User Life Cycle Analysis in the *Overwatch* Gaming Forum

David Skarbrevik
University of California, Berkeley
School of Information
skarbrevik@berkeley.edu

Abstract

Online forums, whether gaming or otherwise themed, are a popular way for communities of people to aggregate around a particular interest or topic. The social norms of individuals and communities as they exist in the real world may have many analogies in the world of online forums. Past work has shown how one's ability to adapt to community specific language exists in online forums and can be a useful predictor of whether or not a user remains active in a community. In this project I explore this linguistic property in a specific gaming forum while laying a path for anyone with forum data to do the same. The work presented here will focus on a particular gaming forum, however the analysis, as well as much of the code found in the associated GitHub repository, will apply well to any online forum.

1 Introduction

The gaming industry as of 2016 surpassed an annual revenue of 90 billion USD. With computing resources becoming more ubiquitous and technologies like virtual reality reaching a mainstream audience it is not surprising that the gaming industry sees substantial growth each year. What is surprising is how difficult it is for a game developer to keep gamers interested in their games. One particularly effective way to do this is to ensure channels for gamers to socialize with each other (whether while playing a game or not). Online discussion forums provide an outlet for gamers to express their passion for a game with others that feel the same. In the past the general model for video games was to build a finished game product, sell it for a fixed price, then move onto the next game. Now, the economy has changed such that game developers have much more opportunity for profit if they release a game early in development and continue to support/improve it over time. Because of this extended lifetime and newfound interplay between developers and consumers, it is not unreasonable to indirectly measure the success of a game by its online community (player user base, forums, competitive scene, etc.). For this reason, game developers have significant motivation to foster social interaction in their games and forums. By analyzing the language used by users, game developers can better identify users that are likely to churn.

2 Background

The work in this paper is heavily motivated by the past work of Danescu et al. 2013 [2]. In this paper, the authors use the text of forum posts to create a framework for analyzing a user within a community. The primary hypothesis is that the language of a community changes constantly over time, but the language of an individual user follows a more interesting trend. Initially as the user begins engaging with the community, their language begins to align more closely with the community. However, at a certain point, the user reaches "linguistic maturity" in the community and becomes less likely to continue to adapt with the ever changing language of the community. The idea is that this trend can be a useful predictor of user churn. By looking at the trajectory of cross-entropy between the language in a user's first few posts and the community as a whole you may have insight to the lifespan of a user.

3 Methods

Data collection (web scraping) occured on 11/18/2017 and thus represents all posts to Blizzard's official Overwatch general discussion forum from its inception until that date. It is of important note that replies to posts (i.e. users writing responses to a post) were not scraped.

To determine the linguistic distance between the community and the user, the metric from Danescu et al. 2013 was applied. This is simply the cross entropy ($H(LM, p)$) between a user's post and the community language model for the month the post was made.

$$H(LM, p) = -\sum_{i}^{n} \log_2 P_{LM}(context)$$

Then the above calculation was used for all of a user's posts (using the LM that corresponds to the month the post was made).

The creation of the language models themselves was also similar to Danescu etl al. 2013 in which two posts are randomly selected from 500 random users that month.

This data was subsequently removed from future analysis. Of difference from Danescu's paper, the language models used in this project are trigram models with add-k smoothing where k=1.

4 Results and Discussion

Because the novelty of this project is in the data that is being used (not the models that are being applied to it), the cleaning/pre-processing of the data and the subsequent analytic workflow were methodically constructed to be quite general. Thus, if one were to continually pull the most current Overwatch forum data, or if one wishes to apply a different, but similar dataset (something with time, user, topic text, and post text type features), most of the functions/workflow would be directly reusable.

Looking at the Overwatch forum dataset, we see there are a few properties that make it promising for analysis (figure 1). There is a reasonably large amount of data (posts) and many users. Unfortunately two things stick out that we would rather not see: few users make multiple posts, and the timeline of the forum as a whole is short.

Total posts in dataset	426,047
Number of users	85,308
Time frame of posts	05/2016 - 11/2017
Average number of posts per user	4.99
Number of users with over 50 posts	1,364
Average number of replies to a post	9.97

Figure 1: Basic EDA on Overwatch forum data.

One idea for remedying these issues is to simply have more data (which actually is possible in this case). As stated in the Methods, replies to posts were not scraped. In hindsight, this is a point of paramount importance because it turns out that the bulk of text in the forum is hidden in replies. As exploratory data analysis showed (figures 1 and 2), very few users post often, but most posts get a large number of replies. It is entirely possible that scraping the replies in this forum could add an order of magnitude more data. Unfortunately this will be a somewhat complex task as hyperlinks to specific forum posts are not as systematic as the pages that display just the posts themselves.

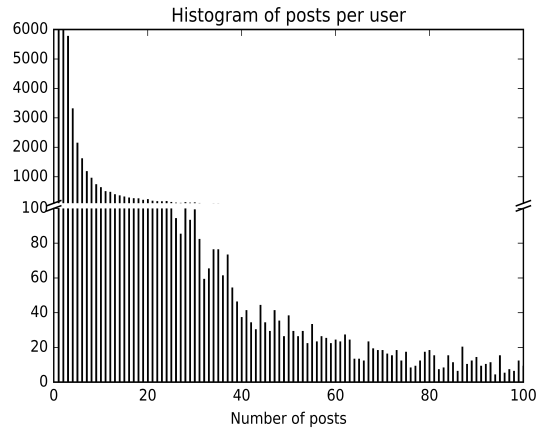

Figure 2: Number of posts per user. Broken x-axis to better show severity of skew.

Moving on to look at the final life cycle analysis that was performed on the specific user "Joe" (actual user's name) we see their posts average cross entropy relative to the snapshot models (figure 3). Upon first inspection there appears to be some interesting trends. It seems that Joe's cross-entropy lowers over the first few months which we hoped to see. It also appears that his cross-entropy begins to rise toward the end of 2017 possibly hinting at the end of his life cycle in the forum.

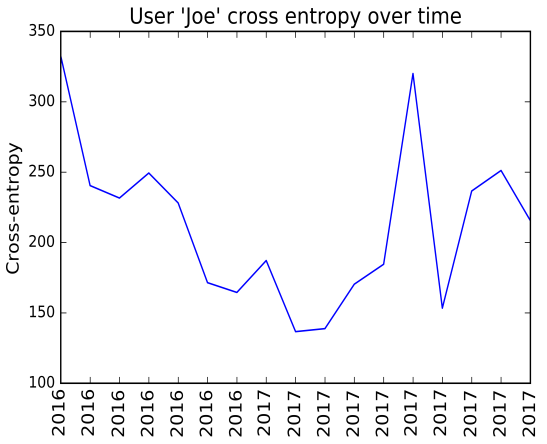

Figure 3: Linguistic distance between a specific user and the community as a whole.

This is all great but before we celebrate there are two issues. One, we need to consider how many posts are determining the cross-entropy at each month (something like standard error bars on the plot would be nice). Second, this is just one user and many other users don't show such a nice trend. However, all users that were analyzed

were still "active" in the forum. Meaning they had still posted in the most current month of the dataset.

To see if a full life cycle can be seen in this young forum we would first need to define a time frame of "abandonment". Perhaps we can say that users who made at least 50 posts and then don't post for one month have effectively churned. Though it is possible that this is too soon to consider someone as having churned. Differing points of view make this a difficult concept to portray accurately [2] [3] [5].

Because of the possibility that the young age of the forum makes it not suitable for this type of analysis, it would not be good to apply the workflow presented in this project to another forum's text. The web scraping script I wrote to gather this data is made to adapt quickly to any of Blizzard's forums so perhaps the Starcraft II or World of Warcraft forums (those with many years of history) will show more favorable results. It may also be interesting to gather a large number of subreddits and put them all through this analysis to see how well the idea of linguistic maturity holds up over many datasets. This may be feasible if Reddit's API is friendly/robust enough.

5 Conclusion

The work performed here presents a useful framework for analyzing any forum's community and users. When considering the usefulness of language models in predicting user life cycle in a forum, it is ideal for the forum to have two properties: a large percentage of users making many posts each, and long forum history (10 years would be nice :)).

The next steps for this project are to expand upon the types of analysis available. More concretely, implementation of better smoothing techniques for language models, more visualizations for whole dataset analysis and implementation of a basic classifier for identifying users that are likely to churn from the forum. Much past work has been done on this that could provide motivation [2] [4] [6]. This step is key for empowering forum owners to maintain the health of their communities. It has been shown that engaging these likely-to-churn users lead to improved retention rates [1]. As a last "nice to have" for this project, I would like to create a dashboard application that could display "at risk" users on a daily or weekly basis along with some other key stats about the health of the forum.

For code associated with this work, please see the associated github repository.

References

[1] Giovanni Luca Ciampaglia and Dario Taraborelli. Moodbar: Increasing new user retention in wikipedia through lightweight socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 734–742, New York, NY, USA, 2015. ACM.

[2] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, New York, NY, USA, 2013. ACM.

[3] Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 673–682, New York, NY, USA, 2012. ACM.

[4] Subhabrata Mukherjee, Stephan Günnemann, and Gerhard Weikum. Continuous experience-aware language model. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1075–1084, New York, NY, USA, 2016. ACM.

[5] Matthew Rowe. Changing with time: Modelling and detecting user lifecycle periods in online community platforms. In *Proceedings of the 5th International Conference on Social Informatics - Volume 8238*, SocInfo 2013, pages 30–39, New York, NY, USA, 2013. Springer-Verlag New York, Inc.

[6] Matthew Rowe. Mining user development signals for online community churner detection. *ACM Trans. Knowl. Discov. Data*, 10(3):21:1–21:28, January 2016.