Online Engagement and Well-being at Higher Education Institutes: A German Case Study

Margeret Hall
Karlsruhe Service Research Institute, KIT
Karlsruhe, Germany
hall@kit.edu

Simon Caton
National College of Ireland
Dublin, Ireland
simon.caton@ncirl.ie

Abstract—Society is increasingly mirrored in the digital sphere. Unknown is how well this maps back in real-life aspects like the feeling of well-being or community engagement. Our tool 'BeWell@KIT' was employed to parse 140 Facebook pages comprising the online social media presence of a large public German university. Discourse attributes are established and investigated to identify the interaction between digital discourse and real-world events. We find evidence of critical system-wide events directly impacting the expressed well-being and engagement of the community, while also displaying an uptick in the psychological concepts 'belongingness' and 'resilience.' Our study indicates that digital expressions of well-being support the healthy functioning of the community at large.

Keywords—Online Social Networks; Facebook; Behavior Modeling; Sentiment Analysis; LIWC

I.   INTRODUCTION

The dividing line between off- and online communities is increasingly blurred. Digital participation and communication has become the 'new normal' [1]. Considering fast-paced online communities there is an institutional interest in knowing if, and which, events have significant effects on the way the community interacts and expresses well-being online, what changes sentiment over time. Similar works name the mapping of digital sentiment and behaviors a 'Social Observatory' [2]. An interesting extension to existing efforts is the mapping of digital interactions of a university community considering expressed well-being. We ask, '*Which attributes of digitally expressed, institutional well-being can be extrapolated from informal online text?*'. To address this, our work evaluates the performance of a university from the perspective of the university community's subjective opinion(s) of itself online, aggregating based on community well-being and expressions of engagement. Such a system establishes a more granular and sensitive feedback system for stakeholders (i.e., administration, students, faculties) to assess and respond to university performance. In response to this a Social Observatory is employed to analyze and report socially-sourced indicators on university quality and satisfaction. The Social Observatory procured data from popularly used public Facebook pages surrounding the Karlsruhe Institute of Technology (KIT) [1], for a tool that is near to real time and sensitive to concerns of both privacy and the desire to participate online.

Section II reviews related work and Section III justifies the design made in the implementation choices and gives the descriptive attributes of the KIT Facebook network. Section IV reviews the macro, meso and micro attributes of communal discourse across the KIT Facebook network. Section V discusses and contextualizes the findings and addresses limitations, and Section VI concludes the chapter.

II.   RELATED WORK

A.   On the application of Social Media Platforms for Social Sentiment Analysis

In the 1960s computational innovations resulted in a challenge shift: The restricting parameter for social researchers was no longer the collection of data. Instead, information grew at a rate faster than researchers could analyze [3]. Developments in people's daily lives are at once more transparent, yet more difficult to understand. This is due in part to the rise of networked online social data. Social media sites in particular have quickly ascended from a novelty of the early 2000's to a fact of life, and daily necessity. Users interact online by creating profiles and providing (semi)personal information in form of text, photos and other media [4]. As social networking and media platforms are generally based on true identities or variants thereof [4], [5], they are well suited for digital community analyses.

Facebook is the largest, most active platform with its 1.308 billion daily active users, with one in every seven minutes worldwide (and for Americans, one in every five minutes) spent on Facebook.[1] In an exhaustive survey, [6] summarized and classified 412 articles written on Facebook for the period 2007-2012 leading to five supra-categories: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. Recognizable is that the usage of Facebook's API by non-Facebook staff or partners to support quantitative, unobtrusive studies is low; when the referenced studies apply quantitative methods, the method of choice tends to be based in survey methods.

Notable studies from Facebook Research look at public expressions of sentiment. [7] used status updates based in the United States to create a composite well-being index. Another

---

[1] http://techcrunch.com/2014/07/23/facebook-usage-time/. Last Accessed: 12 March 2015.

series of studies by Kramer and colleagues [8], [9] reviews emotional contagion on Facebook. These studies report that emotions are indeed contagious in a network. Their findings support that short informal text like Facebook status updates can be used to measure sentiment online. To date, no studies were identified that dissect how educational institutions and their communities use, leverage, and engage computer-mediated communication with Facebook. This article addresses that research gap.

B. Linguistic Inquiry and Word Count

LIWC originally was not intended to be used on short informal text, but to analyze text of expressive and therapeutic writing sessions usually containing more content than the average tweet or Facebook update [10], [11]. However, its expansive psychometric dictionary offers a unique opportunity to reveal the latent emotional context of text-based data. The application of LIWC on documents returns the percentage of words across the categories social processes, affective processes, cognitive processes, perceptual processes, biological processes, work and achievement, as well as punctuation and structural details [11], [12]. Per cent based information gives the researcher a mechanism by which to see the relative worth of categories in speech. This facilitates measuring change, looking for group-based patterns, monitoring individual spikes and dips, and identifying psycholinguistic profiles.

When people share (written) information, there is not only content but also the way they create their message and the linguistic style [13]. They found that function words are well suited to build a systematic picture of this inconceivable dimension as latent indicators. They refer to pronouns, prepositions, articles, conjunctions, and auxiliary verbs and altogether can be imagined as "[…] the linguistic "glue" that hold content words together" [14]. While LIWC focuses on function words it also includes content words. The functionality is based on dictionaries that assign over 4,500 words to 70 different categories, ranging from a simple stylistic (e.g. article, prepositions) to a complex psychological level (e.g. positive emotion, cognitive words). Due to their near constant usage and grammatical weight, use of function words is nearly impossible to manipulate and thus will uncover motives, personality and psychological processes more accurately than analysis of the content [15]. Using computational tools in analyzing function words bears further advantages. Firstly, people's poor awareness of function words is not restricted to their own language. The listener doesn't focus on function word composition, and therefore is unable to rate usage. Hence, computational pattern matching can reveal findings not attainable by human judges. Secondly, less than 0.04% of an average persons' vocabulary are function words [13]. At the same time, they make up more than half of daily language. Consequently, function-word based analyses are well-situated to reveal latent individual states. All in all, the function word's importance on psychological findings justifies the application of its simpler dictionary-based approach wherever emphasis is set on personal traits.

III. Design and Approach

To address research questions several steps must first be taken. Following the methodology established in [1], [2] the raw data is first filtered based on post type, then aggregated to represent groups of the university; is run through LIWC, and finally mapped and assessed. In order to create a comparative baseline, LIWC scores of all data (posts and comments) before the start of the event and after its completion have been aggregated to a single number, weighted by total word counts. All measures in the coming analyses do not show the actual LIWC scores, but relative increases and decreases considering a time-local baseline. The app used to extract Facebook data is open source and can be found in [2].

The first assumption to be addressed it the use of Facebook as opposed to Twitter. The KIT database of Facebook activities features an average text length of 33.96, mainly German, words. Given Twitter's character restrictions and that the average German word length is estimated as 5.7 characters[2], this would exclude 33.57 characters of the average message or otherwise force unnatural brevity or improper spellings; consequently adding complexity and errors to the (text) analysis. The fraction of posts and comments in this procured dataset containing more than 160 letters (28 words on average) represents 80.1% of the corpus, reflecting 39.86% of all comments and posts being longer than Twitter's restriction. Twitter would certainly result in drastically shorter text submissions and consequently in a loss of more complicated, reflective statements. There is an additional restriction of Twitter that lends unknown biases, namely that Twitter grants between 1-10% of the data available from the first request date in a given query [16]–[18], compared to the full Timeline of the Facebook extraction. For KIT, Facebook usage outranked all other Social Media usage for both university-generated and student-generated content, which is in line with the fact that Facebook has an 82% market reach of Germany, whereas Twitter has approximately 20%.[3]

The 140 pages in the dataset represent all open pages which were (1) primarily populated by KIT members and (2) had posted 50 or more words between the timespan of 2011-2014. Finally, four granularities are investigated: post-comment splits, page group splits, administration-faculty splits, and temporal aspects. From this baseline it is possible to see what, if any, spikes and dips appear. Accordingly the next section describes the KIT Facebook community, establishing the attributes that make up the communal discourse.

IV. Macro, Meso, and Micro Granularities of BeWell@KIT

In order to gain a more granular understanding of how the KIT relates and interacts online, the baseline of discourse and latent emotive value are established focusing on the years 2011-2014; while some pages were open longer than this, all pages included in the study were open from 2011 onwards.

[2] http://www.duden.de/sprachwissen/sprachratgeber/durchschnittliche-laenge-eines-deutschen-wortes. Last Accessed: 10 March 2015.
[3] http://www.statista.com/statistics/280176/penetration-rate-of-social-media-sites-in-germany/. Last Accessed: 10 March 2015.

A. Macro Attributes of the KIT Facebook Network

Table I displays descriptive characteristics of the dataset. Likes far outnumber posts and comments, and posts outnumber comments. That posts outnumber comments in this use case is a surprising characteristic as most official pages only permit administrators to post on the timeline; constituent participation is restricted to commenting on those posts.

TABLE I: SUM OF VALUES OF ALL PAGES IN KIT FACEBOOK NETWORK CONSIDERING POSSIBLE INTERACTIONS OF THE PAGES AND AUDIENCES

Page Likes	Status Updates	Wall Posts	Comments	Likes on Posts	Resources Posted	Resources Liked
101,772	26,259	4,284	16,079	179,721	8,817	45,241

Given that KIT is an online community, it is expected that its members communicate in similar time spans. KIT's communal discourse has a cyclic pattern that matches recurring semester cycles: The start of semester, mid-semester, exam weeks and semester holidays. The intensity of interactions also follows this pattern closely, as approximately 66% of interaction occurs inside of the semester. Figure 1 displays the average over academic years considering the timespan 2011-2014 in per cent. There it can be seen that the bulk of discussions occur inside of the semester, with the Winter Semester having slightly more chatter than the Summer Semester. This pattern is flipped for the holiday seasons, which Summer Holidays having a slight boost in activity compared to the Winter Holidays. That remains constant when comparing the exam weeks to the holidays – Winter Holidays have less Facebook interaction than the Winter Exams, and Summer Holidays have more interaction than the Summer Exam period.
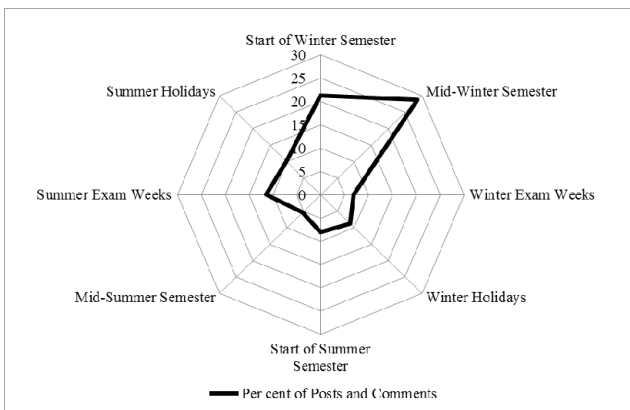

Fig 1. Frequency of posts and comments at KIT, 2011-2014

B. A Meso-assessment of KIT's Discourse Baseline

A group representation is the creation of supra-groups based on commonalities (e.g., administration, faculties, student groups) used to assess the KIT community as a more realistic replication. Regarding group partitioning, two approaches are executed. First, all the 140 available pages are assigned to one of 12 page categories in order to facilitate analyses of the university's Facebook community. The naming of the groups is guided by the KIT website where possible to assure a realistic assessment in reconstructing discourse. In the case of KIT affiliated but not KIT sponsored groups, the most general common name is used. The names of the groups are *KIT (official presence), Library, Schools, Departments and Institutes, Student Clubs, University Clubs, Sports Teams, Innovation and Development, Politics, Career, Music, and Social*. It must be noted that during the course of the study five pages closed and were duly excluded from the analysis; pages with less than 50 words over the four years of assessment are likewise excluded. These groups are further assessed considering if they are run by administrators or students. Splitting the data into these subgroups aims to reproduce an accurate picture of the community, by taking interactions and communal diversities within into account.

Comparing administrators and students reveals interesting differences in the discourse baseline. Results of an Independent Sample Mann-Whitney U test show highly significant differences in the use of Positive Emotion ($U = 6,740$, $z = -4.520$, $p = .0005$) and Negative Emotion ($U = 7,530$, $z = -3.381$, $p = .0005$), using an asymptotic sampling distribution for U. Mann-Whitney U is the non-parametric estimation of a One-Way ANOVA. Administrators show a lower frequency of positive and negative emotional discourse. When these emotions are employed, they tend to be employed by students. Net Affect, a composite variable, is calculated by subtracting negative sentiment categories from positive sentiment. KIT's network is mesokurtic with a positive skew (Figure 2a) and a reversed sigmoid distribution (Figure 2b), hovering at zero but with a long positive tail.
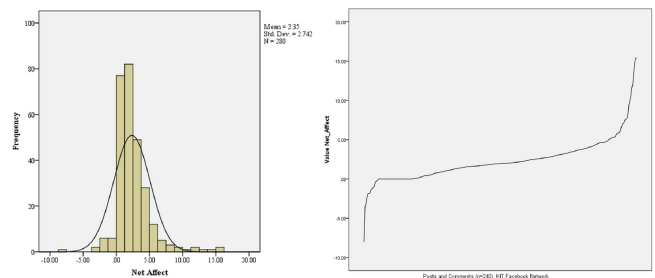

Fig. 2 Net Affect, displaying skewedness and (a) Kurtosis and (b) Distribution

That KIT's Net Affect tends to hover around zero signifies few pages employing extreme emotion. The absolute range is -8.0 from the OSKar- Optics Students Karlsruhe e.V comments to a positive 15.38 from the comments of the Institute of Regional Science. Students tend to make up both ends of the tails, and administrators are grouped in the middle of the distribution (the zero range). This supports the results of the Mann-Whitney U tests that students are move emotive than administrators in their digital discourse.

When this is considered alongside with the tendency of comments to use more cognitively expressive (Figure 3) and emotive discourse in their responses, it can be understood that although this tendency should be expected in most communities, the size of this gap indicates that the university's constituents visit the pages to seek and engage in lively

discussions. Comments display significantly higher cognitive complexity than posts (U = 5,831.5, z = -5.861, p = .005).
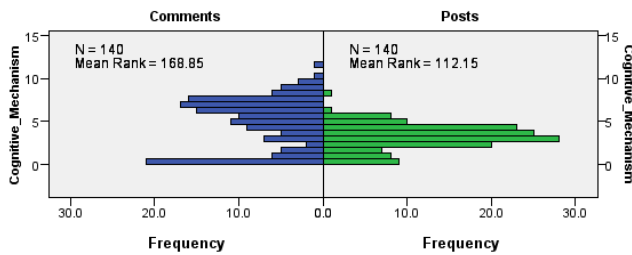

Fig. 3. Results of a Mann-Whitney U test comparing cognitive complexity

C. Micro Representations of Communal Well-being

Given KIT's cycle discourse, a reasonable way to identify events of impact is to inspect sentiment spikes and dips as they are related to the semester intervals. The following analysis addresses the benchmarks of the semester. BeWell@KIT established a critical system shock for students and employees as the denial of the Elite Status on 15 June 2012.[4] The loss acted as a shockwave across the network and was the most common discussion topic the days after the loss, as it was expected to damage the university's prestige and signaled the end of the 'Excellence Money,' a governmental support of 15-20€ million yearly. Students feared decreasing employment opportunities in the highly competitive academic working environment. At the same time, financial consequences threatened the continuing of research projects and existence of administration jobs. The denial impacted students, researchers, and administration employees likewise [1].

The Facebook community's overall activity after publication of the judges' Excellence decision increased strongly. Whilst the week before the announcement counts 7,425 words, this amount increases by one third to 11,070 words during the consecutive week and 15,072 (almost an additional 25%) two weeks after the event. The two weeks representing the event and after the event comprise 1.3% of the corpus' words. The categories reflecting cognitive complexity (Articles, Exclusion, Causation) show a positive trend in the following week of the Excellence loss compared to the overall score before (Table III). Putting this together with the significantly higher scores of Past and Future (measuring verb tense frequency), and the topic categories Money, Occupation,

TABLE III SCORE DEVELOPMENT FOR COMPARISON BETWEEN 1) ALL DATA BEFORE JUNE 15TH 2012, 2) THE FOLLOWING FIRST WEEK AFTER THE EVENT AND 3) THE FOLLOWING THREE WEEKS AFTER THE EVENT WHERE GREEN SHOWS INCREASES AND RED SHOWS DECREASES

	Before Loss	1 Week After	3 Weeks After
Articles	6.68	8.24	7.64
Exclusion	0.86	1.04	1.04
Causation	0.63	0.88	0.72
Past	1.31	1.85	1.71
Future	0.56	0.78	0.71
Money	0.72	0.89	0.68
Occupation	5.49	6.07	5.83
Job	1.89	2.06	2.04
School	2.87	3.37	3.19

[4] http://www.kit.edu/kit/english/5963.php. Last Accessed: 3 January 2015.

Job and School is an indication of intense discussion on the reasons and future impacts of the Elite denial.

The first week shows the most distinct peaks for all cases. Still, a wider timeframe post-event produces the same tendencies for all LIWC categories but Future (Figure 4). The additional three-week timeslot enables observation whether detected peaks presume or ebb away quickly. All categories except Future display significant percentage increased at the p < 0.05 level in the short-term and that Occupation and Past maintain mid-term significance at p < 0.05.
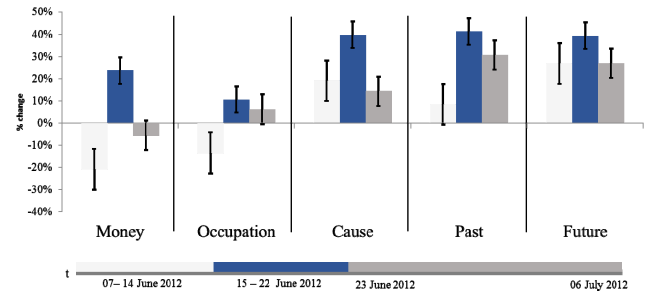

Fig. 4 Affective changes in discourse relating to the KIT Elite loss. All measures show relative changes, not absolute LIWC scores. The blue bars in the middle reflect the event week, while bars to the left (1 week before) and right (3 weeks after) represent temporal deviations from the baseline.

More than impacting professional and practical concerns, the loss of the Excellence status had a major influence on the KIT's digital expressions of well-being (Figure 5). Increased frequencies of the categories Negative Emotion and Sad hint at collective frustration. Positive Feeling depicts a decrease (-35.7%) directly after the announcement. More interesting are the categories Social and Inclusion, which are proxies for the feeling of belonging to the community of reference. The LIWC categories Social and Inclusion increase slightly after the incident, and quickly increases in the following three weeks. Here we see no significant change between the week preceding and the week of announcement. However, significant increases (p < 0.05) in social and inclusive discourse are seen in the month after. This indicates both expressions of increased community belongingness, and its related construct resilience. It is interesting to observe that after the event, zooming out to the following three weeks the categories show a slight upswing indicating communal resilience while reminding us how delicate results based on latent emotional states are.

These two categories are strong reflectors of communal belongingness, thus leading to an interesting finding. Because the loss was unexpected it affected almost all community members: the shock was wide-spread and deep. Former research found that tragic collective experiences often promote feelings of belongingness [7], [19], [20]. This is evident in the KIT dataset, where the loss of the Excellence status acted as a collective crisis. Encouragingly, the community responded with not only shock and negative feelings, but also resilience and an increase in togetherness, signs of well-being according to the definition of [21].

The Excellence initiative reaction suggests that campus-wide incidents affect the way the community interacts. Well-being is affected in the short-run, but the long-term impacts are

minimal. Belongingness increases in spite of the negative feeling overall. This highlights both community resilience, and how delicate the results are.
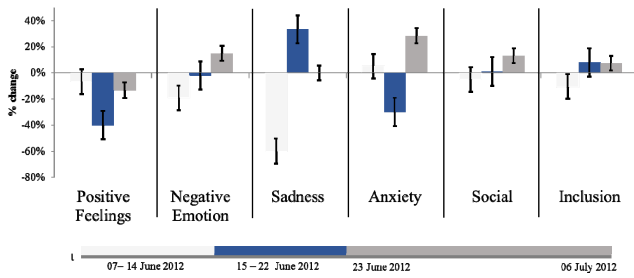

Fig. 5 Emotive sentiment flow in discourse relating to the KIT Elite loss.

V. DISCUSSION

Focusing a Social Observatory on the KIT Facebook network revealed quite clear online discourse patterns among university network members. Post-comment comparison, in which posts represent activities of page administrators and comments participation of page visitors, serve as the analysis's baseline, providing insights into the community characteristics as a whole, and as a guideline for further data partitions.

LIWC results display an overall satisfied community, disclosing indicators of high emotional and mental well-being through various emotional, attentional and cognitive categories. Interestingly, comments are both the most positive and negative aspects of the dataset, indicating that the community has a diversity of emotion even though the net effect is overall positive. To better understand the dynamics of discourse, focus was shifted to differences between comments and posts, considering if it originated on an administrator or student led page. Comments on student pages are more emotional overall. Combining this with the prevalence of cognitive processes in comments, it can be posited that a central motivation for visiting the KIT Facebook pages is seeking lively discussions and discussion of opinions. In contrast, university administrators seem to restrict themselves to 'newsflashes' in a professional, formal manner, avoiding narratives.

With respect to the temporal aspects of the analysis, several interesting patterns were detected. Campus discourse showed dependencies with the recurring semester cycles. KIT's Facebook community is most active when students are returning from holidays to the new semester. Supplementary pressure and study habits seem to reduce social activity in contrast to the middle of the semester, where social processes peak. The denial of the Elite status acted as a shockwave not only on the campus but also across the various pages of the university's Facebook community. Members reacted emotionally with anger, anxiety and sadness summarized by a generally increased density of negative emotion. Positive feelings in the community were marked by a significant drop from the week preceding the announcement. However, the community showed resilience as displayed by an increase in positive emotions, and social and inclusive discourse three weeks after the event. Remarkably, the KIT community

responded with an increase of communal belongingness to this disappointing experience.

A. Limitations and Future Work

This work focuses on spikes and dips with clear data signals in its current iteration. Innumerable smaller and unstudied incidents can add up and be responsible for emotive shifts just as well as significant and sudden dips and spikes. This is due to the fast-changing features of and in social media, including strong dynamics without distinct attributes. The long-term analysis of events seems best suited for large-scale political interventions ([22]) or small, clear communities [1].

Some limitations caused by the tools available do exist. As stated, LIWC was not designed for short informal text like that found in Online Social Media, even though it copes astonishingly well. The importance of multilingualism in Online Social Media is increasingly recognized. Interlanguage comparison or even pages including a mixture of several languages could mislead interpretation of results. To allow for consideration of these inaccuracies further software versions could process an output reflecting word count percentages of contained languages.

A major limitation of this exploratory work is its reliance on estimations of emotional states. This is especially true for dictionary-based approaches that are insensitive to context and thus will misinterpret ambiguous words and certain linguistic constructs like irony or sarcasm. Although there is a high amount of agreement with established literature to indicate this study's validity, better grounding of the dictionary to context and not only latent states would allow for more definitive statements on the general health of the community.

An interesting extension would be a comparative assessment of other universities and technical universities in Germany, as well as (dis)similar global universities. This would enable the establishment of in-depth comparisons of community characteristics and participative behavior, representing a powerful information resource for education institutions worldwide. It would also establish the findings this work as confirmatory rather than exploratory.

VI. CONCLUSION

BeWell has shown that it can detect notable community events by tracking expressed sentiment in Facebook posts and comments. This work's contribution is the binding of a multi-dimensional well-being definition that are otherwise hidden inside a data stream. To achieve this, both benchmarks from literature and unusual sentiment-based spikes and dips were observed and reported.

The results revealed by the temporal analysis indicate that within a community, stakeholders should not be identified in a top-down way. Especially the shockwaves across the digital community after the loss of the Elite status show that the community is both self-nominated, and highly engaged, participating in the events and emotions experienced as a community. Partitioning the data in recurring semester cycles presents information on how communication focus shifts over the year. Due to the fact that people frequently debate about

daily activities and events the results also capture the prevailing topics of daily activities. The way a Facebook page is administrated also seems to affect a basic indicator contributing to well-being, namely the feeling of communal belongingness. This characteristic is especially valuable for institutions since it reflects if constituents can identify themselves with values and views of the organization.

Sentiment analysis on Facebook and the KIT Facebook presences revealed multiple characteristics useful to describe a community. LIWC score interpretation allowed for the identification of the community's well-being, belongingness, resilience and engagement. The description of characteristics was not restricted to capturing macro tendencies but even delivered dynamics over time, sentiment cycles, and differences between various subgroups of the respective community. Results affirm LIWC as an efficient analysis tool for tracking communal sentiment, well-being and aspects of belongingness. The results are quite often nuanced: small percentage points highlight differences for more than one community characteristic. Yet, topic domains and specific other scores allow for detecting more specific interpretations and should not be disregarded.

Information estimated from aggregated social media data may lack some interpretation quality but provides an easy and repeatable way to gain quick insight into the essential factors defining a community. Macro-assessment of social indicators rises from investigation of post-comment distinction, a pre-given structure of any Facebook dataset. This means that the approach is easily replicable for other communities and generalizable. Although some customizing effort concerning data preparation are inevitable if community-specific insights are pursued, many of the employed partitions are to be individualized to further use cases. This aspect of popularly sourced well-being information is ripe for broader adaptation.

REFERENCES

[1] A. Lindner, C. Niemeyer, S. Caton, M. Hall, C. Niemeyer, and S. Caton, "BeWell: A Sentiment Aggregator for Proactive Community Management," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015, pp. 1055–1060.

[2] S. Caton, M. Hall, and C. Weinhardt, "How do politicians use Facebook? An applied Social Observatory," *Big Data Soc.*, vol. 2, no. 2, p. 2053951715612822, Dec. 2015.

[3] C. Cioffi-Revilla, "Computational social science," *Comput. Stat.*, vol. 2, no. 3, pp. 259–271, May 2010.

[4] M. Hall and S. Caton, "A Crowdsourcing Approach to Identify Common Method Bias and Self-representation," Oxford, England, 2014.

[5] J. Lingel, M. Naaman, and danah boyd, "City, self, network: transnational migrants and online identity work," in *CSCW'14*, 2014, pp. 1502–1510.

[6] R. E. Wilson, S. D. Gosling, and L. T. Graham, "A Review of Facebook Research in the Social Sciences," *Perspect. Psychol. Sci.*, vol. 7, no. 3, pp. 203–220, May 2012.

[7] A. Kramer, "An Unobtrusive Behavioral Model of 'Gross National Happiness,'" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 287–290.

[8] A. Kramer, "The spread of emotion via facebook," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2012, pp. 767–770.

[9] A. Kramer, J. E. Guillory, and J. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Natl. Acad. Sci.*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014.

[10] N. Wang, M. Kosinski, D. Stillwell, and J. Rust, "Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National," *Soc. Indic. Res.*, vol. 115, no. 1, pp. 483–491, 2014.

[11] Y. Tausczik and J. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, Dec. 2010.

[12] J. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The Development and Psychometric Properties of LIWC2007," University of Texas, Austin, Austin, TX, 2007.

[13] C. Chung and J. Pennebaker, "The Psychological Functions of Function Words," *Soc. Commun.*, pp. 343–359, 2007.

[14] C. J. Groom and J. Pennebaker, "Words," *J. Res. Pers.*, vol. 36, pp. 615–621, 2002.

[15] J. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us*. New York, New York, USA: Bloomsbury Press, 2013.

[16] S. González-Bailón, N. Wang, A. Rivero, and J. Borge-Holthoefer, "Assessing the bias in samples of large online networks," *Soc. Networks*, vol. 38, no. January, pp. 16–27, 2014.

[17] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science (80-.).*, vol. 346, no. 6213, pp. 1063–1064, 2014.

[18] M. Russell, *Mining the Social Web*, Second. Sebastopol, CA: O'Reily Media, 2013.

[19] J. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: our words, our selves.," *Annu. Rev. Psychol.*, vol. 54, pp. 547–77, Jan. 2003.

[20] J. Pennebaker and T. C. Lay, "Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences.," *Journal of Research in Personality*, vol. 36. pp. 271–282, 2002.

[21] F. Huppert and T. T. C. So, "What percentage of people in Europe are flourishing and what characterises them?," Florence, Italy, 2009.

[22] B. Böcking, M. Hall, and J. Schneider, "Event Prediction With Learning Algorithms—A Study of Events Surrounding the Egyptian Revolution of 2011 on the Basis of Micro Blog Data," *Policy & Internet*, vol. 7, no. 2, pp. 159–184, 2015.