Keep me posted! Human and machine learning analysis of Facebook updates

Franco Delogu
Humanities, Social Sciences & Communication
Lawrence Technological University
21000 W Ten Miles Rd.
Southfield, Michigan, USA
1-248-204-3527
fdelogu@ltu.edu

Marija Franetovic
Department of eLearning
Lawrence Technological University
21000 W Ten Miles Rd.
Southfield, Michigan, USA
1-248-204-3758
mfranetov@ltu.edu

Lior Shamir
Math & Computer Science
Lawrence Technological University
21000 W Ten Miles Rd.
Southfield, Michigan, USA
1-248-204-3512
lshamir@mtu.edu

ABSTRACT
The key element of Facebook social network platform is the status updates, in which the user can upload text or other media such as pictures and videos. In this study, we manually classified more than 3500 Facebook status updates (FSUs) by the subject, the emotional activation, the medium used, the originality and the degree of self-referential content. We then cross-tabulated that information with demographic factors such as gender and occupation. Thirty students participated in the categorization task, each annotating more than 100 FSUs of their Facebook friends' FSUs. Statistical and supervised machine learning analysis was then applied to the categorized features. The text itself was not analyzed further after the annotation for the purpose of preserving the privacy and anonymity of the FSU authors. Results show that FSUs vary in subject, emotional connotation and structure as a function of demographic factors like gender and occupation of the poster. Statistical analysis and supervised machine learning are able to predict the demographic and emotional expressions based on the other features annotated by the participants.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Information Systems – *models and principles*, *software psychology*.

General Terms
Human Factors.

Keywords
Facebook, status update, social media, psychology.

1. INTRODUCTION
The Online Social Network (OSN) of Facebook has over 1,39 billion monthly active users, with approximately 20% of its users residing in United States [6]. Facebook has also become the central hub for sharing information [12]. Studying Facebook allows social scientists to study a wealth of measureable behavior versus the traditional gathering of self-reports that they are used to relying upon [9]. Facebook identity is not only a study of online behavior; it has actually permeated offline reality so that the two are partially integrated [3,10]. This reality extends out to networks from high school, college, workplaces and professional networks, special interest groups, and regions. It also extends out to Facebook real-estate that is integrated with other websites and applications.

In Facebook, status updates (FSUs) are commonly used to broadcast personal states and experiences or to share opinions about something considered interesting or relevant. Consequently, the analysis of FSUs can provide significant information about aspects of people's interactions and mental life in general. In spite of this topic's relevance, there are only a few studies which specifically and directly analyze FSUs' texts with a goal of categorizing their semantic content. In particular, the literature lacks studies in which FSUs are analyzed through human and computerized categorization and in which the effects of the process of categorization analysis of such posts may alter one's use and perception of Facebook. One of the reasons for the limited amount of studies is that, for privacy constraints, the access to FSUs is restricted to Facebook friends of each specific user, precluding by consequence the analysis of large datasets. In a recent study, Garcia and Sikström [7] showed that semantic features of status updates can predict personality traits. In particular, they found that the semantic content of Facebook updates predicted Psychopathy and Narcissism. As there is evidence indicating that Facebook materials are reliable indicators of what people really think and feel [3], it appears probable that data collected through FSUs is a reliable representative of actual ideas and feelings outside the social network. Is it possible to predict how people will use Facebook using demographics? McAndrew & Jeong [11] created a survey to collect data about Facebook habits as a function of age, gender and marital status. Their results indicate that demographics is a reliable predictor of types of Facebook use.

Concerning methods, psychological and sociological research on Facebook relies mostly on surveys [2,11], recruitment of participants via Facebook applications, and data crawling, which involves gathering data from profiles without the active participation of users via automated means [17]. As from March 2011, data had become difficult to collect using automated means due to privacy restrictions and thus this method has become less informative. The method we are proposing in this study includes the integration of data crawling and subject recruiting. Students

MOBIMEDIA 2015, May 25-27, Chengdu, People's Republic of China

DOI 10.4108/icst.mobimedia.2015.259075

involved in the project did an analysis of their Facebook friends' FSUs. Students deleted or recoded any information that could reveal the identity of the posters before saving the text of each FSU. In this way, with our method, we have access to a large amount of FSUs without violating the privacy of the posters.

In the social media age that we live in, constructivist theories [8,16] are usually used to teach and engage students because current learning is most commonly constructed through social and technological mediation. Per Ally [1], a constructivist theoretical orientation allows us to answer the "why" questions, whereas the previously used cognitive orientation - the "how" questions, and the behaviorist orientation - the "what" questions. Throughout this study, students retrieved information and problem-solving through coding as well as assisting in figuring out how to perform automatic text analysis. This method may increase their awareness about rationales for Facebook status updates. Based on what they surmise, it may possibly alter their own Facebook attitudes and usage patterns.

Learning requires a change in long-term memory and can be identified by actions and attitudes which progress from ones of a novice, which have errors and are slow, to ones of an expert, which are effortless. As the novice becomes more familiarized with content, their schemas or long-term memory structures are altered so that working memory can perform more efficiently [14]. The structuring of information and interactivity are important elements in the learning process and thus should influence instructional design [15]. In this study, as novices in media analysis become more and more proficient over prolonged exposure, they learn about Facebook from the perspective of a researcher instead of acting as mere users. Thus, they may change their attitudes and usage patterns. The social media, the method of analysis itself and conditions of exposure may be explored as further determinants of a change in behavior. For this reason, this research has an important pedagogical valence.

As noted by Wilson and collaborators [17], Facebook data analysis has been underestimated in the field of psychology in general, and in identity research in particular. This study uses a unique method to gather and analyze data, one which is a combination of recruited volunteers' friends' FSU analysis and automated analysis of said posts. As such, it circumvents the new stricter privacy policies [5].

2. METHODS AND PROCEDURES

Methods consists in the collection and analysis of content from approximately 4000 FSUs. 30 undergraduate psychology students from Lawrence technological University collected and analyzed posts from their Facebook Friends. Demographic information were collected in terms of the FSU, such as age, gender and #s of friends in order to aid in descriptive analysis.

After the posts were collected and categorized through human judgment the automatic text analysis was applied to all posts with the purpose of automatically categorizing them according to, whenever possible, the same categories used with human categorization.

The annotators determined for each FSU the activation, length, use of quotes, and originality, also ranked the relevance of the FSU to music, sports, religion, food and drink, art and aesthetics, comedy, opinion and society, and daily life.

3. ANALYSIS

3.1 Behavioral classification

Variation in the features of the FSUs were analyzed as a function of gender and occupation of the posters by mean of Analysis of Variance (ANOVA) for the parametric data, and Person's Chi-Square for non-parametric data.

3.2 Automatic classification

The automatic supervised machine learning algorithm used in this study is based on the Weighted Nearest Distance method (Shamir et al., 2008). For each feature of categorization the FSUs were separated based on the category. Then, each of the other categorization features was assigned with its Fisher discriminant score (Bishop, 2006), reflecting its ability to differentiate between the FSUs based on the equation

$$W_f = \frac{\sum_{c=1}^{N}\left(\overline{T_f} - \overline{T_{f,c}}\right)^2}{\sum_{c=1}^{N}\sigma_{f,c}^2} \cdot \frac{N}{N-1}$$

where W_f is the Fisher discriminant score, N is the total number of categories, T_f is the mean of the values of feature f in the entire dataset, $T_{f,c}$ is the mean of the values of feature f in the category c, and $\sigma^2_{f,c}$ is the variance of feature f among all samples of category c. Conceptually, the Fisher discriminant score of a feature is higher if the variation of the feature values within the categories is low, but the variation of the values between the categories is high. Hence, the features are weighted by their ability to differentiate between the FSUs.

After each feature is assigned with its Fisher discriminant score, the distance between a training FSU X and test FSU Y is measured by the equation

$$d = \sqrt{\sum_{f=1}^{|X|} W_f\left(X_f - Y_f\right)^2}$$

where W_f is the assigned Fisher discriminant score of feature f, and d is the computed weighted distance between the two feature vectors. The predicted class of a given FSU is determined by the category of the training FSU that has the shortest weighted distance d to the test FSU.

4. RESULTS

Behavioral classification analysis. The statistical analysis of the behavioral classification showed that, concerning the emotional content, women were more positive than men, $F(1, 3534)=13.756$, $p=.00021$ and workers were more positive than students, $F(1, 3534)=7.02$, $p=.008$. Also, men's FSUs expressed more activation than women's, $F(1, 3453)=4.46$, $p=.034$. No effects of gender and occupation were found on the number of likes received.

Concerning the structural features of the FSUs, a clear gender effect was found on the medium used by the posters: men posted more textual FSUs than women (47% vs. 41% of all FSUs), but

less pictures (36% vs. 43%). It is interesting to notice that women posted more pictures than text. No significant gender difference was found in the amount of video posted (16.5% vs. 15.5%). Also, men tended to be briefer than women, Pearson Chi-square: 39.5319, df=7, p<.0001, as their posts contained less words than women's.

Concerning contents and topics of the FSUs, results show that women's FSUs were more self-referential than men's, Pearson Chi-square: 6.16, df=2, p=.046, and less original, Pearson Chi-square: 9.99, df=1, p=.0016. The preferred topic by all posters was *Daily Life*, which alone was classified as FSU's topic in more than the half of all FSUs. Also, a chi-square analysis showed a clear effect of gender on the frequency of FSUs classified under specific topics (see Figure 1). In particular, women posted more FSUs about their personal *Daily Life* than men, Pearson Chi-square: 6.49, df=1, p=.019.

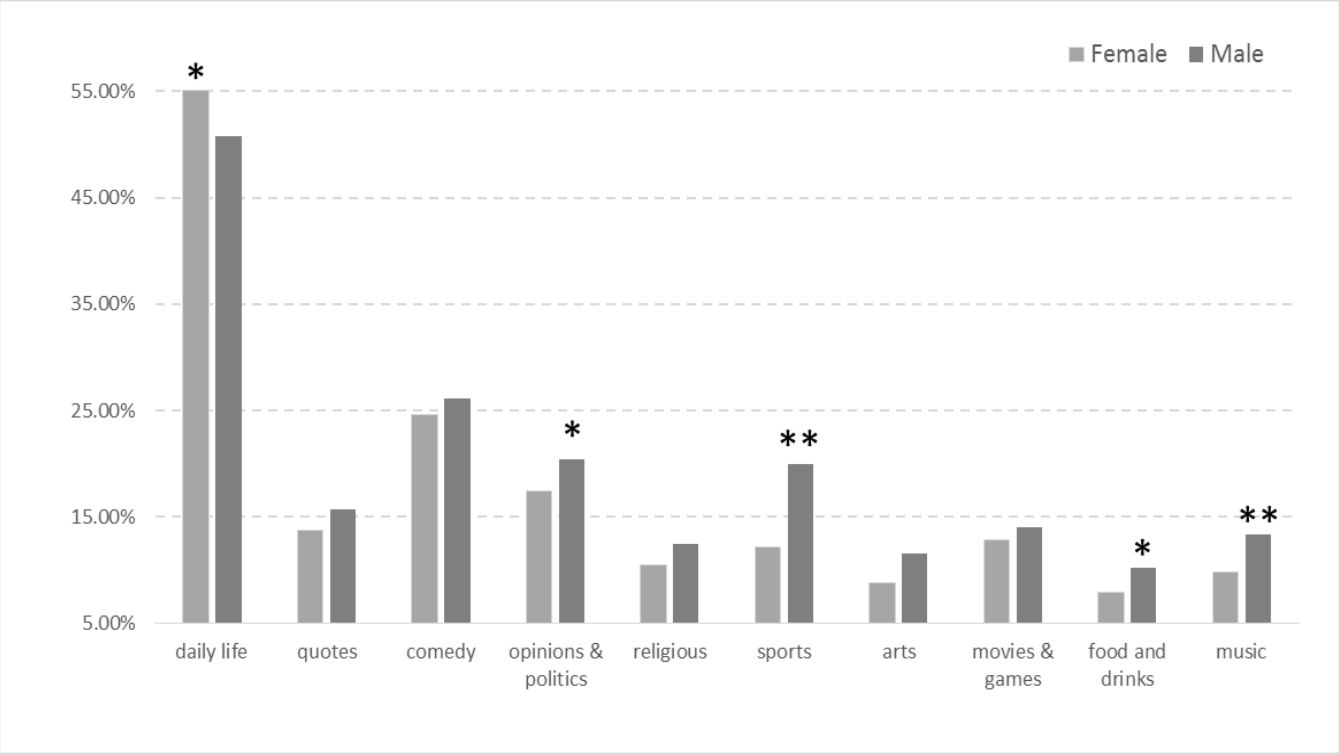

Figure 1: Frequencies of FSU topics as classified by human judges. Single asterisk indicates differences at p<0.05, double asterisk indicates differences at p<0.01.

Men focused more than women on *Opinion & Politics*, Pearson Chi-square: 4.76, df=1, p=.029, *Sports*, Pearson Chi-square: 44.9938, df=3, p<.0001, *Food and Drinks*, Pearson Chi-square: 5.51, df=1, p=.019, and *Music*, Pearson Chi-square: Pearson Chi-square: 10.8565, df=1, p=.000. For all the other categories, including *Quotes, Comedy, Religious, Art, Movies and Games*, there were not significant differences between the two gender groups.

Automatic supervised machine learning analysis. The automatic supervised machine learning analysis tested the automatic categorization of several features, including the gender, activation, number of likes, positive or negative emotion, and subjects such as music and sports. Table 1 shows the number of training and test FSUs that were used in each supervised machine learning experiment, as well as the accuracy of the automatic classification based on the annotations of the other features. Each experiment was repeated 20 times such that in each run the FSU annotations were randomly allocated to training and test sets. The classification accuracy is the mean accuracy of all 20 runs.

As Table 1 shows, in ~57.1% of the FSUs the algorithm could predict the gender of the FSU author based on the other features. The highest classification accuracy was achieved for the number of likes (lower than the average or higher than the average), showing that the number of likes a FSU receives can be predicted based on the other features annotated by the human annotators. These results indicate on dependence between the gender, topic of the FSU, and the active and passive response of the FSU readers.

The most informative features that allowed the prediction of the number of likes, as measured by their Fisher discriminant scores, are the positive and negative response, but also the topic. FSUs related to religion tend to attract more likes, as well as opinions about society.

Table 1. Automatic supervised machine learning categorization of different categorization features based on the other features.

Feature	# Training FSUs	# Test FSUs	Categorization accuracy (%)
Gender	1700	200	57.1
Activation	180	20	61.4
Likes	239	20	70.3
Positive/negative	240	20	62.8
Music	350	20	62.5
Sports	450	20	53.6

The features that separated between genders were mostly related to the topic such as sports, music, and art and aesthetics. Also, FSUs posted by males attracted 11.66 likes on average, while FSUs posted by women received ~10.1 likes on average.

FSUs about music received significantly more likes compared to other FSUs (~21.6 compared to ~8.7), and FSUs about music were more likely to be related to art (~0.31 compared to ~0.07), but also to religion (~0.31 compared to ~0.09).

5. CONCLUSIONS

The FSU functionality is the key element of Facebook, in which users typically share interests and life experiences under the formats of text, pictures and videos. This study presents an original collection and analysis of FSUs according to several dimensions. After collection, the analysis was performed both with traditional statistical techniques and through an automatic supervised machine learning analysis.

The two analyses consistently found a significant gender effect in some of the topics under scrutiny. According to both human classification and computerized analysis, men seem to be significantly more focused than women in sports and music. Less consistent are the results in art, in which the difference is significant only according to the computer analysis and for daily life, food and drinks and opinion and politics, in which the difference is significant only according to the traditional statistical analysis (Pearson's Chi-Square). In summary, the analyses of the influence of gender on the FSUs topics converge to the general evidence that women's FSUs are more focused on their own life than men's. In fact, women are more self-referential and post contents related to their personal experiences compared to men, who tend more to write about topics which are non-immediately related to themselves, and especially about sports and music. The medium is also an important differential factor between the two genders, with women's tendency to prefer pictures to text, and men's prevalence of textual FSUs.

The machine learning analysis tool provides stimulating findings about multidimensional dependencies between features which do not clearly emerge with the traditional statistical analysis. Particularly interesting is the dependence between the emotional valence, the topic of the FSU, and the readers' response. Specifically, positive posts attract more likes, as well as FSUs related to religion or opinions and politics.

In conclusion, confirming and extending previous results [11], our findings support, with a new methodological approach, the evidence that demographic aspects can predict how people share contents in Online Social Networks.

REFERENCES

[1] Ally, M. 2004. Foundations of educational theory for online learning. In T. Anderson & F. Elloumi (Eds.) Theory and practice of online learning. Athabasca, AB: Athabasca University.

[2] Acquisti, A. and Gross, R. 2006. Imagined communities: Awareness, information sharing and privacy on the Facebook. In Proceedings of Privacy Enhancing Technologies Workshop (pp. 36–58), Cambridge, England: Springer.

[3] Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B. and Gosling, S. D. 2010. Facebook profiles reflect actual personality, not self-idealization. Psychological science.

[4] Bishop, C.M. 2006. Pattern Recognition and Machine Learning, Springer Press, New York, NY. Pages 191-192.

[5] Facebook. 2011b. Facebook terms. Palo Alto, CA: Facebook. Retrieved from http://www.facebook.com/terms.php

[6] Facebook. 2015. Statistics of Facebook. Palo Alto, CA: Facebook. Retrieved from http://newsroom.fb.com/content/data#!/

[7] Garcia, D. and Sikström, S. 2014. The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. Personality and Individual Differences, 67, 92-96.

[8] Garrison, D. and Arbaugh, J. 2007. Researching the community of inquiry framework: Review, issues, and future directions. The Internet and Higher Education, 10(3), 157-172.

[9] Graham, L. T., Sandy, C. J. and Gosling, S. D. 2011. Manifestations of individual differences in physical and virtual environments. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), Handbook of individual differences (pp. 773–800). Oxford, Eng-land: Wiley-Blackwell.

[10] Lampe, C., Ellison, N. and Steinfield, C. 2006. A Face(book) in the crowd: Social searching Vs. social browsing. Paper presented at the ACM Special Interest Group on Computer-Supported Cooperative Work, Banff, AB, Canada.

[11] McAndrew, F. T., & Jeong, H. S. 2012. Who does what on Facebook? Age, sex, and relationship status as predictors of Facebook use. Computers in Human Behavior, 28(6), 2359-2365.

[12] Ries, T. 2010. 250 million people engage with Facebook on external sites monthly. Retrieved from http://therealtimereport .com/2010/12/11/250-million-people-engage-with-facebook-on-external-sites-monthly

[13] Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J. and Goldberg, I. 2008. Wndchrm - an open source utility for biological image analysis, BMC Source Code for Biology and Medicine, 3: 13.

[14] Sweller, J. 1983. Cognitive load during problem solving: Effects on learning, Cognitive Science, 12, 257-285.

[15] Sweller, J. 1994. Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4, 295–312.

[16] Wenger, E. 2006. Communities of practice. Retrieved from http://www.ewenger.com/theory/

[17] Wilson, R. E., Gosling, S. D. and Graham, L. T. 2012. A review of Facebook research in the social sciences. Perspectives on Psychological Science, 7(3), 203-220.