When the Filter Bubble Bursts: Collective Evaluation Dynamics in Online Communities

Adiya Abisheva
ETH Zürich
Weinbergstrasse 56/58
8092 Zürich
aabisheva@ethz.ch

David Garcia
ETH Zürich
Weinbergstrasse 56/58
8092 Zürich
dgarcia@ethz.ch

Frank Schweitzer
ETH Zürich
Weinbergstrasse 56/58
8092 Zürich
fschweitzer@ethz.ch

ABSTRACT

Through the analysis of collective upvotes and downvotes in multiple social media, we discover the bimodal regime of collective evaluations. When online content surpasses the local social context by reaching a threshold of collective attention, negativity grows faster with positivity, which serves as a trace of the burst of a filter bubble. To attain a global audience, we show that emotions expressed in online content has a significant effect and also play a key role in creating polarized opinions.

CCS Concepts

•**Human-centered computing** → **Collaborative and social computing; Empirical studies in collaborative and social computing;**

Keywords

Social filtering, emotions, collective dynamics

1. INTRODUCTION

Online participatory media, such as social networking sites and forums, serve as platforms offering a vast majority of online uploaded content. User driven evaluation and rating tools, such as dislike or like buttons, and recommender system algorithms allow users not to sink in the oceans of online news, videos and images, but find content that is relevant to them and to discover those items that reached the global popularity among the online community.

On the sample of videos from `YouTube`, we observed that the collective dynamics of online evaluations follows a bimodal pattern. A video receives initial popularity and initial positive evaluations within a small community, but as the video achieves a larger viewership, negative reactions to it surge faster than in the early moments. Initial popularity in a narrow circle of users can be explained by a social network of an uploader, or by the similarity of the video with their past liked content. This is a result of the personalization of the content for users, which is achieved by the social filtering mechanisms and recommender systems. After some time, the video travels through other media beyond the local circle of early viewers and reaches a global audience. In contrast, larger public is characterised by users who are more critical and who express their negativity more frequent towards the video. We may extrapolate the observed collective dynamics of evaluations in `YouTube` videos to the dynamics of evaluations of various online content in the Internet. Filter mechanisms lead to strengthening the opinions of users and result in positivity bias towards the content, which later is surpassed by the negativity backlash. We view this abrupt shift in evaluations as a trace of the burst of the *filter bubble* [3].

In our study we use multiple online communities to shed light on collective evaluation processes. Through the relationship between likes and dislikes, we look for the existence of a positivity threshold and the following negativity rise, which we use as a quantitative evidence of the duality hypothesis of the collective evaluations.

2. DATA AND METHODS

The data used in this research is the result of our one to five year crawl of four publicly accessible online communities. Each platform specializes on different online content. `YouTube` hosts videos as a service, `Reddit` is viewed as an online news aggregator and a discussion board, `Imgur` serves as an image hosting website and `Urban Dictionary` is a crowdsourced online vocabulary of idiomatic expressions and slang lexicon. All platforms provide functionality for users to evaluate uploaded content positively and negatively by clicking a `like` or `dislike` button respectively.

To quantify emotional expression, we applied two complementary sentiment analysis methods to headers or titles of each item, leaving for a future research the analysis of longer descriptions, transcripts, and comments. First, we apply a lexicon of affective norms [6] to quantify valence, arousal and dominance of each text. Second, we apply SentiStrength [5], a state-of-the-art lexicon-based classifier, to measure a positive and a negative valence for each text. Data processing and filtering resulted in the total number of over **883 million** of likes and over **70 million** of dislikes in all four platforms.

To reveal the existence of statistical regularities of collective evaluations, we fit the distribution of the number of likes and dislikes of each platform to four known statistical distributions related to complex growth phenomena. To conclude

WebSci '16 May 22-25, 2016, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05…$15.00

DOI: http://dx.doi.org/10.1145/2908131.2908180

which parametric distribution provides the best fit available, we perform a comparative test based on the log-likelihood ratio between the two candidate distributions.

To test the existence of duality, or the presence of local and global regime, in collective evaluations, we explore the non-linear properties [2] of the relationship between the amounts of likes and dislikes for each item. We fit a continuous piecewise regression function, known as the multivariate adaptive regression splines. It identifies the knot that joins the locally linear pieces and also outputs the parameters of the two fitted equations of dislikes as a function of likes: $\mathrm{D(L)} = I + \alpha_1 * max(0, \mathrm{L} - L_c) + \alpha_2 * max(0, L_c - \mathrm{L})$, where D is the number of dislikes, L is the number of likes and α_1 and α_2 are the parameters of the two fitted equations projecting to the local and global regimes respectively. The knot L_c corresponds to the attention threshold in the number of likes, where the transition between the local to global regime occurs, or when the bubble of a local popularity bursts.

Finally, we measure the degree of polarization of an item, which is manifested by simultaneous large amounts of positive and negative evaluations. And, we investigate the influence of emotional content of items on reaching the global regime and on its polarization score. Prior to modelling, we check the sentiment scores of an item for multicollinearity.

3. RESULTS

Statistical analysis of the evaluation distributon reveal the statistical regularities in the distribution of the number of likes and the number of dislikes. For all datasets, the results of the log-likelihood pairwise comparisons of the four distributions identified the *log-normal* distribution as the best fit, which confirms early findings of the fits of the popularity distribution to the log-normal distribution [1]. This finding allows us to trace back the observed distribution of collective evaluations to the properties of the multiplicative process.

Dual regime analysis of evaluations identifies the cutoff value of likes in every dataset. These cutoff values divide the system in a local versus a global regime, with the fitted functions of the form $D \propto L^{\lambda}$ and $D \propto L^{\gamma}$ respectively. In all datasets, the exponent of the global regime is larger than the exponent of the local one, $\gamma > \lambda$, which indicates that after surpassing a threshold value of likes L_c, the dislikes given to items grow faster signalling the burst of a filter bubble. Comparison of the dual model against a single regime model yields that the proposed model outperforms the latter model in the coefficient of determination R^2 and the generalized cross-validation prediction error (GCV), validating the existence of two regimes.

Finally, the rank correlation analysis between emotional dimensions discovers a significant high correlation of over 0.7^{***} between valence and dominance. This leads us to discard dominance from the analysis of polarization and the attainment of the global attention threshold by an item. First, we model the probability of the event of an item reaching the global regime as a function of the emotions expressed in the evaluated item. The effect of activation (arousal) and pleasure (valence) are heterogeneous, showing a significant positive effect in some datasets and a weak negative effect in the other communities. The second model describes the relationship between the polarization score of an item and its emotional scores. While there is no significant effect in `Reddit`, in all the other datasets items expressing high arousal lead to a stronger polarized reaction manifested in the simultaneous large number of likes and dislikes. The model using positive and negative sentiment scores further supports this finding: items expressing higher negativity also correlate with a stronger polarization score in the same three cases as for arousal. The results of the effect of arousal and negative emotions of our observational analysis is consistent with the early theory [4] and experimental findings which links the expression of activating and negative feelings to creation of more extreme opinions.

4. CONCLUSION

Our analysis of collective evaluations across various online media shows statistical regularities in the distributions of evaluations and their relationships. First, we reveal that the distribution of collective evaluations are well fitted by log-normal distributions, which posits that possibly multiplicative process drives the generation of such distributions. Second, we find a quantitative threshold of positive evaluations after which the negativity rises faster. This serves as a trace of a burst of the filter bubble, as a strong evidence for our hypothesis on the duality of the relationship between likes and dislikes and the existence of a local and a global regime of collective evaluations. Third, we assert the role of sentimens in expressing more polarized opinions and attaining the global regime, which goes in line with the theories in psychology that attribute the polarization of opinions to emotions.

Our results reveal the emerging properties in the social online system which has been modified by the filtering mechanisms. Among such properties are soaring negativity and increasing polarization levels of discussions, when the online content goes beyond the local circle of social contacts. Sudden shifts in online behaviour, such as negative backlash, might have a damaging and far-reaching effect to user participation, but may become salient only in the long run.

5. ACKNOWLEDGMENTS

This research was funded by the Swiss National Science Foundation (CR21I1_146499/1).

6. REFERENCES

[1] S. Asur, B. A. Huberman, G. Szabò, and C. Wang. Trends in social media: Persistence and decay. In *ICWSM*. The AAAI Press, 2011.

[2] P. V. Mieghem. Human psychology of common appraisal: The reddit score. *IEEE Transactions on Multimedia*, 13(6):1404–1406, 2011.

[3] E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[4] R. Reisenzein. The schachter theory of emotion: two decades later. *Psychological bulletin*, 94(2):239, 1983.

[5] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, Dec. 2010.

[6] A. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.