Proceedings of the Ninth International AAAI Conference on Web and Social Media

# A Stratified Learning Approach for Predicting the Popularity of Twitter Idioms

**Suman Kalyan Maity, Abhishek Gupta, Pawan Goyal,** and **Animesh Mukherjee**
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India - 721302
Email: {sumankalyan.maity, abhishek.gupta, pawang, animeshm}@cse.iitkgp.ernet.in

## Abstract

Twitter Idioms are one of the important types of hashtags that spread in Twitter. In this paper, we propose a classifier that can stratify the Idioms from the other kind of hashtags with 86.93% accuracy and high precision and recall rate. We then learn regression models on the stratified samples (Idioms and non-Idioms) separately to predict the popularity of the Idioms. This stratification not only itself allows us to make more accurate predictions but also makes it possible to include Idiom-specific features to separately improve the accuracy for the Idioms. Experimental results show that such stratification during the training phase followed by inclusion of Idiom-specific features leads to an overall improvement of 11.13% and 19.56% in correlation coefficient over the baseline method after the $7^{th}$ and the $11^{th}$ month respectively.

## Introduction

Twitter Idioms are hashtags referring to a conversational and personal theme that constitutes of a concatenation/merger of at least two words e.g., #10ThingsAboutMe, #4WordsAfterABreakup, #ICantForgetAboutYou. The merger usually does not include names of people or places and the full phrase is not a proper noun or a reference to the title of a song, movie or organization. However hashtags like #IWishICouldGoToBahamas, #10ThingsToDoIfYouAreBelieber etc. are Idioms as they convey wish or desire to be a person or to be at places etc. in contrast to hashtags that convey support for an organization, person, places like #SaveNHS, #SaveAssange etc. that do not qualify as Idioms. We consider hashtags like #peopleoverpolitics, #MusicMonday etc. to be topical (politics, music respectively) and not Idioms.

We consider a two and a half year long 1% Twitter random sample and observe that the Idioms constitute ∼17-25% of this sample in Twitter data stream. If we consider month-wise top 100 hashtags, then we observe ∼20-30% hashtags to be Idioms. These Twitter Idioms are usually used by people for day-to-day gossip, showing personal feelings etc. Therefore, if one needs to understand human conversation dynamics, one has to study the Idioms in isolation. This could open the gateway to understand the opinion, sentiments and emotions of people in more focussed way instead of looking into the entire gamut of Twitter hashtags which researchers have been doing over the years. The above hypothesis is also supported by our initial investigation on the dataset where we found that Idiom related tweets contain almost twice the fraction of opinionated words compared to that in other hashtag related tweets as evidenced through the comparison with the opinion dictionary [1]. One of the first steps to understand this rich data of conversation dynamics is to know what are the popular Idioms, the factors that make these kind of hashtags popular. (Romero, Meeder, and Kleinberg 2011) observed that different topical category (sports, music, Idioms) of hashtags have different propagation pattern. Therefore, there is a natural intuition that they would gain popularity through different mechanism as the underlying spreading pattern is different. These observations form the central motivation of our current work where we investigate the detailed mechanics of the spread of the Idioms and, subsequently, develop a model that can automatically predict at different future time points, those Idioms that are going to be popular. Prediction of Idioms at different time points can be directly applied to identify the temporal scope of sentiments/opinions in the associated tweets and the way the scope changes. This also helps us understand the community sentiment as people usually follow the trending Idioms.

## Dataset description

Twitter provides 1% random sample of all the tweets via its sample API in real time. We consider tweets from $1^{st} August, 2011$ to $31^{st} May, 2014$ for authors who have mentioned English as their language in their profile and then performed a second level filtering of the tweets by a language detection software (Lui and Baldwin 2012). We then separate out the hashtags that have been coined between $1^{st} January, 2012$ and $31^{st} May, 2014$ (or were absent in the time period from $1^{st} August, 2011$ to $31^{st} December, 2011$) and have frequency of at least 100. From this hashtag set, a random sample of 9000 hashtags is chosen and labeled manually by two of the authors (over 90% agreement) into two categories 'Idiom' and 'non-

[1] http://www.cs. uic.edu/∼liub/FBS/sentiment-analysis.html

Idiom'. Out of these 9000 hashtags, 1575 hashtags are found to be 'Idioms'. In order to deal with the imbalance between the two categories, we randomly choose 1575 non-Idioms from the set of labeled non-Idioms. From this set of 3150 hashtags, we found some hashtags with 90% of their tweets as the same retweet. We remove these hashtags and finally we have 2931 hashtags. In Table 1, we show some example hashtags from Idiom and non-Idiom categories.

| Category | Examples |
|---|---|
| Idioms | GuysBeLike, TeenagersFact, 10ThingsIHateAboutMyself, TuesdayTease, Things90skidsSaid, WorstBreakupExcuse |
| non-Idioms | SummerBall, TVDParty, SussexHour, GOTSeason4, Election2013, BelieveMovie |

Table 1: Examples of Idioms and other type of hashtags.

## Experimental Framework

Our goal is to predict the popularity of Idiom hashtag $h$ at a time period $\Delta t$ after its birth. We approach the problem in two stages. In the first stage, we train a classifier with a set of features (first level) to separate the test sample into Idiom and non-Idiom category. We use Support Vector Machine (SVM) and logistic regression for the classification. In the next stage, we attempt to predict the popularity of the Idiom class by learning a regression model with a set of specialized features (second level) to predict the popularity (see fig 1). The normalized count (Tsur and Rappoport 2012) which we presume as a measure of popularity is defined as follows:

$$N(h^i) = \sum_{j \in months} count(h^i_j) \times \frac{m'}{m_j}$$

where $count(h^i_j)$ is the number of appearances of the hashtag $h^i$ in the $j^{th}$ month. $m_j$ is the sample size of new hashtags that appear in month $j$. $m'$ is a constant whose value is considered to be $m_1$, i.e., the sample size of these hashtags in the $1^{st}$ month. We use Support Vector Regression (SVR) for predicting the log normalized counts.

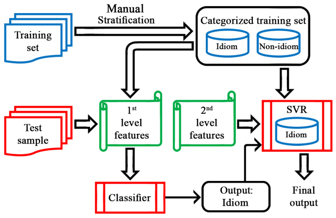

Figure 1: (Color online) A schematic of our proposed framework.

## Hashtag Classification

For the task of classification, we use three major types of features - Hashtag content, Tweet content and User features. These are:

**Hashtag character length**

**Number of words in the hashtag**

**Avg. and maximum character length of the words in the hashtag**

**Presence of days of the week and numerals in the hashtag**

**Presence of constituent words of the hashtag in the dictionary**

**Hashtag collocation**

**Parts-of-Speech tag diversity** - We use CMU POS tagger (Owoputi et al. 2013) for identifying the POS tags after segmenting the hashtag. We consider 14 important POS tags as feature for our classification model. We define the POS diversity (PosDiv) as $PosDiv(h_i) = -\sum_{j \in pos_{set}} p_j \times \log(p_j)$ where $h_i$ is the $i^{th}$ hashtag and $p_j$ is the probability of the $j^{th}$ POS in the set of POS tags. We use this diversity metric as a feature for our classifier.

**Presence of common personal pronouns, verbs in the hashtag**

**Presence of n-grams in English texts** - We segment the words in the hashtags and search for 2, 3, 4, 5 grams of the constituent words in the corpus of 1 million contemporary American English words[2]. We use the presence of any of these n-grams as a feature for the classifier.

**Hashtag Clarity** - Hashtag clarity as defined in (Ma, Sun, and Cong 2012)

**Word Diversity** - If $H_i$ is the document containing all the tweets in which hashtag $i$ appears and $p(w|H_i)$ is the probability of a word belonging to the document $H_i$ then word diversity is defined as $WordDiv(i) = -\sum_{w \in H_i} p(w|H_i) \times \log p(w|H_i)$.

**Topical Diversity** - We adopt Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), a renowned generative probabilistic model for discovery of latent subtopics. We set number of subtopics as K = 10, 20, 30 and find out $p(topic_k|H_i)$ for a document $H_i$ containing all the tweets in which the $i^{th}$ hashtag appears and then compute the topical diversity ($TopDiv$) for $i^{th}$ hashtag as $TopDiv(i) = \sum_{k=1}^{K} p(topic_k|H_i) \times \log p(topic_k|H_i)$.

**Cognitive Dimension** - The cognitive dimension (linguistic and psychological) for different kind of hashtags is captured through the different categories provided by LIWC software (Pennebaker, Francis, and Booth 2001).

**In-vocabulary words and Out-of-Vocabulary words ratio in the tweets for a hashtag**

**Mention multiplicity** - Fraction of tweets containing the hashtags in which no mention, one mention, two mentions, three or more mentions are used.

**Retweet multiplicity** - Fraction of times a tweet containing the hashtag is retweeted multiple times. We have used 4 features corresponding to different retweet multiplicities (0, 1, 2, 3 or more).

[2] http://www.ngrams.info/samples_coca1.asp

## Performance evaluation of the classifier

We use SVM and logistic regression classifier available in Weka Toolkit (Hall et al. 2009). We perform a 10-fold cross-validation on the entire data and achieve 86.9% accuracy with high precision and recall rates (see Table 2 for details). Both the classifiers yield very similar classification performance. We observe that the number of topics of LDA does not have significant effect on the classification results. Therefore, for the second stage prediction we consider no. of topics as 10 and train the SVM classifier on separate training set and test it on a separate test set. We choose a random sample of 1000 hashtags as training set and another 1000 hashtags as test set. We achieve an accuracy of 86.1% on this test sample. Hashtag content features (hashtag length, POS tag diversity, n-gram presence, hashtag collocation etc.) are the most discriminative ones followed by tweet features (especially the LIWC features).

| Method | Classifier | $K$ | Accuracy | Precision | Recall | F-Score | ROC Area |
|---|---|---|---|---|---|---|---|
| 10-fold Cross Validation | SVM | 10 | **86.9** | 0.868 | 0.868 | 0.868 | 0.868 |
| | | 20 | 86.25 | 0.863 | 0.863 | 0.863 | 0.862 |
| | | 30 | 86.49 | 0.865 | 0.865 | 0.865 | 0.865 |
| | Logistic Regression | 10 | 86.76 | 0.868 | 0.868 | 0.868 | 0.939 |
| | | 20 | 86.49 | 0.865 | 0.865 | 0.865 | 0.938 |
| | | 30 | 86.46 | 0.865 | 0.865 | 0.865 | 0.938 |
| Separate training and test set | SVM | 10 | **86.1** | 0.861 | 0.861 | 0.861 | 0.862 |
| | Logistic Regression | 10 | 86.01 | 0.86 | 0.86 | 0.86 | 0.933 |

Table 2: Performance of various classifier for different topic selection for LDA feature with number of topics ($K$ = 10, 20, 30).

## Predicting the popularity of Twitter Idioms

In this section, we learn a regression model (SVR) to predict the popularity of the classified Twitter Idiom hashtag set by the SVM classifier within a given time frame $\Delta t$. The popularity of a hashtag within time window $\Delta t$ is defined by log normalized count of the hashtag within the same time window (Tsur and Rappoport 2012). We use the features described in (Tsur and Rappoport 2012) as baseline features for our model.

**Idiom-specific features:** We incorporate a set of new features which are instrumental in popularity prediction for Idioms. We categorize these features in four groups: hashtag content features, tweet content features, user and temporal features.

**Hashtag content features** These features are related to the words that constitute the hashtag.

*Presence of common verb and personal pronouns*

*Frequency of the n-grams in the English texts*

*Diversity of the POS tags of the words in the hashtag*

*Hashtag repetition* For each hashtag, we calculate fraction of tweets in which the hashtag appears multiple times. We use this hashtag repetition ratio as a feature for our regression model.

*Word clarity in a hashtag* Word clarity is a similar measure to the hashtag clarity. It is formally defined as KL divergence of word distribution within the hashtag and global word segment collection containing all the hashtags $W$; $WClarity_i = -\sum_{w \in D_i} p(w|D_i) \times \log \frac{p(w|D_i)}{p(w|W)}$ where $w$ are the words within the word collection of the hashtag ($D_i$), $p(w|D_i)$ is the local probability of a word to be present in the hashtag and $p(w|W)$ is the probability of belongingness of a the word $w$ in the segmented word pool.

**Tweet content features** These features are related to the content of the tweet in which the hashtag appears.

*Hashtag clarity* Hashtag clarity is an important feature in predicting user adoption of a hashtag (Ma, Sun, and Cong 2012).

*Pagerank of the hashtags* We constructed a co-occurrence graph considering all the words and hashtags in the tweet collection. An edge represents collocation in the same tweet. The number of collocations serves as edge weight. We prune out all such edges which have very small weight. We then run Pagerank algorithm on this graph to obtain the pagerank values of all the hashtags. This measure indicates how "central" a hashtag is within the collection of tweets.

**User features** User's tweeting behavior is also instrumental in propagation of hashtags. Here, we describe user features.

*Mention Multiplicity* We measure fractions of tweets in which no mention, one mention, two mentions and three or more mentions are made. Higher the mention multiplicity, higher is the chance that a hashtag will gain popularity.

*Retweet Multiplicity* Like mentions, we measure retweet multiplicity as fraction of tweets in which no, one, two and three or more retweets been made. We use these 4 attributes as 4 different features.

**Temporal features** In the prediction of the popularity of a hashtag, the early patterns of spread are important in knowing how the hashtag will survive in long term. In this subsection, we shall discuss some of these interesting temporal features.

*Early coinage* The hashtags that are coined earlier have higher chance to survive if there are potential competitors. For example, #ThingsAboutMe has multiple variants #10ThingsAboutMe, #20ThingsAboutMe etc. Similarly, #FlashbackFriday has all other possible variants with other days of the week. We ranked the hashtags according to their time of birth and use normalized rank as a feature.

*Exposure to different set of users* In order to identify the level of exposure of hashtags, we compute the fraction of unique users who have tweeted using a hashtag in the first 50 tweets. This serves as a feature for our model. We also find out number of unique users who have been mentioned by someone in a tweet containing the hashtag and number of users who have been exposed multiple times in the first 50 tweets as other two features.

*Time gap between first few tweets* We consider first few tweets (41 tweets) for each hashtag. We then find out the time difference between each of these tweets at leads of 5. We calculate percentage change in the time i.e, $\frac{t_6 - t_1}{t_1}$, $\frac{t_{11} - t_6}{t_6}$, $\frac{t_{16} - t_{11}}{t_{11}}$ . . . . In summary, we have 8 such percentage change values as features.

### Performance of our regression model

For a regression task, the most significant evaluation metrics are correlation coefficient ($\rho$) and root mean square error ($\theta$). In this paper, we predict the popularity value at two different time instances (after 7 months and after 11 months). First, we apply the baseline features to train the regression model on 1000 training samples and test on a 1000 test sample. We achieve a correlation coefficient and root mean square error of 0.6058 and 0.7674 respectively. We then use our first-stage classifier to stratify the test samples into Idioms and non-Idioms. On this separated Idiom test set ($\sim$ 500), we use the baseline features using same sized training sample of Idiom data selected from the corpus. This stratification helps in improving the correlation coefficient ($\rho$) by 7.1% and reducing the root mean square error ($\theta$) by 6.36%. As a subsequent step, we separately apply the Idiom specific features on the same training and test data. Incorporating these features boosts up the correlation coefficient by 3.76%, and reduces the root mean square error by 2.25% (see table 3 for details). Overall, this stratification strategy helps in improving correlation coefficient from 0.6058 to 0.6732, i.e., by 11.13%. We also learn the regression model for predicting the popularity value at a longer time point (after 11 months). The improvement is relatively higher in comparison to the short range predictions[3]. The stratification approach improves the correlation coefficient by 12.2% and the incorporation of the Idiom specific features boosts it by another 6.55% leading to an overall improvement of 19.56% in correlation coefficient over the baseline method. We use $RELIEFF$ feature selection algorithm in Weka Toolkit to rank the features. The rank order indicates that for popularity prediction (at least in the early stages) usually the user and the temporal features are more important than content features. For prediction at later time point, certain content features like hashtag clarity also gain significant importance. This possibly points to the fact that as hashtags grow old, their contents become an important determinant of their popularity.

## Conclusions

In this paper, we have adopted a two-stage framework for predicting popularity of Twitter Idiom hashtags at different time intervals after the birth of the hashtags. We developed a binary classifier incorporating a wide variety of features to stratify the hashtag set into disjoint Idiom and non-Idiom set. The classifier achieves 86.9% accuracy with a high precision and recall rate. We observe that the hashtag content features have the most discriminative power compared to the other types of features.

In the first stage of stratification, if one applies the baseline method on the stratified non-Idiom set, the correlation coefficient decreases. For predicting the popularity after $7^{th}$ month, we observe that on the stratified non-Idiom set, the correlation coefficient decreases from 0.6058 to 0.5917 and the root mean square error increases to 0.7756 from 0.7674. A possible reason for this phenomena is that the baseline features are more suitable to the Idiom class compared to the non-Idiom class. However, we would like to remark that the type of hashtags belonging to this non-Idiom class are mostly event-based, so their popularity indicators are different.

| Time frame | Method | $\rho$ | $\theta$ | % improvement in $\rho$ | % improvement in $\theta$ |
|---|---|---|---|---|---|
| After 7 months | Baseline | 0.6058 | 0.7674 | | |
| | Stratification | 0.6488 | 0.7186 | 7.1 | 6.36 |
| | **Idiom specific features** | **0.6732** | **0.7024** | **11.13** | **8.4** |
| After 11 months | Baseline | 0.3932 | 0.7635 | | |
| | Stratification | 0.4412 | 0.7321 | 12.2 | 4.11 |
| | **Idiom specific features** | **0.4701** | **0.7121** | **19.56** | **6.73** |

Table 3: Performance of the regression model. The % improvements are shown over the baseline.

## Acknowledgments

The authors would like to thank Prof. Chris Biemann, TU Darmstadt for providing them with a historical Twitter 1% random sample data. This work has been supported by Microsoft Corporation and Microsoft Research India under the Microsoft India PhD fellowship Award.

## References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.* 11(1):10–18.

Lui, M., and Baldwin, T. 2012. Langid.py: An off-the-shelf language identification tool. ACL '12, 25–30.

Ma, Z.; Sun, A.; and Cong, G. 2012. Will this #hashtag be popular tomorrow? SIGIR '12, 1173–1174. New York, NY, USA: ACM.

Owoputi, O.; Dyer, C.; Gimpel, K.; Schneider, N.; and Smith, N. A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. NAACL '13.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. *Linguistic Inquiry and Word Count.* Mahwah, NJ: Lawerence Erlbaum Associates.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. WWW '11, 695–704. New York, NY, USA: ACM.

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. WSDM '12, 643–652. New York, NY, USA: ACM.

[3] We also did experiments on a much shorter range (predicting at $10^{th}$ week after observing 6 weeks data), where also we achieve improvement over the baseline; however, the improvement is relatively less due to scarcity of data per hashtag on weekly basis.