Attention Prediction on Social Media Brand Pages

Himabindu Lakkaraju
IBM Research - India
Bangalore, India
klakkara@in.ibm.com

Jitendra Ajmera
IBM Research - India
New Delhi, India
jajmera1@in.ibm.com

ABSTRACT

In this paper, we deal with the problem of predicting how much attention a newly submitted post would receive from fellow community members of closed communities in social networking sites. Though the concept of attention is subjective, the number of comments received by a post serves as a very good indicator of the same. Unlike previous work which primarily made use of either content features or the network features (friendship links on the network), we exploit both the content features and community level features (for instance, what time of the day is the community more active) for tackling this problem. Further, we focus on dedicated pages of corporate brands on social media websites and accordingly extract important features from the content and community activity of such brand pages. The attention prediction task finds direct application in the listening, monitoring and engaging activities of the businesses that have such brand-pages. In this paper, we formulate the problem of attention prediction on social media brand pages. We further propose Attention Prediction (AP) framework which integrates the various features that influence the attention received by a post using classification and regression based approaches. Experimental results on real world data extracted from some highly active brand pages on Facebook demonstrate the efficacy of the proposed framework.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Experimentation

Keywords
Social Media Analysis, Brand Pages, Popularity Prediction

1. INTRODUCTION

The ever growing amount of user generated content and active participation over social media have recently attracted considerable public and scientific interest. A number of different studies estimate that the time spent on social networking websites is 20-25% of the total time spent on the internet. A significant portion of this social media usage is critical for businesses to listen, monitor and respond. There exist many forums, blogs, news-sites and message-boards where customers express their opinions, complaints, questions, suggestions etc. regarding products and services.

Businesses are increasingly becoming aware of this trend and devising strategies to best make use of this social media channel. It is now common to see ‘pages’ of various businesses on commonly used social networking websites. These pages are used by companies to make announcements and get feedback from customers about their products or services. Customers themselves use these brand pages for availing discounts, read reviews, access information, submit opinion etc. In some cases, companies (customers) are also using these pages for providing (accessing) customer service. Since the nature of brand page usage varies across ‘authors’ of user generated content, it is desirable for businesses to be able to prioritize the content for listening and responding. While the exact prioritization schema may depend on an application requirement, we focus on prioritizing the posts based on their popularity. Although this prediction may not entirely solve the problem of prioritization, it can serve as an important feature in the overall prioritization schema. Such prediction may be helpful for both the corporate brands as well as their customers. Companies can use this for targeted advertising, filtering, monitoring and prioritized responding while the customers can access relevant filtered content.

Our main contributions through this paper are:

- We formulate the problem of predicting the popularity of a particular post on a brand page on a social networking site.
- We present a detailed study of the features which would aid us in solving the problem. Further, we present a classification / regression based approach to solve the problem at hand.
- We carry out extensive experimentation on some of the most active brand pages on Facebook thereby demonstrating the effectiveness of the proposed framework.

2. RELATED WORK

Predicting the popularity of newly submitted online content has been of great interest to the research community.

We refer to such pages as brand pages through out this paper.
Several Twitter specific scientific studies have also studied a related problem of predicting whether a user would comment on a given blog post. They proposed the usage of latent Dricllet association (LDA) probabilistic model [1] for generation of blog posts and comments jointly. They used the information from the previous posts that a user has commented on in order to predict future responses of the user. Although related to our work, it is not easy to apply this kind of approach to social media brand pages, the reason being that it is impossible to check if each user would be interested in commenting on a particular post due to the sheer volume of the users in the community.

Several Twitter specific scientific studies have also studied a related problem of predicting whether a tweet will be re-tweeted. Most of these studies [8, 5, 11] concluded that the social features (number of followers, friends, lists etc.) have the maximum effect on retweetability. A collaborative filtering approach was used in [11] to study the retweet behaviour persistent among pairs of users. They used author information, number of the followers and number of words in the tweet for this study. A passive-aggressive algorithm was used in [5] to predict the retweetability. Although it was found in [5] that while the social features alone perform very well, there is substantial gain in also using content specific features. These content specific features comprised of novelty, trending words, and other meta information of the tweet (whether the tweet is a reply to some other tweet etc.).

As discussed above, though most of the previous work tackled the problem of predicting the content popularity, it is not straightforward to apply these existing techniques in the context of social networking sites. Thus we propose an Attention Prediction framework for dealing with this problem. This framework combines social and content specific features in order to facilitate the prediction of content popularity in online social networking sites.

3. ATTENTION PREDICTION FRAMEWORK

In this section, we first outline the problem definition and then describe in detail the approach we employ to solve the post popularity prediction problem in the context of social networking sites.

Post Popularity Prediction Problem : Given a brand page $B$ on a social networking site comprising of a set of posts $P$ of cardinality $N$ with the associated popularity measure $C$, predict the popularity $c_{N+1}$ of a new incoming post $p_{N+1}$.

Our Approach : We make use of a combination of content specific, temporal and author specific features that would enable better capture of the underlying patterns of the data, thus, facilitating better predictions.

- **Content Specific Features**: Post content plays a very important role in terms of defining what would be interesting to a community. Our Attention Prediction framework makes use of the following content based features:

- **Presence of Characteristic Keywords**: Bulk of the posts on brand pages pertain to querying information about the products/services associated with the brand. Further, people also express their opinions about various products and services. The presence of products and services in a particular post itself serves as a valuable input. However, capturing the presence of such interesting entities is a very challenging problem. In order to do this, we adopt an approach loosely based on the approach outlined in [7]. This approach primarily makes use of the principle that those words which are occurring frequently in the discussions on the brand page, but are not frequent enough in any general post (for instance, that posted by some user on his own wall) qualify as characteristic keywords. To instantiate this, let us consider the word ‘Pontiac’, it is easy to see that this word would typically appear relatively frequently on brand pages of car manufacturers but is less frequent in normal discussions or posts on any other page. Hence this would qualify as a potential keyword.

- **Presence of Sentiments**: Positive and negative sentiment expressions usually act as very good indicators of attention prediction. Whenever someone expresses a negative sentiment about a particular product/service, supporters (or refuters) of that statement tend to comment on the post. In our framework, we employ sentiment wordnet [2] to identify some of the explicit sentiment words. However, since we are dealing with microtext, where people usually tend to write short hand notations for most of the words, we also employ a loose spelling and context based approach to extract more sentiment words. According to this approach, we first identify the actual sentiment words using sentiment wordnet. We further take the appropriate context words of these words into account and identify the words which co-occur with them, these are likely to be other sentiment expressions. Finally, in order to obtain the polarity of the sentiment expression, we do a phonetic
spelling based comparison with the words which have been gathered from sentiwordnet and find the nearest neighbor in each case.

**Presence of Additional Metadata:** Additional metadata like presence of links/videos/pictures etc. is also used as a feature. During our analysis, we observed that presence of some of such metadata boosts the attention that would be garnered by a particular post.

**Post Classification:** Identifying if a given post is spam, query, opinion expression forms another crucial content based feature. Note that a post may actually belong to both query and opinion expression classes. On the other hand, a class like spam is exclusive. If a given post belongs to a query class or not is determined by rule based methods which involve detecting the presence of words like ‘can’, ‘will’ etc. at the beginning of a particular sentence and also the presence of question marks. Opinion expressions are detected using the method discussed above. Any post which does not fall into one of these two categories is put into the ‘non-technical’ category which may involve informal greetings or spam which we do not try to predict the popularity for as it is highly unlikely that they impact the brand in anyway.

- **Author Specific Features:** In most of the online discussion forums, brand pages being no exception, the identity of the author is very crucial in estimating the popularity of a post. For example, if there is an announcement from the brand page, then it is likely to garner a lot of attention. Further, if a highly influential user in the community points out a flaw in a product/service, it is highly likely that the post would become popular with a lot of people agreeing/disagreeing with the same. For our purpose, we keep track of the average number of comments a post from a particular user would receive and use this as one of the features.

- **Temporal Activity in the Community:** Finally, we turn our attention to one of the most important factors, the temporal activity in the community. This factor is very important because during certain peak times like when there is a new product being launched by a brand, there is typically a lot more activity in the forum. In order to capture this as one of the features, we make use of the average number of comments that the forum received in the preceding time interval ‘t’ (We have experimented with various values of t ranging from an hour to two days and found that a period of 6 hours was best suited).

On obtaining all the features discussed above, we learn SVM based classification and regression models².

### 4. EXPERIMENTATION

In this section, we briefly discuss the experimentation that we carried out to demonstrate the effectiveness of our **Attention Prediction** framework. In order to evaluate the task of popularity prediction, we consider the number of comments that a post receives as an indicator of popularity. So, we pose the task of attention prediction of a particular post as predicting the number of comments that a post would garner in particular temporal and contextual environment. Further, we report the performance of our framework over -

- **Classification:** In order to pose the task of attention prediction as a classification problem, we consider 5 different classes for the number of comments - Very less attention, Less attention, Mediocre attention, High Attention, Very High attention.

- **Regression:** In order to identify how much our prediction is off by the actual number of comments, we evaluate the attention prediction problem from a regression perspective.

We also train the classification/regression mechanisms to evaluate how well they perform in predicting the attention for different time periods i.e how effective is our approach to answer questions like how much of attention is this post going to receive within the first hour of it being posted.

**Datasets:** In order to validate our approach, we collected the posts and comments from three different highly active brand pages on Facebook for a period of about one week. We collected a total of 35,809 posts and their associated comments. A detailed description of the dataset is given in Table 1. For all our experiments, we have used the last 25% of the posts on each brand page as the test set. The initial 75% of the set was used for training the models.

**Baselines:** To compare the performance of our approach with the state-of-the-art, we implemented an user-centric approach as proposed in [6] and [10]. For each user, we determine if the user would be interested in commenting on a given post or not, based upon the user’s topics of interest. These topics of interest are identified using Latent Dirichlet Allocation [1] for each user.

### 4.1 Attention Prediction as Classification Problem

As explained above, we posed the problem at hand as a classification problem with 4 different classes each corresponding to the number of comments - Very less attention (<10% of the maximum of comments garnered) Less attention (<25% of the maximum of comments garnered), Mediocre attention (<50% of the maximum of comments garnered), High Attention (<75% of the maximum of comments garnered), Very High attention (>75% of the maximum of comments garnered).

**Experimentation:** After tagging the posts in the dataset with the appropriate classes (based on the number of comments received), we run the SVM classification training module using the initial 75% of the posts from each brand page as our training set. This aided us in learning the weights for each of the individual features corresponding to content, author and temporal activity. We tested the performance of our framework on the remaining 25% of the data for each brand page. We simulated the process of comments pouring in according to the time and appropriately updated the training set with the same every one hour. Also, we compare

<table>
<thead>
<tr>
<th>Brand Page</th>
<th># of users</th>
<th># of posts</th>
<th># of active users</th>
<th># of comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Motors</td>
<td>250941</td>
<td>7022</td>
<td>4056</td>
<td>18975</td>
</tr>
<tr>
<td>Bestbuy</td>
<td>3055096</td>
<td>10142</td>
<td>7986</td>
<td>35876</td>
</tr>
<tr>
<td>Nokia</td>
<td>3520071</td>
<td>18645</td>
<td>63032</td>
<td>18645</td>
</tr>
</tbody>
</table>

²http://svmlight.joachims.org/
the predictions of our approach in terms of the number of comments received by each post in the test set at intervals of 1 hour, 3 hours, 8 hours, 1 day, 2 days. We present the results for the same in Table 2.

**Discussion:** As can be seen from Table 2, our attention prediction framework outperforms the baseline in predicting the popularity of the posts at different time intervals since the time they are first posted. Analysis of this kind of behavior revealed that this problem of predicting the post popularity cannot be tackled well using user centric techniques highlighted in [6] and [10]. These user centric techniques primarily make use of individual user interests in order to predict if the user is likely to comment on a post or not. However, more often than not, whether a particular user would comment on a post or not is governed not just by the topics of interest, but also by various other external issues like time the user logs in and checks his social networking profile, how active the user is in the current period of time etc. Bearing these issues in mind, dynamic behavior of users is not always predictive of the post popularity, hence our experiments also reveal the need for aggregate techniques as opposed to user specific ones for attention prediction.

### 4.2 Attention Prediction as Regression Problem

In this subsection, we discuss the results of posing attention prediction problem as a regression problem. We again present our results in terms of the number of comments at different time intervals for each post.

**Experimentation:** We make use of support vector regression model for the training phase. As earlier, we make use of 75% of the data from each brand page for training and the remaining for testing. We use the predictive-\(R^2\) statistic to evaluate the performance. We present the results for the same in Table 3.

**Discussion:** As can be seen, our framework outperforms the baseline even in this case of regression. As in the previous case, the best performance is achieved in the case of both AP framework and the baseline when predicting the number of comments at the end of 2 days. During our analysis, we found that the task of predicting comments a post will receive at initial time intervals is much more noisy compared to the stable state prediction of popularity for each post. The same is reflected in the results in Tables 2 and 3.

### 5. CONCLUSIONS AND FUTURE WORK

In this work, we have formulated the problem of predicting a post’s popularity on brand pages in social networking sites. We further presented a detailed analysis of the features that would aid us in doing the same. We posed this as a classification and regression problem and tried to study the performance in terms of accurately predicting the number of comments a post would receive. Further, we have tried to keep our approach generic enough to be extended to most communities or discussion forums, however, on typical discussion forums, user specific interests on commenting are likely to play a bigger role. An interesting direction to this work would be to gauge how a combination of user specific interests when coupled with the aggregate analysis would perform on the task of attention prediction on social media brand pages.

### 6. REFERENCES


