Sentiment as a Predictor of Wikipedia Editor Activity*

Sergio Martinez-Ortuno[1], Deepak Menghani[2], Lars Roemheld[3]

Abstract— **We perform sentiment analysis on messages exchanged between Wikipedia editors in the so-called "user talk pages," to predict future user editing behavior. We found a reasonably well-performing model to predict the number of edits next week on a per-user level by applying the GBM algorithm, and we discuss the relatively limited impact our sentiment scores had for this model. Our findings could be used better engage editors, potentially resulting in better article quality.**

I. INTRODUCTION

Wikipedia, the worlds largest encyclopedia, is created by millions of unpaid editors online. Every user can edit every article, and the project is protected against vandalism and low-quality contributions only through version control and a system of (again unpaid) reviewers. Somewhat hidden to most casual readers of the encyclopedia, Wikipedia also features a simple social network: every user has a personal user profile and a "user talk page" which acts as a publicly accessible guestbook where users can leave messages to each other.

The messages exchanged in user talk pages are often related to a user's editing behavior. For example, senior users may welcome new users, or congratulate them on their first edits. Administrators may officially warn culprits after transgressions of Wikipedias content guidelines or policies. Users may also thank one another for certain edits, and, of course, users engage in heated debates about what the ground truth reflected in a certain article should be. Not all such debates are pleasant, although the community as a whole has been noted for its considerable resilience against both anarchy and uncontrolled aggression [1]–[3].

Social feedback has long been known to be a strong influencer of intrinsic motivation [4], [5]. Observing praise and gratitude may be a strong incentive for Wikipedia editors to "keep up the good work," whereas repeated unpleasant discussions, official warnings, or even personal insults may discourage further editing behavior. With this intuition in mind, we formulated our hypothesis: we ask if received message sentiment can help predict editor activity on Wikipedia. In so doing, we create the opportunity to engage with frustrated editors—for example, motivating emails could be sent to users who are expected to significantly reduce their editing due to received message sentiment. If effective, this could increase overall editing activity (and editor happiness) on Wikipedia; for the scope of this paper we assume a high number of edits to be desirable, since it enables the encyclopedia to better reflect an everchanging world.

Previous work analyzed the sentimental content of the conversations between Wikipedia editors [3], [6]. Our work is unique in that we focused on the personal messages exchanged through user talk pages and their relation to future editor activity, rather than analyzing the more factual discussions on the so-called article talk pages.

II. DATA AND METHODS

A. Source Data and Sentiment Scores

We accessed an anonymized replica of the English Wikipedia database through the Wikimedia Foundation's research servers [7]. For all users who registered in 2013, and who had made at least one article edit ($n \approx 620,000$), we downloaded the contents, date and author of each message received by these users through their user talk pages, and the number of article edits made by each user per week in the complete year 2013.

Wikipedia's talk pages are implemented in such a way that any user can not only add text, hyperlinks and images; but also delete anything, even content added by other users. We collected the revision history of each user talk page and we performed a diff data comparison between each revision and the previous one. If there was any text added or replaced on a particular revision, then we considered these additions to be the content of the "user talk message". To prepare the messages for sentiment analysis, we stripped them from any formatting markup [8] and applied Porter's stemming algorithm [11] to each word using Python's NLKT library.

We used two different sentiment dictionaries for message scoring: well-known Bing Liu's Opinion Lexicon [9] provided two lists of words, one of 2006 words with positive emotional connotation, and one of 4783 words with negative connotation. We further used the NRC Word-Emotion Association Lexicon [10], a lexicon of 13901 words, where each word is labeled with 10 interesting emotional dimensions: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust. We compressed every message into 12 numerical features, each one providing the relative strength of the measured sentiment and calculated as share of words in a message appearing in one particular dictionary.[1]

In mid-2013, a new feature was introduced on Wikipedia that allowed users to express "thanks" for specific edits

* We thank the Wikimedia Foundation, and in particular Leila Zia of the Research and Data team, for their generous help and support of our project. This is a course project for Andrew Ng's course CS 229 and (by approval of both teaching teams) for Sharad Goel's course MS&E 331 at Stanford University. We are grateful to both courses for an inspiring quarter.

[1] sergiomo@stanford.edu
[2] deemeng@stanford.edu
[3] roemheld@stanford.edu

[1] To be specific, $\text{score}_m = \frac{1}{|\text{sentence}|} \sum_{w \in \text{sentence}} \mathbb{1}(w \in \text{dictionary})$

with a single click. The thanked editor would then see a notification the next time he or she logged in. We enriched our dataset with the number of thanks received by each user per week.

B. Data Analysis and Models

Manual data inspection showed the weekly number of edits per user to be extremely noisy, with strong spiking patterns even for the most active users (see figure 1). On a weekly basis, the number of edits per user is clearly dependent on the preceding week's number of edits (edits per week are a time series). Typically, an editor would receive user talk messages after periods of increased activity, reacting on his editing behavior. Under our hypothesis, these messages would then influence his subsequent activity. Given this observation, we decided to model weekly differences in editing behavior, and to employ a threshold model. Specifically, we wanted to predict weeks where editing behavior is outside a τ-neighborhood of the previous week, τ being a model parameter. Our basic model was to predict the binary event

$$\mathbb{1}(|e_i(w) - e_i(w-1)| > \tau_i(w))$$

where $e_i(w)$ denotes the number of edits by user i in week w. Different functions for the $\tau_i(w)$ values give different predictors. We tried relative thresholds, historic rolling midrange, and historic standard deviation; the latter performed best, and we defined $\tau_i(w)$ as the empirical standard deviation in weekly edit count for user i, up until week $(w-1)$:

$$\tau_i(w) = \frac{1}{w-1}\sum_{t=1}^{w-1}(e_i(t) - \overline{e_i})^2$$

Figure 2 shows such a $\tau_i(w)$-neighborhood for one exemplary user, and marks the events to be predicted with circles.

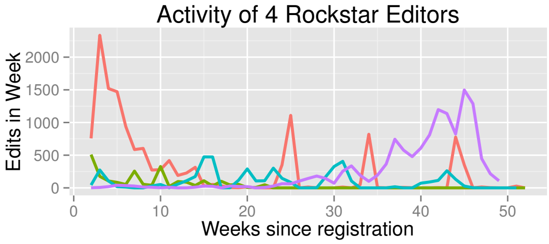

Fig. 1. Exemplary behavior of 4 of Wikipedia's most active editors signed up in 2013. Editors have periods of very high engagement, followed by times of relative inactivity.

In different model iterations we considered different subsets of our feature space (see ablative analysis in table II). Our total feature space comprised the following features:

- Number of weeks since user registration and $exp(-[\text{weeks since registration}])$: the average number of edits across all users appeared inversely exponential in the user account age, leading us to explicitly include this factor
- Historic count of edits

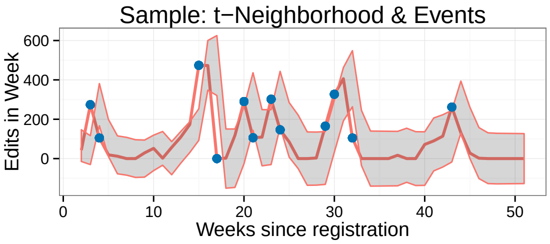

Fig. 2. Model visualization: we predicted the events signified by blue circles, when a user's edit count was outside of a 1σ-neighborhood of the preceding week (grey ribbon).

- Historic count of thanks received
- Historic count of messages received
- Historic count of message-words received
- Historic average message-sentiment received (along 12 emotional dimensions)

Finally, we defined the relevant history for the model to be two weeks, using weekly bins to reduce computational complexity and to smooth over the somewhat sparse data. To predict a "significant change in editing activity" event in week w, we included editor activity and communication history in week $(w-1)$ and week $(w-2)$, and treated the two weeks separately to allow our models to catch temporal patterns. We did not include more than two weeks to reduce the risk of overfitting, implicitly assuming that motivational effects of messages will not be significant beyond a time span of two weeks. Our total feature vector spanned 34 features.

The vast majority of Wikipedia user accounts is inactive, in that most users perform at most one edit (typically within the first week of registration, after which the user accounts become inactive). Most users in our data will therefore not receive any messages, or show any other signs of activity. In fact, only about 3% of the users in our dataset had more than one edit and received more than 2 messages on their user talk page (the first one typically being an automated welcome-message). In order to get more meaningful effects in our sentiment models, we reduced our dataset to these users ($n \approx 17,500$ users).

We split our data into a training set (80%) and a test set (20%) to perform holdout validation. Unless otherwise noted, all performance metrics are calculated on the test set.

Our definition of output variable and feature space framed our research question as a binary classification problem, for which we tried three algorithms: a logistic-regression GLM, a support vector machine with linear kernel (SVM), and a bernoulli-distributed gradient boosted tree model (GBM).

III. RESULTS

We worked with relatively sparse and unorganized data, and probing into the sentment-dictionary scores showed them to align only roughly our subjective ratings. Given this, our model performed surprisingly well, and we obtained reasonable predictive power. The ROC-curve in figure 3 summarizes the performance of the three algorithms tested:

GBM performed significantly better than the algorithms that attempted to find a linear boundary between the positive and negative classes.

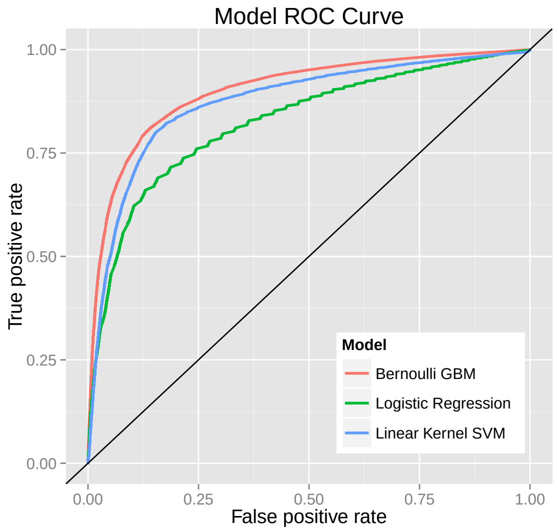

Fig. 3. ROC curve for GLM, SVM, and GBM (as calculated on test set). Our Gradient Boosted Model showed the best performance overall.

Interestingly, this hints at a strongly non-linear relationship between our input features and the output event. This is not immediately intuitive, since one could have expected that "more anger" will always deter further edits—we attribute part of the nonlinearity to time-patterns in the historic data. GBM's superior performance is not due to overfitting, as can be observed in table I (test error ≈ training error).

In table I, we present performance indicators for the test and train datasets. Additionally, we tested on a subset of our data, which excluded all weeks without received message history ("Test Comments", only weeks with at least one message in weeks $(w-1)$ and $(w-2)$). For all algorithms, the recall corresponding to the "Test Comments" set stands out for being significantly higher than the rest, while the precision remains about the same across the board. In other words, it appears that the models are more successful at detecting a significant change in week-on-week activity when they have information about the messages received by the users during the past 2 weeks.

	TEST			TRAIN			TEST COMMENTS		
Model	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
GLM	90%	64%	20%	91%	66%	20%	63%	65%	56%
SVM	91%	58%	41%	91%	59%	42%	65%	62%	74%
GBM	92%	69%	47%	93%	70%	48%	73%	70%	79%

TABLE I

MODEL COMPARISON: "TEST COMMENTS" IS A SUBSET OF OUR DATA, EXCLUDING ALL WEEKS WITHOUT RECEIVED MESSAGE HISTORY. NO SIGNIFICANT OVERFITTING CAN BE OBSERVED.

To further investigate this, we performed ablative analysis by removing features from our model and observing the resulting test statistics (table II). We found that removing all message sentiment causes only a minute decrease in model performance, whereas the historic number of edits carried much greater importance (underlining the time-series character of our data once more).

Adding sentiment back onto the overwise "empty" model, the sentiment scores by themselves were enough to reach a performance level comparable to that of the fully-featured model—especially for the "Test Comments" set. Since receiving any messages at all correlates with previous editing activity, we attribute some of this effect to the implicit inclusion of historic edit counts by including historic sentiment. Especially given the effects on the less sparse "Test Comments" set, we conclude that sentiment scores have some predictive value for editor activity. They are not at all sufficient, however, to build a reliable model.

Figure 4 offers further insight into our best-performing GBM model: the model is well-calibrated, and makes relatively many highly confident predictions. In particular, it captures the fact that the bulk of users did not have significant changes in week-on-week behavior.

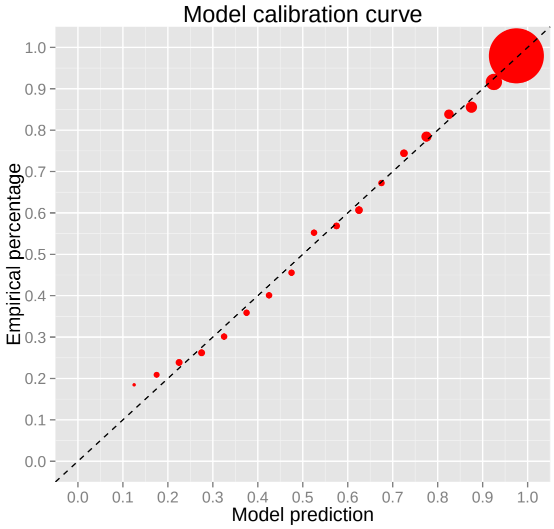

Fig. 4. The calibration curve of our GBM model. The model was well-calibrated overall. The large circle in the top right corner implies that our model makes a strongly confident and correct prediction for the bulk of the data where no significant change in week-on-week editing occurred.

IV. DISCUSSION AND FUTURE WORK

While we found some predictive value for future behavior in the sentimental content of messages received by Wikipedia editors, we do not have evidence to establish a causal relationship between these variables.

Furthermore, we note that our conclusions do not necessarily generalize outside of Wikipedia and similar crowdsourced environments. The first limitation that we see is that most

[2] The model with this feature set predicted the negative class for all training examples.

	TEST			TEST COMMENTS		
Model	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
Time since registration, 2 weeks of edits, thanks and messages	92%	69%	47%	73%	70%	79%
Time since registration, 2 weeks of edits and thanks	92%	68%	46%	72%	72%	69%
Time since registration, 1 week of edits and thanks	91%	61%	42%	65%	65%	63%
1 week of edits and thanks	91%	59%	41%	63%	63%	61%
Time since registration	90%	–[2]	0%	50%	–[2]	0%
Time since registration, 2 weeks of messages	90%	64%	21%	64%	64%	63%
Time since registration, 2 weeks of messages and thanks	90%	64%	21%	63%	64%	62%

TABLE II

Ablative Analysis

new Wikipedia users do not make frequent use of the user talk pages. This by itself limits the macro-level impact of any interactions that occur in these pages. Second, our definition of sentimental content was rather limited in that we only performed a simple word-matching analysis. We would like to perform more elaborate bag-of-words classifiers in future work to build on stronger messaging patterns.

Most messages exchanged through user talk pages are not sentimentally-loaded, but rather talk about the Wikipedia guidelines and policies in a neutral manner. More sophisticated natural language processing techniques could help identify more complex patterns in these messages. Another possible refinement could come from performing cluster analysis on the contents of the messages, to find the set of ideas that are most-commonly exchanged in the user talk pages (an example of one of such ideas might be: I reverted you edit because you did not cite any sources, or I deleted your image because it violated copyright law).

We would have liked to take into account the quality (measured by later reverts) and size of the edits performed. Unfortunately, we were unable to obtain this data for this study. That is, for this investigation we could not distinguish between an edit that corrected a typo versus a new article creation, for example.

Nonetheless, we were able to detect macro-level patterns of behavior that appear to discredit the hypothesis that the sentimental content of user talk pages is a main driver of user churn on Wikipedia. Additionally, we undertook the first steps in building a useful model to predict when a user is about to suddenly stop making contributions to the encyclopedia.

References

[1] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). ACM, New York, NY, USA, 453-462.

[2] Viegas, F.B.; Wattenberg, M.; Kriss, J.; van Ham, F., "Talk Before You Type: Coordination in Wikipedia," System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on , vol., no., pp.78,78, Jan. 2007

[3] David Laniado, Carlos Castillo, Andreas Kaltenbrunner and Mayo Fuster-Morell (2012). Emotions and dialogue in a peer-production community: the case of Wikipedia. WikiSym '12 - 8th International Symposium on Wikis and Open Collaboration, Linz, Austria, August 2012.

[4] Cameron, J., & Pierce, W. D. (1994). Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis. Review of Educational Research, 64(3), 363423. doi:10.3102/00346543064003363

[5] Geister, S. (2006). Effects of Process Feedback on Motivation, Satisfaction, and Performance in Virtual Teams. Small Group Research, 37(5), 459489. doi:10.1177/1046496406292337

[6] D. Iosub et al, "Emotions under Discussion: Gender, Status and Communication in Online Collaboration", in PLoS ONE 9(8):e104880, 2014.

[7] Wikimedia Tool Labs, `https://wikitech.wikimedia.org/wiki/Help:Tool_Labs`

[8] Giuseppe Attardi, Wikipedia Extractor Code, `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

[9] M. Hu and B. Liu, Mining and Summarizing Customer Reviews, in Knowledge discovery and data mining, 2004, pp. 168-177.

[10] S. Mohammad and P. Turney, Crowdsourcing a Word-Emotion Association Lexicon, in Computational Intelligence, 39(3), 2013, pp.555-590.

[11] M.F.Porter, An algorithm for suffix stripping, in Program 14 no. 3, 1980, pp.130-137