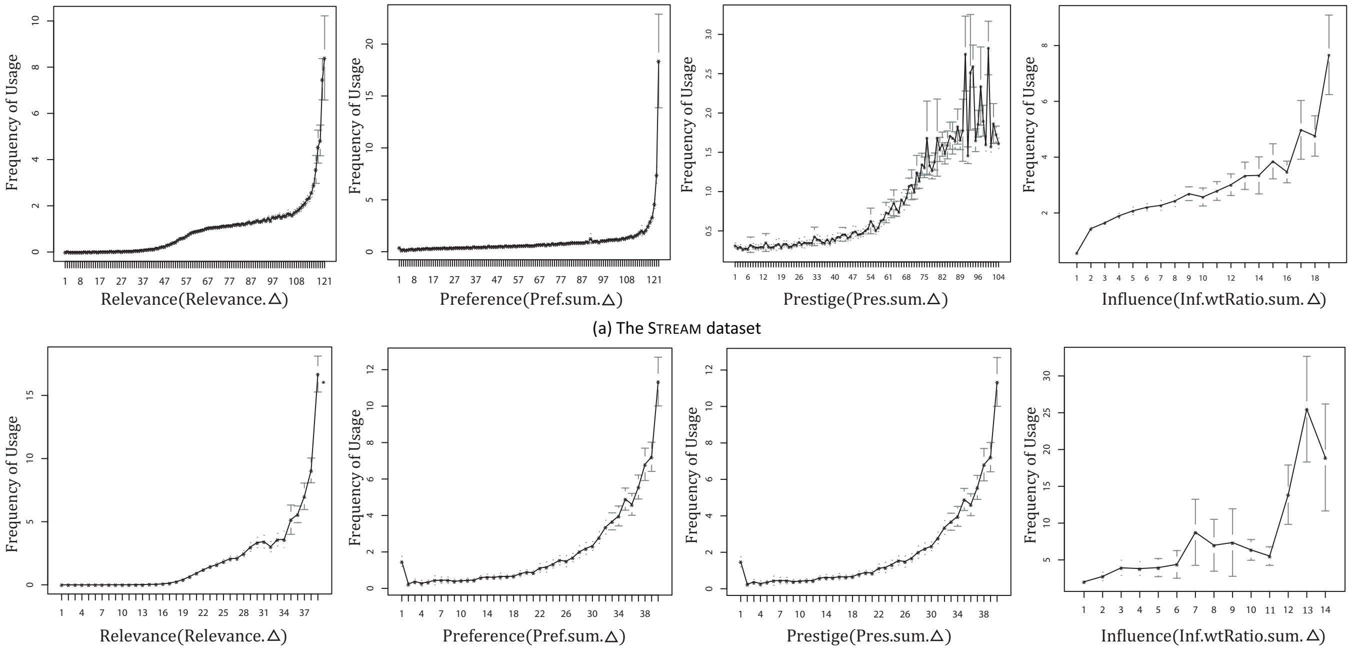

(a) The Stream dataset

(b) The Politics dataset

Figure 2: The degree of interest in hashtags correlates to the role-specific factors.

the end, we obtain 8,029 users for this group, and we name the group as the Movie group.

Random sampling

Another sampling strategy is to randomly select users from the entire dataset. We gather all users that appear in the Stream dataset, and randomly select a set of users who have posted at least one hashtag within 10 days prior to Δ_1. In the end, we collected 15,038 users for this group, which is named as the Random group.

5.1.2 Sample Negative Examples

Given a set of users, it is straightforward to generate data points (i.e., user-hashtag pairs) with a positive outcome of hashtag adoption. The data points with a negative outcome (i.e., the event that a hashtag is not adopted), however, is trickier. The simplest way is to create a negative example for each existing hashtag (i.e., within 10 days before the held-out period Δ_1) that a user did not use during Δ_1. However, the massive number of hashtags simply makes the negative examples dominating the collection, diluting all useful signal. Thus, for each user, we randomly sample negative examples proportional to the number of positive examples (i.e., number of hashtags adopted by this user).

5.2 Results of Regression Analysis

We employ *logistic* regression to predict the adoption of hashtags in the held-out time period Δ_1 from features computed based on Δ_0 prior to Δ_1. One instantiation of each measure is included. We set the time window Δ_0 as the period of 30 days prior to Δ_1. To comprehensively understand the predictive power of the role-specific measures, we also incorporate baseline features into the regression analysis. Table 2 presents the results of the regression analysis.

Apparently, all the four role-specific measures have a significant and positive predictive power of hashtag adoption, even when merged together with all baseline measures. One baseline measure, the variable *Popularity* in the Movie and Random group, yields a negative coefficient in the regression. This is because the popularity of a hashtag is highly correlated with the sum of prestige of hashtag users (e.g., with a correlation over 0.9196 in the Movie group). However, such correlation in the Politics group is lower, thus yielding a positive coefficient for *Popularity*. Indeed, when we remove the *Prestige* variable from the regression, the coefficient of *Popularity* becomes positive in the two groups (and remain significant). The *Length* of hashtags presents a negative relationship with hashtag adoption, possibly because people tend to adopt short and concise hashtags.

The results of the regression analysis thus provide a much stronger evidence that both the content role and the community role affect hashtag adoption. All four role-specific factors we presented have a significant positive predictive power to the adoption of hashtags.

6. PREDICTION ANALYSIS

The regression analysis has proved that all the role-specific measures are predictive to the adoption of hashtags. This reassures our hypothesis that the dual role of a hashtag affects the adoption. We are, however, moving forward to investigate the feasibility of constructing an effective prediction and recommender systems. Unfortunately, the regression analysis doesn't tell how an effective prediction system can be constructed based on these measures, or how hashtag recommendation can be done. To test the feasibility of hashtag recommendation, we proceed with building a machine learning model to predict the adoption of new hashtags in the future. Compared to the regression analysis that serves as a white box to interpret the effect of the measures, this analysis provides a black box that aims at optimizing the accuracy of of predictions.

Table 2: Regression Results.

Group	Feature abbr.	β	Significance
Politics	Popularity.Δ_0	2.911e-04	
	Indegree.Δ_0	-1.096e-01	
	Outdegree.Δ_0	-9.604e-02	
	Length	-4.345e-02	* * *
	N.uniTag.Δ_0	-2.831e-03	
	Inf.num.Δ_0	1.991e-00	* * *
	Pref.sum.Δ_0	1.211e-02	* * *
	Relevance.Δ_0	2.262e+01	* * *
	Pres.sum.Δ_0	8.466e+01	* * *
	Sample size = 3,753		
Movie	Popularity.Δ_0	-6.955e-05	* * *
	Indegree.Δ_0	-1.181e-03	* * *
	Outdegree.Δ_0	-1.975e-03	**
	Length	-4.613e-02	* * *
	N.uniTag.Δ_0	1.200e-03	* * *
	Inf.num.Δ_0	1.030e+00	* * *
	Pref.sum.Δ_0	2.472e-02	* * *
	Relevance.Δ_0	3.962e+01	* * *
	Pres.sum.Δ_0	8.086e+03	* * *
	Sample size = 26,188		
Random	Popularity.Δ_0	-3.512e-05	* * *
	Indegree.Δ_0	-3.777e-03	* * *
	Outdegree.Δ_0	-7.536e-04	* * *
	Length	-9.904e-02	* * *
	N.uniTag.Δ_0	1.107e-03	* * *
	Inf.num.Δ_0	1.720e+00	* * *
	Pref.sum.Δ_0	2.514e-02	* * *
	Relevance.Δ_0	5.186e+01	* * *
	Pres.sum.Δ_0	7.478e+03	* * *
	Sample size = 27,878		

Significant at the: *** 0.01, ** 0.05, or * 0.1 level. Bold: role-specific measures.

Table 3: Statistics of training and test datasets.

	Movie		Random		Politics	
All	Train	Test	Train	Test	Train	Test
# of (+)	13,071	11,932	13,969	12,393	1,886	1,086
# of (-)	13,117	11,884	13,909	12,464	1,867	1,071
NonRTs	Train	Test	Train	Test	Train	Test
# of (+)	6,348	5,912	8,093	7,233	1,612	928
# of (-)	6,358	5,842	8,106	7,272	1,600	931
RTs	Train	Test	Train	Test	Train	Test
# of (+)	7,332	6,550	6,368	5,536	335	207
# of (-)	7,397	6,618	6,346	5,472	334	208

NonRTs = Non-retweets, RTs = Retweets

Table 4: Accuracy of the SVM predictor.

Group	Measures	Accuracy (%)		
		All Tweets	Non-RTs	Retweets
Politics	(B)aseline	68.15	66.97	65.54
	B+Rel.	75.29 ***	74.23 ***	72.53 ***
	B+Pref.	70.84 ***	71.17 ***	67.23 ***
	B+Inf.	69.31 ***	68.42 ***	67.23 ***
	B+Pres.	75.52 ***	74.88 ***	71.32 ***
	All	**78.25** ***	**78.32** ***	**74.93** ***
Movie	(B)aseline	75.98	74.43	77.10
	B+Rel.	80.42 ***	78.93 ***	81.66 **
	B+Pref.	79.63 ***	77.66 ***	80.62 ***
	B+Inf.	79.93 ***	76.89 ***	81.04 ***
	B+Pres.	74.09 ***	71.57 ***	74.12 ***
	All	**80.64** ***	**79.13** ***	**82.80** ***
Random	(B)aseline	74.66	73.30	75.41
	B+Rel.	83.19 ***	82.64 ***	84.50 ***
	B+Pref.	81.39 ***	79.97 ***	83.39 ***
	B+Inf.	77.42 ***	75.56	80.18 ***
	B+Pres.	74.37 ***	73.39 ***	75.72 ***
	All	**84.03** ***	**82.45** ***	**85.64** ***

Significant at the: *** 0.01, ** 0.05, or * 0.1 level, paired t-test.
Rel.: Relevance; Pref.: Preference; Inf.: Influence; Pres.: Prestige

6.1 Experiment Setup

Consistent to the regression analysis, we setup three experiments over three groups of users, and sample negative examples accordingly. Held-out time periods are employed to surrogate the "future." The prediction problem is cast as a binary classification task: whether or not a user will adopt a new hashtag in a held-out period. The performance of such a classifier is evaluated by the accuracy of the predicted classes. Strictly, the behavior of a user to adopt a hashtag in her own tweets is quite different from adopting a hashtag by retweeting others. The former is spontaneous and the latter is passive. We further differentiate the tasks of predicting hashtag adoption in all tweets, retweets, and user-composed tweets (i.e., non-retweets) respectively.

Different from the regression analysis, we train the prediction model with a training dataset and assess the effectiveness with a separate test dataset. To do this, we select different pairs of "history" (Δ_0) and "future" (Δ_1) time windows. We then use some "history-future" pairs to train the prediction model, and use other time window pairs to test the model. From each pair of time windows, we sample a collection of user-hashtag pairs as data examples. For each data example, we then compute **all** the features in Table 1 based on Δ_0, and identify the label (positive or negative) based on whether the hashtag is adopted by the user during Δ_1. The number of positive and negative samples in our training and test datasets are presented in Table 3.

6.2 Results and Discussion

We first employ an SVM classifier with all baseline features in Table 1. The RBF kernel is adopted, and 5-fold cross-validation is used to select parameters. The performance of the SVM predictor is presented in Table 4. Such a baseline model performs reasonably well - with 68% accuracy on the Politics group and around 75% on both the Movie group and the Random group, comparing to a 50% accuracy of random guess. We then add the all features instantiating each of the four role-specific measures into the classifier, to test the four families of measures one by one, and then altogether. The inclusion of each of the four families of measures has improved the prediction performance significantly, with a few exceptions of the Prestige measures. In all three groups, the mixture of all four families of features has further improved the prediction accuracy.

The only exception is the *prestige* measure on the Movie and Random group. Surprisingly, the inclusion of the prestige measure even decreases the prediction performance when all tweets are considered. While on the Politics group, the prestige features performed significantly well. This is possibly due to the way prestige is computed. Of both the Movie group and the Random group, prestige is computed based on the global retweet network of millions of users, most of which are not part of the group. Therefore, the prestige is global and unspecific to the community of the group. Of the Politics group, however, the retweet network is specific to the community. The prestige is thus relatively "localized" and more specific to the user group [4]. Therefore, we further conducted an experiment to remove the prestige measures from the Movie and the Random group. The accuracy of the prediction task over Movie group increased from 80.64% to 82.00%, and the accuracy over Random group increased from 84.03% to 84.35%.

Interestingly, predicting the hashtag adoption in retweets seems to be easier than predicting the behavior in spontaneous tweets (Movie and Random). The exception is the Politics group, where the candidates prefer to tweet rather than to retweet (Table 3). In general, it is reasonable that the prediction of content in retweets is easier than the prediction of spontaneous tweeting behavior.

With all features included, the best prediction model achieves an accuracy around 80% among all datasets. This is a promising result, suggesting the feasibility of building effective recommender systems of hashtags in Twitter.

7. CONCLUSION

We presented a formal empirical analysis to test how the dual role of hashtags in Twitter affects hashtag adoption. Results of a correlation analysis, a regression analysis, and a prediction analysis all suggest that a hashtag serves as both a tag of content and a symbol of membership of a community. The measures we propose to quantify the factors of tagging content or joining communities all present significant predictive power to the adoption of hashtags. The prediction analysis using a SVM predictor provides a feasibility study of hashtag recommender systems, which suggests a promising future direction of research.

8. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under grants no. IIS-0968489 and IIS-1054199.

9. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*, pages 44–54, 2006.

[2] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35:29–38, 1992.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW '11*, pages 675–684, 2011.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: the million follower fallacy. In *ICWSM '10*, 2010.

[5] H.-C. Chang. A new perspective on twitter hashtag use: diffusion of innovation theory. In *ASIS&T '10*, pages 85:1–85:4, 2010.

[6] J. Chen, R. Nairn, and E. Chi. Speak little and well: recommending conversations in online social streams. In *CHI '11*, pages 217–226, 2011.

[7] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *CHI '10*, pages 1185–1194, 2010.

[8] D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *CoNLL '10*, pages 107–116, 2010.

[9] L. D'Monte. Swine flu's tweet tweet causes online flutter. *Business Standard*, 2011.

[10] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[11] M. Ebner, H. Mühlburger, and et al. Getting granular on twitter : tweets from a conference and their limited usefulness for non-participants. *KCKS '10*, pages 102–113, 2010.

[12] G. Golovchinsky and M. Efron. Making sense of twitter search. In *CHI '10 Workshop on Microblogging*, 2010.

[13] M. Graves. The 2010 world cup: a global conversation. *Twitter Blog*, 2010.

[14] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *SIGIR '09*, pages 540–547, 2009.

[15] M. Gupta, R. Li, Z. Yin, and J. Han. Survey on social tagging techniques. *SIGKDD Explor. 10*, 12:58–72, 2010.

[16] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys '10*, pages 199–206, 2010.

[17] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *CHI '11*, pages 237–246, 2011.

[18] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08*, pages 531–538, 2008.

[19] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *HT '10*, pages 173–178, 2010.

[20] R. Jäschke, L. Marinho, A. Hotho, S.-T. Lars, and S. Gerd. Tag recommendations in folksonomies. In *PKDD '07*, pages 506–514, 2007.

[21] Y. Jiang, C. X. Lin, and Q. Mei. Context comparison of bursty events in web search and online media. In *EMNLP '10*, pages 1077–1087, 2010.

[22] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how twitter is used to widely spread scientific messages. In *WWW '10 Workshop WebSci10*, 2010.

[23] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, 2007.

[24] A. Livne, M. Simmons, E. Adar, and L. Adamic. The party is over here: structure and content in the 2010 election. In *ICWSM '11*, 2011.

[25] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT '06*, pages 31–40, 2006.

[26] Q. Mei, D. Zhang, and C. Zhai. A general optimization framework for smoothing language models on graph structures. In *SIGIR '08*, pages 611–618, 2008.

[27] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[28] R. Noon and H. Ulmer. Analyzing conferences in twitter with social aviary. *Stanford University CS 322*, 2009.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

[30] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: identifying misinformation in microblogs. In *EMNLP '11*, pages 1589–1599, 2011.

[31] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *CSCW '94*, pages 175–186, 1994.

[32] E. M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 5th edition, 2003.

[33] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In *CSCW '06*, pages 181–190, 2006.

[34] K. Starbird and L. Palen. "voluntweeters": self-organizing by digital volunteers in times of crisis. In *CHI '11*, pages 1071–1080, 2011.

[35] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *KDD '10*, pages 1049–1058, 2010.

[36] C. White. Reaching 200 million accounts: twitter's explosive growth. *Mashable*, 2011.