Tweet! – And I Can Tell How Many Followers You Have

Christine Klotz[1], Annie Ross[2], Elizabeth Clark[3], and Craig Martell[4]

[1] FernUniversität in Hagen, Fakultät für Mathematik und Informatik, Hagen, Germany
[2] Colorado State University, United States
[3] Middlebury College, United States
[4] Naval Postgraduate School, Department of Computer Science, Monterey, United States
c._klotz@web.de, {anniesross,eaclark07,craig.martell}@gmail.com

Abstract. Follower relations are the new currency in the social web. User-generated content plays an important role for the tie formation process. We report an approach to predict the follower counts of Twitter users by looking at a small amount of their tweets. We also found a pattern of textual features that demonstrates the correlation between Twitter specific communication and the number of followers. Our study is a step forward in understanding relations between social behavior and language in online social networks.

Keywords: Twitter, follower, user characteristics, text mining, n-grams, Naïve Bayes, tf-idf, online social networks.

1 Introduction

The rise of social media has created a new platform for viral and brand marketing, public relations and political activities. Therefore, "giving the right message to the right messengers in the right environment" [1] is essential for effective campaigns. The number of followers is often interpreted as sign of popularity and prestige, follower relations count as currency within the social web [2]. A marketplace for followers is already evolving [3].

From the perspective of information diffusion, Twitter users with high counts for followers can be categorized as 'information sources' or 'generators' and with high counts for followings as 'information seekers' or 'receptors' [4, 5]. Users with equal numbers of followers and followings are called 'acquaintances' to emphasize the reciprocity of the relationship [6]. Retweeting is associated with information sharing, commenting or agreeing on other peoples' messages and entertaining followers [7]. Tweet content may fall in the categories: 'Daily Chatter', 'Conversations' (tweets containing the @-symbol), 'Sharing Information' (tweets containing at least one URL) or 'Reporting News' (tweets via RSS feeds) [4].

User-generated content obviously plays an important role for tie formation on the social web. Hutto et al. [2] monitored Twitter users over a period of 15 month and concluded that "social behavior and message content are just as impactful as factors related to social network structure" for the prediction of follower growth rates. Kwak et al. [8] examined factors that impacted the stability of follower relationships. They

S. Boonkrong et al. (eds.), *Recent Advances in Information and Communication Technology*, Advances in Intelligent Systems and Computing 265,
DOI: 10.1007/978-3-319-06538-0_24,

found that, beside a dense social network, actions of mutual behavior assured the stability of relations. These include retweeting and replying as well as the usage of similar hashtags in tweets and the inclusion of URLs. Both approaches needed a high amount of data points which were gathered over long periods of time.

In this study, we concentrate on a single point in time. We report an approach to predict the follower rates of randomly picked Twitter users by looking at a small amount of data. Therefore, we apply naïve Bayes classification on tweet content using character trigrams as feature sets. The contribution of this paper is three-fold. First, we show how to predict follower levels, and then compare the predictability of different data points that are often used for the detection of latent user attributes in a second step. Thirdly, we report an interesting pattern of textual features that shows the correlation between Twitter specific functionality (retweeting, replying, etc.) and the number of followers. Follower relations seem to be the new currency in the social web. Our findings show how tweets reflect the popularity of users. Understanding relations between behavior and language usage will lead to an improved understanding of the users.

2 Related Work

Kwak et al. [9] rank user influence with follower counts and retweet rates, and compare the results with PageRank values for the nodes. In a similar approach, Cha et al. [10] rank user influence in three categories by counting followers, retweet rate (forwarding of tweets) and mentions (referring to someone), respectively. Quercia et al. [11] extracted the numbers of followers, followings and lists (how many times a user appears on other reading lists) to predict personality with regression analysis. Klotz et al. [12] applied clustering algorithms on Twitter user data to identify personality types. Five features were chosen related to behavior (numbers of tweets and favorites) or social network (same as [11]). The results of those studies demonstrate personality prediction based on publicly available data. Golbeck et al. [13] used multiple feature sets including behavioral, structural and textual data for regression. The results showed best predictability for openness and worst for neuroticism, but the prediction tasks beat the baseline only slightly. Their low performance may be a result of the diffuse feature sets creating additional noise. Pennacchiotti et al. [14] compare different feature sets to predict socio-demographic attributes (ethnicity, political affiliation, etc.) with Gradient Boosted Decision Trees. In all three tasks, linguistic features performed well, with hashtags and topic words adding beneficial information. Liu et al. [15] developed a threshold classifier to predict gender on Twitter that takes first names as features, and adds other features only in case of low distinctiveness. This second feature set contains textual features, e.g. character trigrams, which show good performance.

Although many studies utilize usage data and user-generated content to identify user characteristics, little is known about the relations between these data points. For many applications, it is useful to target the right people in the social network. Our study is a step forward in understanding relations between social behavior and language in online social networks.

3 Experimental Setup

3.1 Data Collection

A sample of 1400 users was gathered by using random numbers with up to eight digits, which has been acknowledged as a range for valid user-IDs [6, 10]. All tweets posted by these users during a one-month-sampling period (19.06–21.07.2012) were stored as data collection 1. An extended data collection contained all tweets posted during a two-month-sampling period (19.06–29.08.2012). After filtering for users with at least 20 English tweets, data collection 1 contained a total of 527 users and 77,131 tweets, and data collection 2 contained 547 users and 131,575 tweets. Histograms plotted in log-log-space for the number of followers and friends/followings show an approximately normal distribution, while the histograms for favorites, lists, and tweets per day are Zipfian.

3.2 Classification Method and Feature Selection

For a pilot study a naïve Bayes classifier is a beneficial approach. The probabilistic model assumes that features are independent which simplifies the estimation [16]. Although this is often not true, i.e. in case of natural language, the classifier shows good performance and may even outperform more sophisticated models [16].

We did Laplace smoothing to account for trigrams that occur only within tweets of the testing set. The smoothing factor was $\alpha = 0.005$ and the size of vocabulary was $V = 256^3$, because a character trigram can be any permutation of three characters.

The equation for our classification process is as follows:

$$\underset{level_j}{argmax} \left(\ln \left(P(level_j) + \sum_i^N \ln \left(P(trigram_i | level_j) \right) \right) \right) \quad (1)$$

with $$P(level_j) = \frac{number\ of\ users\ in\ level_j}{total\ number\ of\ users}$$

$$P(trigram_i | level_j) = \frac{(number\ of\ times\ trigram_i\ occurs\ in\ level_j) + \alpha}{(total\ number\ of\ trigram\ tokens\ in\ level_j) + V}$$

Text mining approaches differ in the complexity of the classification method, and can also be conducted on varying sizes of feature sets. A feature set may be too simple to capture signal, but also too diffuse causing it to create additional noise. When performing natural language processing on social media text, it should be kept in mind that most techniques are designed for more grammatical and well-formed text sources than for short, noisy ones [17]. We chose character n-grams, in particular trigrams, as a simple, but well-performing feature set. Character n-grams account for uncommon textual features such as "Emoticons, abbreviations, and creative punctuation use (which) may carry morphological information useful in stylistic discrimination" [18]. They also provide good word coverage, and preserve word order. The latter compensates for elaborate text-processing. Character trigrams have

been used in different prediction tasks on Twitter, e.g. as features for gender inference [15] and authorship attribution [18], and showed good performances. We used the frequencies of character trigrams as features in combination with techniques for noise reduction.

Messages on Twitter contain specific entities that refer to platform functionality. These are: the abbreviation 'RT' in retweets (forwarded messages), @-signs in replies and mentions, hashtags marking topics, and the inclusion of URLs. We filtered for these entities because they could create noise or provide false signaling. For example, including hashtag trigrams could result in subject matter detection, rather than user trait detection. However, the usage of platform functionality can be an important cue, so we replaced the entities by placeholders. This strategy led to the discovery of an interesting pattern for follower prediction (see Section 4.3).

3.3 Noise Reduction with Tf-idf

Tf-idf (term frequency-inverse document frequency) is a measurement for the representativeness of a token in a particular document in comparison with all documents of a corpus. The score increases with the number of times a token appears in the particular document (term frequency), but decreases with the number of times it appears in other documents of the corpus (inverse document frequency). The measurement is popular for keyword extraction, because high scores reflect representativeness of a term for the document, but low scores can be utilized for noise reduction.

We used augmented term frequency to prevent bias between users with many collected tweets (and therefore a long document) and users with few tweets. In this context, a document is the concatenation of all tweets we sampled for one user. As we tested our classifier with 10-fold cross-validation, we considered only the training set as corpus leaving out the documents of the test sample. This is the right way to perform cross-validation with multivariate classifier [16].

The scores are calculated as follows:

$$\textit{tf-idf-value} = TF * IDF \tag{2}$$

with

$$TF = 0.5 + \frac{0.5*(\textit{raw count of trigram in document})}{(\textit{raw count of most common trigram in document})}$$

$$IDF = \ln\left(\frac{(\textit{number of documents in corpus})}{(\textit{number of documents containing trigram})}\right)$$

For varying cutoff percentages $p=\{0, 0.05,\ldots,0.3\}$, any trigram whose tf-idf-value fell under the threshold was excluded from the feature set. These can be either trigrams that occur seldomly in one particular document (low term frequency) accounting for uncommon words/misspellings or very frequently in the corpus (low inverse document frequency) accounting for stopwords. As the idf-score decreases faster than the tf-score, we mostly filtered for very frequent words that do not bear much information.

4 Results

4.1 Prediction of Followers

To predict follower rates, we performed a supervised binary classification with a sliding $\lambda = 0.1, 0.2, \ldots, 0.9$ resulting in two bins with the ranges $[0, maximum * \lambda]$ and $]maximum * \lambda, maximum]$. This division technique produces unbalanced bins, so we limited the number of users in the training set so that each bin would have approximately the same number of users, giving baseline around 50%. We worked on data collection 2, to provide sufficient data. The naïve Bayes classifier used character trigrams as features. Platform specific entities were filtered out. To reduce noise, we filtered for trigrams with low tf-idf-scores as described in Section 3.3. We tested our classifier with 10-fold cross-validation. Table 1 gives the results for three runs with different setups.

Table 1. Binary classification of follower counts including noise reduction

	Bin division $\lambda = 0.4$ Baseline Accuracy = 51.7		Bin division $\lambda = 0.5$ Baseline Accuracy = 51.3		Bin division $\lambda = 0.7$ Baseline Accuracy = 51.5	
Tf-idf cutoff	**Accuracy**	**Improvement above Baseline**	**Accuracy**	**Improvement above Baseline**	**Accuracy**	**Improvement above Baseline**
0.00	63.9	23.6%	60.6	18.1%	58.2	13.0%
0.05	63.7	23.2%	60.4	17.7%	58.2	13.0%
0.10	63.5	22.8%	60.6	18.1%	58.2	13.0%
0.15	65.1	25,9%	61.7	20.3%	60.8	18.1%
0.20	65.5	26.7%	61.7	20.3%	60.9	18.3%
0.25	65.3	26.3%	60.9	18.7%	59.3	15.1%
0.30	65.0	25,7%	60.4	17.7%	59.9	16.3%

The classifier performed over baseline for all λ and best for $\lambda = 0.4$. With very low or very high values for λ the performance drops. The good overall performance in predicting five levels of followers (Section 4.2) makes it unlikely that very high or very low numbers of followers are harder to predict. The drop in performance is more likely caused by the reduction of training data when balancing the bins.

Furthermore, the reduction of noise with tf-idf filtering improves the performance. Social media texts are often noisy because of non-linguistic fragments as well as spelling mistakes [17], so the removal of least occurring trigrams like misspellings as well as stopwords bearing no information is a valuable approach. We got best results for a threshold of $p = 0.2$. When increasing the cutoff further, the performance is slightly reduced. As Twitter messages are very short and full of slang [17], a further reduction may remove information rather than noise. Furthermore, using too high a percentage may fit the classifier too specifically to the training set and not be effective for classification of the users in the testing subset. This is called overfitting.

The results of our experiment demonstrate that there is signal inside tweets to classify users according to their level of followers. For future research, a more advanced classification method could improve the performance.

4.2 Comparing the Predictability of Different User Traits

Approaches that predict user attributes mine behaviour data from online social networks as these sources have been proven to bear valuable information [19]. In [12], the five-dimensional feature space contains the tweet rate for communicativeness, the numbers of follower and following representing the position within the social network, the number of lists as further measurement for participation on the platform as well as bookmarking in terms of the number of favorites. We attempt to predict these five traits with textual features as explorations suggested correlations between behaviour and language usage. Each trait was considered an independent experiment and no interaction between traits was investigated. We computed the trait ranges, and divided the scales in five equal sized levels for a 5-way classification task. Since each level has an equal range of values, the boundary of $level_i$ is defined as:

$$\left[\frac{i}{5} * \text{maximum}, \frac{i+1}{5} * \text{maximum}\right) \tag{3}$$

where $i=0,1,\ldots,5$ and *maximum* is the maximum value for the specific trait. The task was performed as supervised 5-way classification with character trigrams as the features. Platform specific entities were filtered out, all other trigrams were preserved. We ran 10-fold cross-validation and used data collection 1, because the work reported in [12] is based on that sample. Table 2 shows the results in terms of accuracy (column 2) and improvement above baseline (column 4). Column 3 gives the accuracy baselines. Baselines are computed as the percentage of objects in the largest bin for each trait. The baselines differ, as the bins are equal sized in terms of range, but not in terms of number of objects, and the five traits possess different ranges. The prediction was successful for followers and for the number of lists. The other tasks failed.

Table 2. Results of 5-way-classification task for five different user traits

Trait	Accuracy	Baseline	Improvement above baseline
Followers	45.6	35.0	30%
Followings	56.8	58.4	-
Favorites	39.7	41.4	-
Lists	40.6	28.2	44%
Tweets Per Day	28.4	33.5	-

Textual features, in particular, character trigrams do not bear beneficial information for the prediction of followings, favorites and tweet rates. On the other hand, the content of tweets has signal to predict how many users will receive the tweet directly by following the blogger or indirectly by the number of lists the blogger is added to. The popularity of the blogger is obviously correlated with the content of the tweet.

4.3 Patterns of Specific Textual Features

The 5-way classification task for follower prediction uncovered an interesting pattern of textual features which may be a reason for the good classification results. The pattern is drawn in Figure 1. Whereas the average number of URLs in tweets increases with the level of followers, the correlation function between follower level and retweet rate is monotonically decreasing. URLs in tweets signal information sharing [4]. Retweeting behavior is associated with commenting or agreeing on messages, entertaining and showing presence, as well as information sharing [7], whereby the content is forwarded, not contributed. Our results indicate that users with lower follower counts are more likely promoters than originators of information. In other words, gaining followers is mainly achieved through producing original content.

At the same time, we found a nearly stable rate of @-references for all follower levels. Tweets containing the @-sign mark conversations [4], because the sign is used to explicitly reply to or mention someone else. The amount of @-signs did not differ significantly among the five follower levels, so the proportion of conversations is much likely the same for people with low follower vs. high follower rates.

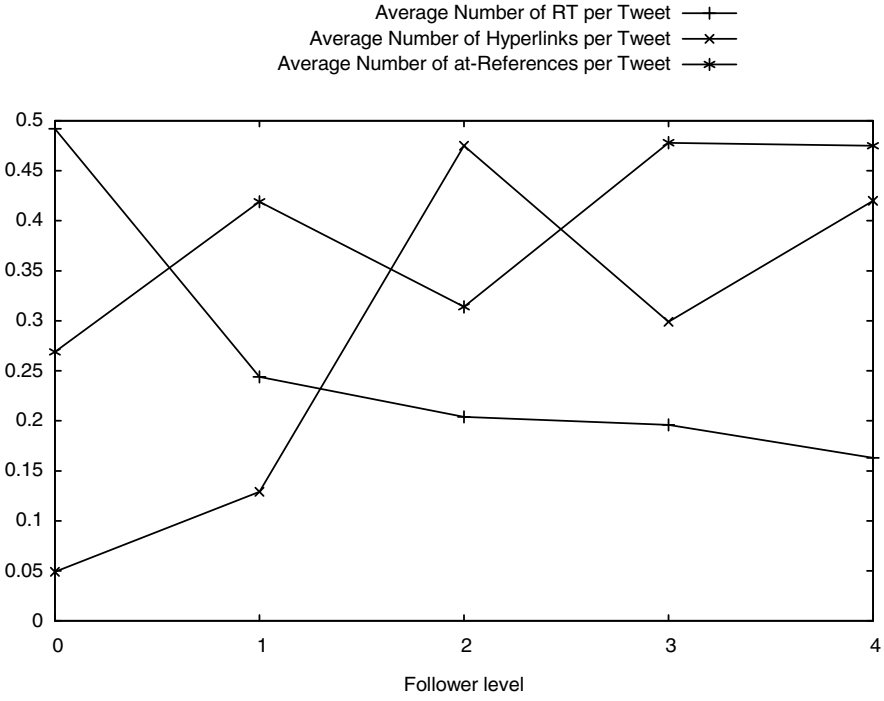

Fig. 1. Pattern of textual features

5 Conclusions

We reported an approach to predict the follower counts of Twitter users by looking at a small amount of their tweets. We used a naïve Bayes classifier with character trigrams as feature set, and reduced noise with tf-idf scores. Our results show that tweet content bears enough information for the prediction of followers and lists, but not for followings, favorites and tweet rates. The popularity of the blogger is obviously correlated with the content of the tweet. We also found a pattern of textual features that demonstrates the correlation between Twitter specific communication and the number of followers. Our results indicate that gaining followers is mainly achieved through producing original content and that the proportion of conversations is much likely the same for people with low follower vs. high follower rates. Our study is a step forward in understanding relations between social behavior and language in online social networks. In the future, more advanced classification methods could improve the performance, and uncover new relations between features and user classes.

References

1. Kaplan, A.M., Haenlein, M.: Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. Business Horizons 54, 253–263 (2011)
2. Hutto, C.J., Yardi, S., Gilbert, E.: A Longitudinal Study of Follow Predictors on Twitter. In: Mackay, W.E., Brewster, S.A., Bødker, S. (eds.) Proceedings of the 2013 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013), pp. 821–830. ACM (2013)
3. Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., Zhao, B.Y.: Follow the green: growth and dynamics in twitter follower markets. In: Proceedings of the 2013 Conference on Internet Measurement Conference, pp. 163–176. ACM (2013)
4. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: An Analysis of a Microblogging Community. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., Yen, J. (eds.) WebKDD/SNA-KDD 2007. LNCS, vol. 5439, pp. 118–138. Springer, Heidelberg (2009)
5. De Choudhury, M., Sundaram, H., John, A., Seligmann, D.: Dynamic prediction of communication flow using social context. In: Brusilovsky, P., Davis, H.C. (eds.) Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, pp. 49–54. ACM (2008)
6. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, pp. 19–24. ACM (2008)
7. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: Proceedings of the 43nd Hawaii International Conference on System Sciences (HICSS-43), pp. 1–10. IEEE Computer Society (2010)
8. Kwak, H., Moon, S., Lee, W.: More of a Receiver Than a Giver: Why Do People Unfollow in Twitter? In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM). The AAAI Press (2012)
9. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)

10. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in Twitter: The million follower fallacy. In: Cohen, W.W., Gosling, S.D. (eds.) Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, pp. 10–17. The AAAI Press (2010)
11. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In: 2011 IEEE International Conference on Privacy, Security, Risk, and Trust (PASSAT), and IEEE International Conference on Social Computing (SocialCom), pp. 180–185. IEEE (2011)
12. Klotz, C., Akinalp, C.: Identifying Limbic Characteristics on Twitter. In: Meesad, P., Unger, H., Boonkrong, S. (eds.) IC^{2}IT2013. AISC, vol. 209, pp. 19–27. Springer, Heidelberg (2013)
13. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting Personality from Twitter. In: 2011 IEEE International Conference on Privacy, Security, Risk, and Trust (PASSAT), and IEEE International Conference on Social Computing (SocialCom), pp. 149–156. IEEE (2011)
14. Pennacchiotti, M., Popescu, A.-M.: A Machine Learning Approach to Twitter User Classification. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) Proceedings of the Fifth International Conference on Weblogs and Social Media, Catalonia, Spain, July 17-21. The AAAI Press (2011)
15. Liu, W., Ruths, D.: What's in a Name? Using First Names as Features for Gender Inference in Twitter. In: Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium. The AAAI Press (2013)
16. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning. Data mining, inference, and prediction. Springer, Heidelberg (2009)
17. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How Noisy Social Media Text, How Diffrnt Social Media Sources? In: Mitkov, R., Park, J.C. (eds.) Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), pp. 356–364. Asian Federation of Natural Language Processing (2013)
18. Boutwell, S.R.: Authorship Attribution of Short Messages Using Multimodal Features. Naval Postgraduate School, Monterey (2011)
19. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 110, 5802–5805 (2013)