Buzz in Social Media: Detection of Short-lived Viral Phenomena

Clemens Deusser
Chair of Privacy and Security, TU Dresden
clemens.deusser@tu-dresden.de

Nora Jansen
Chair of Information Systems and Information Management, Goethe University Frankfurt
jansen@wiwi.uni-frankfurt.de

Jan Reubold
Chair of Privacy and Security, TU Dresden
jan.reubold@tu-dresden.de

Benjamin Schiller
Chair of Privacy and Security, TU Dresden
benjamin.schiller1@tu-dresden.de

Oliver Hinz
Chair of Information Systems and Information Management, Goethe University Frankfurt
ohinz@wiwi.uni-frankfurt.de

Thorsten Strufe
Chair of Privacy and Security, TU Dresden
thorsten.strufe@tu-dresden.de

ABSTRACT

Social media interaction happens in a broad variety of context and magnitude. The vast majority of posts cause little to no discussion, while some start trends and become viral. We study the virality, explicitly of '*Buzzes*' - posts that evoke intense interaction over a short period of time, as they have been observed frequently, sometimes with severe consequences for individuals and companies in the physical world. Early detection of a *Buzz* may help mitigate or prevent negative consequences of large scale social media outrage against companies or persons, by giving them a chance to react at an early stage.

Collecting a labeled set of over 100,000 posts on Facebook pages, we first explore properties that define a *Buzz* using logistic regression. This method helps us to interpret the results and derive practical recommendations. We subsequently train classifiers and apply machine learning based classification techniques to demonstrate the potential capabilities of automated prediction. We achieve high recall with moderate precision, where feature boosting on broad feature sets yields the most promising results.

Our study reveals that *Buzzes* are well described by a high number of comments from previously passive users, a high number of likes given to comments, and a prolonged discussion period - properties that can be used to distinguish inconsequential posts from potentially volatile ones.

ACM Reference Format:
Clemens Deusser, Nora Jansen, Jan Reubold, Benjamin Schiller, Oliver Hinz, and Thorsten Strufe. 2018. Buzz in Social Media: Detection of Short-lived Viral Phenomena. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3184558.3191591

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191591

1 INTRODUCTION

Online social network (OSN) communication is frequently dominated by a small number of *Buzzes* - posts that *"go viral"* and spread far beyond their initial source.

Such a *Buzz* can have exceptional influence on popular discourse, and has the potential to strongly shape the public perception of products, companies, institutions, public figures, and regular individuals - both in a positive and a negative sense.

For example in 2013 Justine Sacco, a PR executive, sent a tweet before boarding a plane: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!". The associated hashtags quickly rose to the number 1 trending spot on Twitter in a global outrage. She described the aftermath as "incredibly traumatic"[1], cutting short her trip and losing her job. In another example a popular lion had been shot and killed by an American dentist[2], again prompting outrage that led to vandalism of his home and office. Whatever ones personal opinion on the actions of those individuals may be, these persecutions are irreconcilable with a lawful modern society. Private persons should not be subjected to dramatic consequences on their lives without undergoing any sort of due process or being allowed a chance to defend themselves against a worldwide mob. Recognizing the underlying dynamics of this phenomenon and giving the involved parties more time to react could help mitigate some of the fallout.

A negative *Buzz* can be aimed at corporations as well, as happened in the case of Fedex, where an employee was filmed being less than gentle delivering a package[3]. Damage to the brand in such cases can be substantial and difficult to prevent through traditional PR methods. Companies have reacted to this by employing social media managers who not only quickly react to negative online stories concerning the company with clarifications or apologies, but also by amplifying positive developments. An early warning system can be effectively used by such managers to make faster and better decisions.

As a final example, a *Buzz* does not need to be positive or negative. It can be benign as well, as happened in the case of "The Dress"[4]; a picture of a dress caused worldwide confusion as to which colors it was made of.

[1] https://tinyurl.com/zmzhu59
[2] https://tinyurl.com/wp-cecil
[3] https://tinyurl.com/mirror-fedex
[4] https://tinyurl.com/wp-dress

This study analyzes such *Buzzes*. To provide scalable early detection, we aim at detecting them using only metadata from publicly available Facebook pages. Exploring their characteristics and the respective features using logistic regression first, we subsequently employ more advanced Machine Learning (ML) techniques. Logistic regression is useful for interpretation of the results and can be used to directly derive recommendations for human decision makers, whereas the ML techniques yield superior classification results.

2 RELATED WORK

Virality has been studied frequently in recent literature in a number of formats. We will focus on publications with an emphasis on discovering or predicting virality in some form through classification techniques, as those are most relevant to our work.

Kaltenbrunner et al.[8] analyze the news site Slashdot in order to predict the number of comments a certain post will receive. Ma et al. [11] and Tsur and Rappoport [16] predict the popularity of hashtags on Twitter. Specifically they predict the number of times a given hashtag appears within a certain time frame. They find that contextual, or structural, features are more significant than content features. While this finding supports our approach of considering metadata instead of content, we did not include rich structural data in our dataset.

A number of publications are researching a phenomenon they call "cascades" [4]. The idea being that a piece of content spreads by being shared and re-shared in a cascade like fashion. Cheng et al. [2] predict the magnitude of such cascades, meaning the number of times a piece of content is shared in such a structure. Notably they avoid the problems that occur when examining extremely rare phenomena by reformulating the prediction problem such that instead of predicting the eventual size, they predict the likelihood of growing beyond the median size of all cascades that have grown at least as large.

Other works examine the virality of specific types of content. Specifically image virality has been studied in several formats. Guerini et al. [6] study visual content on Google Plus, while Deza and Parikh [3] examine the social news aggregation site Reddit and Szabo and Huberman [15] analyze the portals Digg and YouTube.

None of these publications approach the virality dynamic from a manually labeled dataset. Instead they rely on straight forward metrics, usually counting the number of shares, votes or likes, sometimes in combination with other metrics. We are not aware of publications that classify or predict virality in some form with a manually labeled ground truth.

3 THE DATASET AND ITS FEATURES

In this section we introduce our dataset, the method of collection and the derived features.

3.1 Data Collection and Description

To study *Buzzes*, and to substantiate and evaluate our claims, we started by collecting metadata from 76 public Facebook pages. We used the Facebook Graph API for this process, as it allows access to a rich set of metadata on its pages, particularly concerning posts. The specific data fields we gathered were:

- name and likes of each page,
- id, content, timestamp, likes and share count of each post and
- id, content, timestamp, likes and commenter id of each comment on these posts.

We selected these pages based on observation of online newspapers and other websites that report on viral social media events.

#pages (with *Buzzes*)	76 (50)
#posts	119,910
#comments	12,938,690
#*Buzzes* (>2 weeks)	105 (100)

We sanitized this data set, removing posts with inconsistent information reported by Facebook and pages that solely reposted content from elsewhere on Facebook. Post sanitization the final dataset consisted of a total of 119, 910 posts. To ensure enough time for the activity on a *Buzz* to develop, we also excluded all content that was posted within the last two weeks before collecting the sample. We analyzed the remaining dataset subsequently and four coders manually labeled each post as a *Buzz* or non-*Buzz*according to the following definition of *Buzzes*:

"A *Buzz* is a specific post, behaviour, or topic which initially spreads via social media and that suddenly draws surprising as well as extraordinary attention leading to many views. In this respect, a post is a text, a video, and/or a picture. It can lead to certain online reactions such as liking, sharing, commenting on, emulating, and/or offline reactions such as participating in events or purchasing a certain item. In most cases this exceptional attention lasts for a few days or weeks. In rare cases it lasts for a few months. At the beginning, in most cases there is no news agency involvement. However, later on single media outlets might report on the hype."

Thereby, more than one fourth of the data was labelled by all four coders in order to control for their reliability. All the other posts were coded by at least two persons. In the case of disagreement, two of the authors discussed and decided on the final label.

3.2 Features

In the following section we explain the different features that we collected for further analysis (Table 1 gives an overview of all the features we use).

We collect all directly related metadata of posts, like the number of likes, shares, and comments on a post (*nolikes, noshares, nocomments*), as well as the average length of the comments on a post (*contentlength*) and the number of likes on the different comments (*commentlikes*).

Considering the fact that some pages are much more popular than others, we add a feature of their division by the number of average likes, reshares, and comments for the respective page (*postlikes*, *postshares*, *comments*).

We also discovered that the number of comments in direct reaction to the post (*firstlevel*) and replies to those comments (*replies*) vary strongly over the course of the online discussion and are each thus included as features.

A *Buzz* is often characterized by many people discussing a certain topic and commenting on a given post. Holt shows that most participation in discussions is given when people debate different

General Metadata Features

pagelikes	Total Number of page likes at time of data retrieval
postlikes	Number of post likes, divided by the average number of post-likes per page
postshares	Number of post shares, divided by the average number of post-shares per page
comments	Number of comments per post, divided by the average number of comments of posts on that page
firstlevel	Number of comments that reply to the initial post
replies	Number of comments that reply to other comments
replyratio	*replies* divided by number of comments
repliedtoratio	Number of *firstlevel* comments that have been replied to, divided by number of comments
users	Number of different users that commented
oldusers	Number of users that had previously commented on the page
newusers	*users - oldusers*
newusersratio	*newusers* / *users*
repeatusers	Number of users that commented more than once
commentlikes	Mean number of likes per comments
authorcomments	Number of comments made by the post author, divided by the average number of author comments
contentlength	Mean length of the comment message string

Temporal Features

firsthourinterval	Mean number of seconds between comments in the first hour
firsthourcomments	Number of comments posted in the first hour, divided by the number of comments
lasthour	Index of the last hour containing at least 1% of the total number of comments
maxhourcomments	Number of comments posted in the hour of maximum commenting activity, divided by number of comments
maxhour	Index of the hour of maximum commenting activity
maxderivative	Maximum value of the derivative of the number of comments over time function
minderivative	Minimum value of the derivative of the number of comments over time function
completenessindex	Index of the point, in hours, after which at least a certain percentage of the total comments was posted (for 5%, 10%, 25%, 50%, 75%, 90% and 95%)
extremas	Number of minima and maxima in the comments over time function, per hour

Table 1: Feature Description

viewpoints [7]. Considering a post that deals with a controversial topic, many users will actually engage in discussing these "competing or alternative ideologies" [7]. News agencies picking up on and reporting the *Buzz* can act as an amplifier of the activity on the phenomenon, attracting additional, potentially commonly passive users to participate [13, 14]. As a consequence we also measure commonly active users (*oldusers*) and users that previously have not actively participated in discussions (*newusers*) on that page, as well as the percentage of new users (*newuserratio*).

To cover the engagement in the discussions, we count the number of users that repeatedly comment or reply on the same post (*repeatusers*), and we divide the number of their comments on the post by the average number of comments of all users on the post.

To reflect the temporal properties of *Buzzes*, we finally measure different temporal properties of the discussions. These include the frequency of interactions within the first hour after publication of a post ($firsthourinterval$), as well as the number of comments during the first hour ($firsthourcomments$) and during the peak activity of a discussion (*maxhourcomments*), each divided by the overall activity. For *Buzzes* that seem to have reached the end of their discussion we measure the number of hours it lasted (*lasthour*), and we estimate a cumulative activity distribution by calculating the hours during which certain percentiles of the discussion has taken place ($completenessindex$). Finally, we calculate simple statistics of the temporal activity, by counting the number of extreme cases (minimum and maximum number of comments per hour, $extrema$), and the maximum and minimum gradient of activities (*maxderivative, minderivative*).

4 STUDYING AND DETECTING A BUZZ

To study *Buzzes* we explore the dataset using logistic regression, first. We subsequently train common classifiers to demonstrate the feasibility of the automatic early detection of a *Buzz*, after.

4.1 Exploration using Logistic Regression

Classifying posts as *Buzzes* or non-*Buzzes*, we have a dichotomous outcome variable. We therefore estimate several logistic regression models [12]. The logistic regression estimates the likelihood of occurrence of the outcome variable depending on different explanatory variables. The regression is described as:

$$z_k = \alpha + \sum_{j=1}^{J} \beta_j * X_{jk} \tag{1}$$

where α is the Y intercept, the βs are the regression coefficients, and the Xs are a set of predictors. α and β are estimated by the maximum likelihood method.

Modeling the likelihood of occurrence is not based on a linear regression approach but on a logistic function. It is given as:

$$\pi(Y) = \frac{e^z}{(1 + e^z)} \quad (2)$$

where π denotes the probability of the outcome variable *Buzz*.

To assess their dependence, we checked the correlation among all explanatory variables and obtain a set of variables where the correlation never exceeds critical values (cmp. Table 2 for details). We measure the quality of the estimated models (cmp. Table 3) using Pseudo R^2, and we use the Bayesian Information Criterion (BIC) to determine the improvement between two models [1].

Considering the discussion of features, we estimated the first model including solely the feature "*newuserratio*", reflecting the engagement of previously passive users. Model 1 indicates that a post is more likely a *Buzz* when the numbers of newly active users interacting with it increases. It already exhibits a Pseudo R^2 of 0.2007 (cmp. Table 3).

When analyzing the results of the first model further, we discover that it indeed tends to classify certain posts of celebrities as well as live Q&A sessions and lotteries as *Buzzes*. We hence extend the second model by the mean number of likes per comment (*commentlikes*) and the number of comments by the author of the original within the discussion (*authorcomments*). The results support this approach, the second model exhibits a slight positive correlation of the two features with the dependent variable, as well as a higher Pseudo R^2 of .25 and a slightly decreased BIC. Both underline the character of lively discussions, with more reactions on the given comments than in the case of non-*Buzzes*. Thus, we can state that by adding these two further features, we obtain better results for the classification of posts as *Buzzes*.

To reflect the temporal character of short-lived activity at unexpected levels, we extend the third model by the number of posts during the most active hour of the discussion (*maxhourcomments*). The results initially seem counterintuitive, as this feature exhibits negative correlation to the dependent variable. It shows, however, that a *Buzz* receives a high number of comments throughout its entire time of existence, thus leading to low differences between the number of posts during peak and average hours. The discussion on other posts tends to exhibit more variance and hence more pronounced differences, yielding higher feature values. This fact is illustrated in Figure 1. While activity on all posts is highest at the beginning, *Buzz*-activity is sustained for longer than non-*Buzz* activity, which tends to die out within a few hours. With Pseudo R^2 increasing and BIC decreasing we conclude that adding this feature helps improve the classification.

We finally want to test if the overall activity contains supportive information. The *likes, shares, and comments* are all potential features for this purpose. However, testing them for collinearity, we find that they are not independent of each other and hence used each of them exclusively. These tests show that the number of comments has the highest explanatory power, and we hence add *comments* as a feature to Model 4. This fourth Model finally yields the highest Pseudo R^2 and the lowest BIC. It also reflects our definition nicely: it demonstrates that the overall direct and indirect activity and a constant high level of interaction with a post, which also attracts otherwise passive users are defining the *Buzz* phenomenon very well.

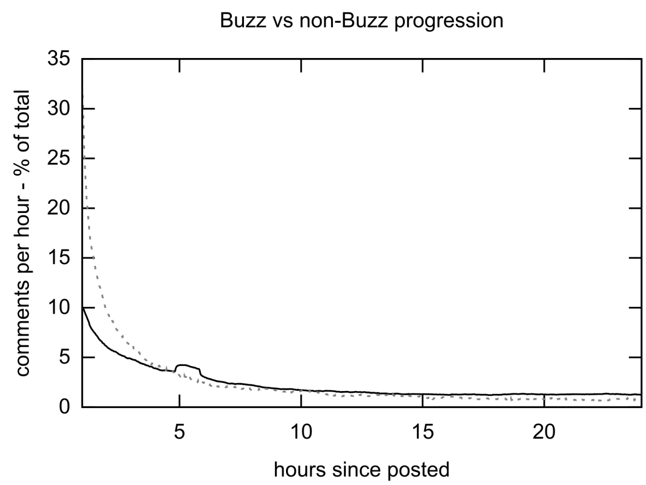

Figure 1: Comment progression of Buzzes (solid line) plotted against non-Buzzes (dotted line).

Subsequently to this exploration, we apply the fourth model for an initial classification experiment on an additional, independent test set. This test set includes almost 24,000 posts with 20 *Buzzes*. Table 4 summarizes the results.

π acts as a threshold on the classification using the models from our logistic regression. Regarding various values of π for the experiments, we naturally obtain different classification results. Choosing a rather large π of 0.2, only eight *Buzzes* are identified, whereas twelve labeled *Buzzes* are mis-classified. Only six posts are falsely classified as *Buzzes*, although they are not labeled accordingly. Thus, a large π value yields a precision of over 50% and a recall of under 50%.

Choosing a small π of 0.01 on the other hand yields higher recall at lower precision: 16 out of 20 *Buzzes* are correctly identified, which translates to a recall of 0.8. The number of false positives unfortunately is also rather high, which leads to a low precision of 0.14.

This supports our expectation of a trade-off between identifying *Buzzes* correctly on the one hand, and a high number of false positive classifications on the other hand.

Selecting an intermediate π-value of 0.05 leads to results that are better balanced. At 13 out of 20 *Buzzes* detected correctly the recall is rather high at .65, and the number of false positives is comparatively lower with a precision of 0.43.

As our data set entails an imbalance concerning *Buzzes* and non-*Buzzes* due to a true reflection of reality, we have to note that standard logistic regression is considered of only limited suitability for studies of such rare events. We hence also estimate a model using the "relogit" function [10], which is explicitly designed for such

	newusersratio	*commentlikes*	*authorcomments*	*maxhourcomments*	*comments*
newusersratio	1.0000				
commentlikes	0.1209*	1.0000			
authorcomments	0.0525*	0.0518*	1.0000		
maxhourcomments	0.2136*	0.1092*	0.0049	1.0000	
comments	0.0940*	0.0609*	0.1543*	-0.0259*	1.0000

Table 2: Feature correlation matrix (p < 0.01(*))

dependent variable: *Buzz*	Model 1	Model 2	Model 3	Model 4
newusersratio	6.898* (11.90)	7.490* (11.95)	7.890* (10.44)	5.909* (7.77)
commentlikes		0.205* (7.56)	0.247* (7.78)	0.222* (6.72)
authorcomments		0.136* (6.91)	0.128* (5.03)	0.110* (4.11)
maxhourcomments			-8.603* (-7.86)	-4.521* (-4.66)
comments				0.0734* (13.52)
constant	-11.46* (-22.19)	-12.27* (-21.52)	-10.39* (-14.40)	-10.52* (-14.77)
Pseudo R^2	0.2007	0.2518	0.3866	0.6113
BIC	1056.5	1012.9	850.1	571.2

Table 3: Summary of estimated parameters for our models ($N = 94927$) using logistic regression with p < 0.01 (*). t-statistics in parentheses; respective Pseudo-R^2 and BIC

		$\pi = 0.01$			$\pi = 0.05$			$\pi = 0.2$		
		Detection			Detection			Detection		
		Buzz	No Buzz	Recall	Buzz	No Buzz	Recall	Buzz	No Buzz	Recall
Observation	Buzz	16	4	0.80	13	7	0.65	8	12	0.40
	No Buzz	96	23867	0.9960	17	23946	0.9993	6	23957	0.9997
	Precision	0.1429	0.9998		0.4333	0.9997		0.5714	0.9995	

Table 4: Results of the Logistic Regression

rare cases. However, comparing the respective results yields either identical quality or a slight bias towards higher recall and lower precision in the case of relogit. This comparison can be considered a test for robustness, and we conclude that even the results based on the "logit" function are valid in general.

The logistic regression provides great insight into the properties that *Buzzes* exhibit, and hence provides useful insights for decision makers and business analysts. Considering the classification results we obtained, we see that there is still some room for improvement.

4.2 Training Further Classifiers

To validate our model and investigate improvements to the classification task we also used more complex machine learning techniques. Random Forests[9], Support Vector Machines and AdaBoost[5] provided the most robust results.

We optimized the hyperparameters of all our approaches using traditional grid search - a process of determining parameters through exhaustive search in a predetermined interval of possible hyperparameter values. We assigned a bias toward recall during this grid search, as avoiding false negatives is of higher priority than avoiding false positives due to *Buzz* rarity. In the case of Random Forests, we randomly selected a subset of 1,600 non-*Buzzes* and averaged the results over a number of iterations, thus removing outlier results, to account for the fact that *Buzzes* are rare events.

Training Random Forests, the results on optimized parameters showed that the most relevant features are *comments, postlikes, postshares, firstlevel, replies, users, newusers, repeatusers, pagelikes* and *extremas*. This again reflects our definition of a *Buzz* being a post with high activity and the observation of an engagement of otherwise passive users.

AdaBoost and SVMs used all available data and features, with scaled parameters.

4.3 Results and Comparison

Training these classifiers, we obtain better results compared to the logistic regression (cmp. Table 5), as expected. Random Forest, AdaBoost, and SVM all achieve a high recall of *Buzzes*, with AdaBoost also achieving moderate precision and the highest F1 score of *Buzz* recognition.

		Random Forest			Support Vector Machine			Adaptive Boost		
		Detection			Detection			Detection		
		Buzz	No Buzz	Recall	Buzz	No Buzz	Recall	Buzz	No Buzz	Recall
Observation	Buzz	18	2	0.90	18	2	0.90	16	4	0.80
	No Buzz	51	23912	0.9979	60	23903	0.9975	23	23940	0.9990
	Precision	0.2609	0.9999		0.2308	0.9999		0.4103	0.9998	
	F1	0.4045	0.9989	0.9984	0.3673	0.9987	0.9982	0.5424	0.9994	0.9990

Table 5: Results using additional classification approaches.

5 LIMITATIONS

Without contradicting previous statements or inhibiting the validity of our results, certain restrictions apply to the methods and results used in this paper. The limitations mentioned here are as exhaustive as we are aware.

Selecting Facebook pages with a higher likelihood of *Buzzes* occurring introduces a bias to the dataset where the number of *Buzzes* is overrepresented. This prevents us from using the data to deduce the overall prevalence of a *Buzz*, but should have no impact on the classification model.

In addition to the page selection, using a single OSN may raise concerns about the universal applicability of our results. It could be argued, for instance, that other OSNs may offer additional dimensions of data or present different correlations due to a different modus operandi (e.g., Twitter functioning very differently and possibly inducing different behavior). While it is true that more data from more sources would likely improve our results, choosing to base our features solely on metadata allows us some generality and Facebook enjoys the most widespread adoption among OSNs. It thus offers by far the largest user base and to the best of our knowledge its audience is no less representative of the online community than other OSNs'.

Beside the choices we made for the source of our data, the amount of data is always a possible point of contention as well. Despite the large amount of data we gathered, we were only able to identify a small number of *Buzzes*, a result of their rarity. In this case a small sample size could not be avoided and our results do show high significance.

Another concern we wish to address here is the choice of features, specifically the fact that we restricted ourselves to metadata. Facebook not only exhibits rich thread data in public pages, but also an extensive amount of structural and content data. However, Facebook does not offer easy access to said data, specifically structural data is difficult to acquire. This not only limits our efforts in retrieving such data, it also makes our results less applicable as others will likely not have ready access to such data either. Content related data, on the other hand, opens a number of complexity dimensions that are difficult and resource intensive to interpret, again restricting the applicability of our results and also yielding an entirely different contribution. Our goal in restricting ourselves to metadata, then, is primarily to offer usable results both for future research and applications.

6 CONCLUSION

This paper deals with *Buzzes* – phenomena rooted in social media, where posts receive extraordinary attention over a short period of time. *Buzzes* have frequently been observed and discussed in media, but never been defined or analyzed scientifically, to the best of our knowledge. Observing that *Buzzes* have exceptional influence on public discourse and opinion, we assert that their early detection is desirable to help prevent brand damage or even vilification of individuals and institutions. It may also aid marketing campaigns to detect and leverage positive *Buzzes*.

Upon discussing a suitable definition of *Buzzes* in relation to similar phenomena, we analyzed *Buzzes* on Facebook.

Collecting and annotating a large dataset of over 100,000 posts from 76 Facebook pages, we used logistic regression solely on the metadata of the posts to identify features that define, and can help classify posts as *Buzzes*. The results indicate that *Buzzes* indeed are characterized by intense activity over short periods of time – which varies from generally popular discussion, or *trends*, in the engagement of otherwise passive users and a rather constant activity over time (features *comments*, *newusersratio*, *commentlikes*, and *maxhourcomments*).

Understanding the characteristics of *Buzzes*, we subsequently aimed at providing better early detection by training well-known classifiers. We provided them with the overall collection of features, as opposed to the manual selection for logistic regression. It hence turned out that the number of shares and likes a post receives compared to the average of posts on the same Facebook page helps enhance the classification. Training Random Forests and SVM using grid search immediately yielded better recall, and applying Adaptive Boosting led to the best results (measured using $F1$ scores).

We see a variety of venues to extend our work in the future. Having focused on Facebook pages within this paper, we shall explore its applicability to consider more structural data, including the posts and ego networks of individuals. Also the applicability to other social media, like for instance Twitter, Instagram, and G+ raises interesting questions. First, it will help to validate the phenomenon and our approach, and it may also disclose the similarities and differences of the characteristics of discussion on different sites, focusing on different types of content. Finally, we are currently working on live, rather than the post-hoc detection of *Buzzes*.

REFERENCES

[1] A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2009.

[2] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *WWW*, pages 925–936, 2014.

[3] A. Deza and D. Parikh. Understanding image virality. In *CVPR*, pages 1818–1826, 2015.

[4] P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In *ICWSM*, 2013.

[5] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[6] M. Guerini, J. Staiano, and D. Albanese. Exploring image virality in google plus. In *Social Computing*, pages 671–678, 2013.

[7] R. Holt. *Dialogue on the Internet: Language, Civic Identity, and Computer-mediated Communication.* Greenwood Publishing Group, 2004.

[8] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Web Conference, LA-WEB 2007. Latin American IEEE*, pages 57–66.

[9] H. T. Kam. Random decision forest. In *Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August*, pages 14–18, 1995.

[10] G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[11] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 64(7):1399–1410, 2013.

[12] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1):3–14, 2002.

[13] A. Reynolds. News influence pictures of the. *Media Effects: Advances in Theory and Research*, 10(1), 2002.

[14] M. Roberts, W. Wanta, and T.-H. D. Dzwo. Agenda setting and issue salience online. *Communication Research*, 29(4):452–465, 2002.

[15] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[16] O. Tsur and A. Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *International conference on Web Search and Data Mining*, pages 643–652, 2012.