistical entropy (denoted by $H(\mathbf{v}) = \sum \mathbf{v} \log \mathbf{v}$) of the in-/out-degree vectors $\text{deg}_{in}, \text{deg}_{out}$ normalized to one. The normalized entropy measures are an attempt to take into account the difference of the observed frequency distribution with respect to a theoretical maximum entropy distribution, given the number of observed comments. We assume as maximum entropy to be the entropy calculated from a distribution that is formed by trying to fill a uniform distribution given k observed comments. We call the latter uni_k distribution. Finally, we do not use the Wiener index in the case of user graphs due to its high computational complexity when calculated for general graphs, but we include the Randić index.

Table 2: User Graph Features

Feature	Definition
user_count [11]	$f_{u_0} = \lvert U\rvert$
comment_hirsch_index	$f_{u_1} = \max(\lvert U_{sub}\rvert)$ $: \forall i \in U_{sub}, \lvert U(i)\rvert \geq \lvert U_{sub}\rvert$
randić_index	$f_{u_2} = \sum\limits_{i,j \in E} \frac{1}{\sqrt{\lvert U(i)\rvert\lvert U(j)\rvert}}$
outdegree_entropy	$f_{u_3} = H(\text{deg}_{out})$
indegree_entropy	$f_{u_4} = H(\text{deg}_{in})$
norm_outdegree_entropy	$f_{u_5} = \frac{H(\text{deg}_{out})}{H(uni_k)}$
norm_indegree_entropy	$f_{u_6} = \frac{H(\text{deg}_{in})}{H(uni_k)}$

Table 3: Temporal Features

Feature	Definition
first_half_time_diff_ave [3]	$f_{t_0} = \text{ave}(\Delta t_{first})$
second_half_time_diff_ave [3]	$f_{t_1} = \text{ave}(\Delta t_{last})$
time_diff_std [3]	$f_{t_2} = \text{std}(\Delta t)$
post_lifetime [11, 3]	$f_{t_3} = t_k - t_0$

3.3 Temporal features

We leverage a number of features that have been used in prediction tasks in the past [3]. We retain a list of the lifetime differences between each successive comment including the cascade root, which we denote by Δt. We utilize two chunks from this list denoted by $\Delta t_{first}, \Delta t_{last}$, corresponding to the first and the last half of the comment set. Finally, the post's current lifetime is defined as the difference between the timestamps of the latest comment and the cascade root. The temporal features are listed in Table 3.

4. EVALUATION

4.1 Dataset

4.1.1 Data collection

We tested our methods on three online news post datasets collected from Reddit, SlashDot and BarraPunto. The latter two have been used in discussion evolution analysis in [7]. We will now describe the collection of the *RedditNews* dataset. Initially, we collected as many Reddit post metadata from the year 2014 as possible resulting in 33.1 million Reddit posts[1] out of the total 54.9 million[2]. Due to the extensive time required to collect a full discussion tree for all posts, we sampled randomly and collected the discussions for approximately one million Reddit posts, of which we use in this study 35,844 discussions that were posted in news-related subreddits. Table 4 shows some basic statistics for the datasets we used and Table 5 shows the number of discussions from each subreddit we targeted.

4.1.2 Prediction targets

The prediction targets in the final stage of the discussion (i.e., at t_∞) are: a) number of **comments**, b) number of

[1] We had enhanced access of a Reddit user agent, courtesy of the DERP Institute (`http://derp.institute/`).

[2] `http://expandedramblings.com/index.php/reddit-stats/3/`

Table 4: Dataset statistics. $|C|$ is the number of vertices in a comment tree, $|U|$ the number of vertices in a user graph and ave(t_∞) is the average time required for a post to gather 99% of its total comments.

Dataset	N	min $\|C\|$	max $\|C\|$	min $\|U\|$	max $\|U\|$	ave(t_∞)
RedditNews	35,844	1	15,378	0	5,152	17.2 hrs
SlashDot	9,222	3	1,568	3	1,031	81.2 hrs
BarraPunto	7241	2	842	1	180	103.3 hrs

Table 5: News subreddits

Subreddit	N	Subreddit	N	Subreddit	N
POLITIC	7,827	betternews	5,566	worldnews	3,055
news	2,299	conspiracy	2,007	FIFA	1,987
politics	1,744	soccer	1,583	nba	1,537
realtech	1,532	technology	1,263	googlenewsfeed	1,164
worldpolitics	1,146	hockey	1,012	nfl	919
science	616	hackernews	587		

users, c) vote **score** (RedditNews only), and d) **controversiality** (RedditNews only). Whereas the score and controversiality targets have certain definitions in Reddit, we define them in a somewhat different way, in order to regularize the proportion of positive votes with the uncertainty of a small number of upvotes/downvotes. So if we denote by $z^2_{95\%}$ the 5% quantile (95% confidence) of the normal distribution, we can calculate the lower bound of Wilson score confidence interval as shown in Equation 1:

$$lower\_bound = \frac{\hat{p} + \frac{z^2_{95\%}}{2\cdot tot} - z^2_{95\%}\sqrt{{}^{[\hat{p}(1-\hat{p})+\frac{z^2_{95\%}}{4\cdot tot}]}/_{tot}}}{1+\frac{z^2_{95\%}}{tot}} \qquad (1)$$

where $\hat{p} = {}^{v_u}/_{v_u+v_d}$, $tot = v_u+v_d$ for score and $\hat{p} = \min(v_u, v_d)$, $tot = \text{floor}({}^{v_u+v_d}/_2)$ for controversiality calculation given that v_u, v_d are the number of upvotes and downvotes.

4.1.3 Supervised regression

We used Random Forest [2] regression for learning the hypotheses for their natural handling of inhomogeneous features. We minimize the mean squared error (**MSE**), use 50 trees in each forest and for each tree we consider the full number of features when searching for the best split. As for the evaluation, we use MSE.

4.2 Experiments

We perform two series of experiments in which we compare the discussions at equal lifetimes t_*. We first calculate for a dataset the mean final lifetime ($\bar{t}_\infty$) that is required for discussions to reach 99% of their final comment count, and report the MSE for lifetimes equal to $t_* \in \{1\% \cdot \bar{t}_\infty, 2\% \cdot \bar{t}_\infty, ...14\% \cdot \bar{t}_\infty\}$. Note that the RedditNews dataset includes discussions with zero comments, which leads to a rather small mean final lifetime. We use 10-fold cross-validation in all experiments. The code and pre-processed data for the experiments can be found on GitHub[3].

4.2.1 Predictive value of graph-based features

In the first series of experiments, we assess whether the addition of the proposed graph features leads to improvement compared to existing ones. The methods we use are: a) **mean** target baseline, b) only comment and user count features at t_* (**n_c+n_u**), c) only **simple graph** features by [11], and d) all comment tree and user graph features (**all graph**). The results are shown for RedditNews comment and user prediction in Figure 2a and for score and controversiality in 2b, while for SlashDot they are shown in Figure 2c, and for BarraPunto in Figure 2d.

We expect a correlation between early and late comment/user counts and as such the *n_c+n_u* method to be a significant baseline in comment and user prediction. Indeed, although we see that *all graph* outperforms the two baselines in all cases except the smaller lifetimes in Reddit comment and user prediction, the addition of simple graph features in *simple graph* leads to marginally better (if any) performance than *n_c+n_u*, or in the case of BarraPunto and RedditNews score/controversiality a *decrease* in performance. That being said, *all graph* has a clear lead in RedditNews score/controversiality prediction, which means that the new features are very potent predictors of the non-observable qualities that are related to user votes. The small mean final lifetime in RedditNews are expected to lead to rather "jagged" prediction curves. This explains the difference in how the curves look in comment/user prediction among RedditNews and SlashDot/BarraPunto. However, it reinforces the fact that the proposed features bring improvements in non-observable prediction targets by the clear lead of *all graph* in RedditNews score/controversiality prediction (also applies to the next series of experiments).

4.2.2 Feature modality comparison

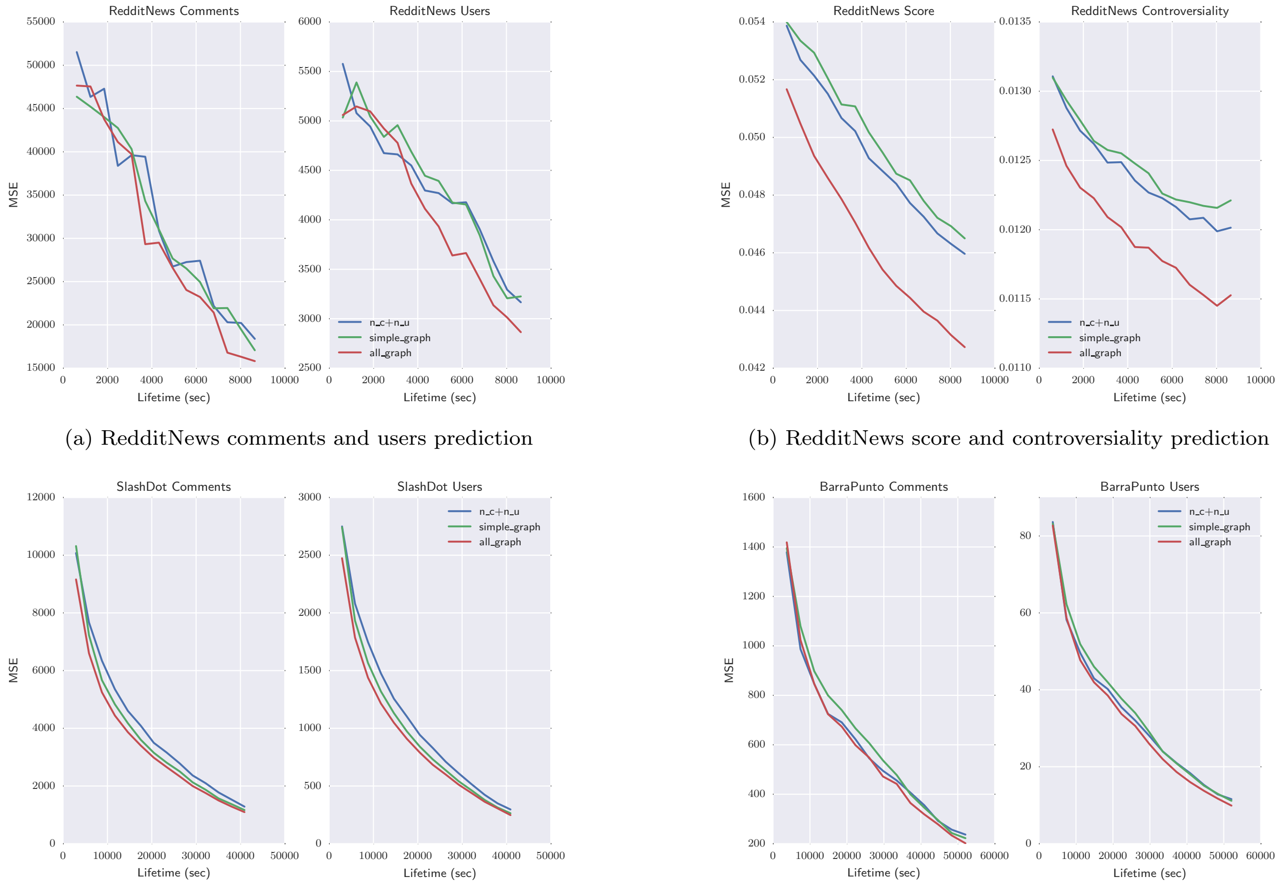
The second series of experiments compares the performance of the three different feature groups of Section 3 (**comment tree**, **user graph** and the **temporal** features from [3]), and combinations thereof, in order to highlight their predictive potential on different popularity indicators. We include the **all graph** method as before and finally the method in which we include the features from all three modalities and which we denote by **all**. The results are depicted in Figures 3a and 3b for comments/users and score/controversiality in RedditNews and in Figures 3c and 3d for SlashDot and BarraPunto respectively.

The *temporal* features are shown to perform well for comment prediction and especially in Reddit, where for the smaller lifetimes, they are actually the leading method until the point where the graphs differentiate enough. In all other cases of comment and user prediction the inclusion of our new graph features in *all* brings an improvement. As expected, the *comment tree* and *user graph* methods are important predictors for the respective comment and user prediction tasks (not so much reversely), but they never outperform *all graph* and *all*. As for score prediction, we see that *all graph* is the leading method for small lifetimes and that the inclusion of temporal features in *all* brings some improvement in larger lifetimes. Finally, we see that for controversiality prediction, the temporal features decrease the performance, as *all graph* is the leading method and *user graph* seems to be the most significant contributor.

4.2.3 Top story prediction

We additionally show the ability of our proposed methodology in predicting which stories will reach the top. We report the Jaccard coefficient (multiplied by 100) between

[3] `https://github.com/MKLab-ITI/news-popularity-prediction`



(a) RedditNews comments and users prediction

(b) RedditNews score and controversiality prediction

(c) SlashDot comments and users prediction

(d) BarraPunto comments and users prediction

Figure 2: Comparison with previously proposed features.

the top-100 most controversial stories in RedditNews and the top-100 most controversial as predicted by the *all graph*, *temporal* and *all* methods in Table 6. The random baseline would be less than 0.3%.

For example, for the relatively early lifetime $t_* = 5\% \cdot \bar{t}_\infty$, the most controversial story predicted by the *all graph* method is titled *"Gun deaths for U.S. officers rose by 56 percent in 2014: report."*[4]. We see that many users contribute to the discussion by linking to more specific information, although some disagree by claiming that the title is worded to evoke sensationalism and others that discuss how civilian gun deaths is a related and under-reported statistic.

5. CONCLUSION AND FUTURE WORK

We showed that the extraction of both the comment tree and the user graph implicated in online discussions and the extraction of features from them leads to improved popularity prediction compared to the use of rudimentary comment tree features. We further showed that the proposed features significantly outperform the competition on the problem of controversiality prediction in a Reddit news dataset. We intend to further add information modalities, such as the connectivity of underlying user friendship, subscription or reply graphs in order to capture common user preferences when predicting news story popularity or controversiality as in [4]. We wish to further examine whether there is difference in the prediction potential dependent on the topic category of a story (e.g., subreddit) or the type of the post (e.g., text post or multimedia).

6. ACKNOWLEDGMENTS

This work is supported by the REVEAL FP7 project, partially funded by the EC under contract number 610928. The authors would also like to thank Drs. Vicenç Gómez and Andreas Kaltenbrunner for sharing the SlashDot and BarraPunto datasets. Finally, the RedditNews data collection was expediated via enhanced usage of the Reddit API, courtesy of the Digital Ecologies Research Partnership Institute.

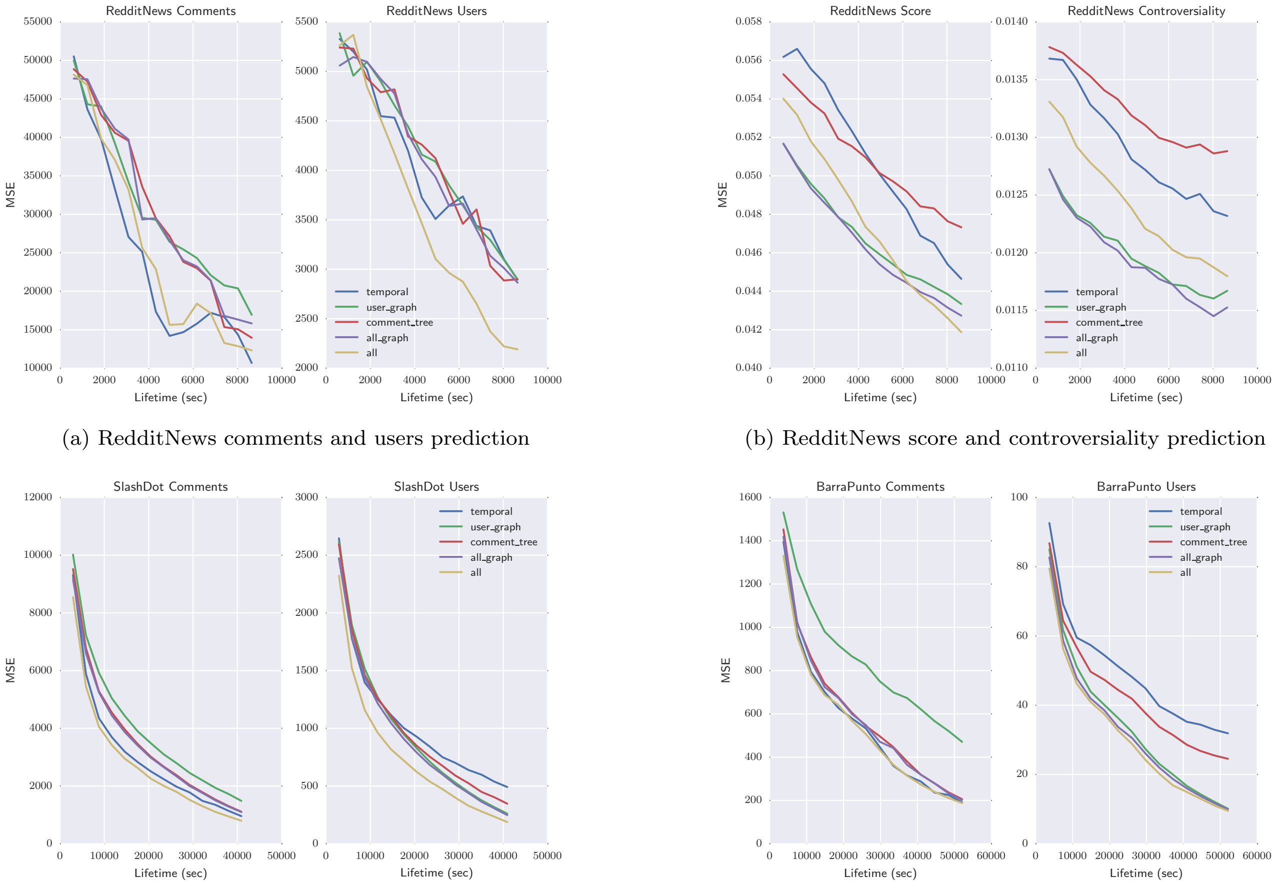
7. REFERENCES

[1] S. Bora, H. Singh, A. Sen, A. Bagchi, and P. Singla. On the role of conductance, geography and topology in predicting hashtag virality. *arXiv preprint arXiv:1504.05351*, 2015.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] 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*, pages 925–936, 2014.

[4] `https://www.reddit.com/2qtfdu/`



(a) RedditNews comments and users prediction

(b) RedditNews score and controversiality prediction

(c) SlashDot comments and users prediction

(d) BarraPunto comments and users prediction

Figure 3: Comparison among different features.

Table 6: Identifying top-100 controversial stories

	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%
all graph	**4.4**	**3.9**	**4.4**	4.4	**4.3**	**4.4**	**5.4**	**5.0**	**6.1**	**4.8**	4.7	**6.5**	**6.7**	**5.9**
temporal	2.8	2.9	3.2	4.4	3.5	4.0	4.2	3.4	3.6	3.3	3.4	3.9	4.3	4.3
all	3.3	3.7	3.8	**4.9**	3.9	3.3	3.9	4.6	4.6	4.3	**4.8**	4.9	4.9	4.7

[4] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. *arXiv preprint arXiv:1507.05224*, 2015.

[5] S. Goel, A. Anderson, J. Hofman, and D. Watts. The structural virality of online diffusion. *Preprint*, 22:26, 2013.

[6] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th intern. conference on World Wide Web*, pages 645–654. ACM, 2008.

[7] V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, 16(5-6):645–675, 2013.

[8] S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs. The structure of political discussion networks: a model for the analysis of online deliberation. *Journal of Information Technology*, 25(2):230–243, 2010.

[9] C. Hsieh, C. Moghbel, J. Fang, and J. Cho. Experts vs the crowd: Examining popular news prediction performance on twitter. In *Proceedings of the WWW'13 conference*, 2013.

[10] A. Leavitt and J. A. Clark. Upvoting hurricane sandy: event-based news production processes on a social news site. In *Proceedings of the SIGCHI conf. on human factors in computing systems*, pages 1495–1504. ACM, 2014.

[11] J. Lee, M. Yang, and H. Rim. Discovering high-Quality threaded discussions in online forums. *Journal of Computer Science and Technology*, 29(3):519–531, 2014.

[12] M. Randić. Characterization of molecular branching. *Journal of the American Chemical Society*, 97(23):6609–6615, 1975.

[13] M. Schutte and M. Dehmer. Large-scale analysis of structural branching measures. *Journal of Mathematical Chemistry*, 52(3):805–819, 2014.

[14] A. Tatar, P. Antoniadis, M. D. De Amorim, and S. Fdida. From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1):1–12, 2014.

[15] M. Tsagkias, W. Weerkamp, and M. De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 1765–1768. ACM, 2009.

[16] L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting successful memes using network and community structure. *arXiv preprint arXiv:1403.6199*, 2014.