Extracting Social List Hashtags from Twitter

Ankan Mullick*, Pawan Goyal*, Niloy Ganguly*, Manish Gupta†
*IIT Kharagpur, India. {ankanm, pawang, niloy}@cse.iitkgp.ernet.in
†Microsoft, Hyderabad, India. gmanish@microsoft.com

***Abstract*—Social list queries like 'valentines day gift ideas', 'best anniversary messages for your parents', etc. are quite popular on web search engines. Users expect instant answers comprising of a list of relevant items (social list) for such a query. Surprisingly, current search engines do not provide any crisp instant answers for queries in this critical query segment. To the best of our knowledge, we propose the first system that tackles such queries. Although such social factors are heavily discussed on online social networks like Twitter, extracting such lists from tweets is quite challenging. How to discover such lists from tweets? We present a system that identifies these 'social lists' from a large number of Twitter hashtags using a high recall classifier trained using novel task-specific features with good accuracy. Further, we briefly discuss how list items can be extracted from related tweets. Experiments over a dataset of ∼4M tweets show that our recall-optimized system can obtain up to 75.5% precision at 95.3% recall.**

I. Introduction

In terms of search volume, one particular segment of queries that has grown significantly over the years is the segment of social list queries. Social list queries seek a list of items as an answer and are often related to social topics. A few examples include "greeting lines for your boss", "ways to wish good morning", "things to do in goa", etc. Users put up such queries in the hope of obtaining innovative, witty, popular, and informative answers from opinions and experiences expressed on matching webpages. While some of these queries seek information about social events (e.g., "things to ask an interior decorator", "things to remember for wedding planning"), many others are related to social situations and moods (e.g., "first date questions", "tricks to feed your baby", "lies girls use a lot"). Fig. 1 shows percentage of 'social list queries' on Bing since 2011. It shows significant increment of 36.7% and 25.9% in 2015 and 2016 respectively[1].

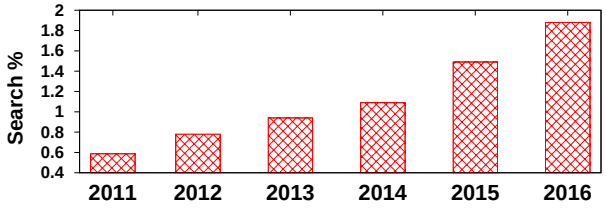

Fig. 1: Year-wise Percentages of 'Social List' Queries

[1]These values were computed by taking the Bing query log for Dec 1 for each year, which was filtered using patterns for list type queries. A random sample of 200 queries from this filtered set were then manually labeled as 'social list queries' or not. The % shown in the graph denotes the fraction of queries satisfying the filter among all the queries times the fraction of queries labeled as 'social list queries' in the random sample.

While search engines are effective at showing instant answers for factual queries like "weather paris" or "temperature sunnyvale", they hardly support instant answers for such social list queries.

For a large number of other social lists we queried, no good list items can be obtained even after browsing through the top few results[2]. However, such social topics are discussed extensively on social media like Twitter, Quora etc. For example, Fig. 2 shows few tweets for the query "advice for young journalists". The aim of this paper is to retrieve structured information from those discussions and produce a list of items corresponding to each social list. We implement the methodology on Twitter data.

Fig. 2: Some Tweets for "advice for young journalists"

Twitter hashtags can be clearly exploited towards shortlisting items for such social list queries. Among different types of conversational tweets, a class of tweets discusses 'social topics' - we term the hashtags anchoring them as social list hashtags (SL-hashtags in short). Examples of SL-hashtags are #10tipsforinteriordesign, #5thingspeoplehate, #adviceforyoungjournalists, etc. Formally, we define SL-hashtags as a specific category of hashtags, where users discuss about a list of things, specific to the underlying social topic. SL-hashtags can mainly be of two types - 1) Objective - that contain factual objects as list items, e.g., names of places in #placestovisitinjapan, or movie names in #3moviesthatyoulove. 2) Subjective - that contain list items conveying people's views on a topic, e.g., #childhoodmemories comes with subjective list items that convey various childhood memories.

There are however several challenges which need to be addressed. (a) SL-hashtags are a miniscule minority among all hashtags, hence detecting these sparse identities is difficult. We need a system to detect as much social hashtags as possible

[2]Please check the webpage http://www.cnergres.iitkgp.ac.in/sociallist/ for examples

i.e. a high recall system. (b) Tweet texts are unstructured and contain noisy texts, URLs, so extraction of valid list items from the unstructured texts and URLs is nontrivial. (c) The valid list items derived may not be relevant. For example, corresponding to 'dinner plans' query, the derived item *why dont you tell me about #dinnerplans* doesn't produce any relevant information, hence deriving relevant items from the valid list is challenging. (d) Finally several SL-hashtags occur only in few tweets leading to retrieval of only a small number of relevant list items. The challenge lies in populating such tail lists with relevant list items from other similar hashtags.

In this paper, we work towards addressing the first challenge, i.e., SL-hashtag discovery, and briefly discuss about extraction of list items. Overall, we make the following contributions.

- We discuss a large set of hashtag and tweet-level features to build a classifier to extract social list hashtags from Twitter.
- We experiment with multiple classifiers using a dataset of ∼4M tweets. Our recall optimized classifier provides a precision of 75.5% at a recall of 95.3%.
- We briefly discuss about extraction of list items for the identified social list hashtags.
- We manually curated 1001 social list hashtags, which we release publicly along with code[3].

The paper is organized as follows. We discuss related work in Section II. We present dataset details and design the social list hashtag detection classifier in Sections III and IV respectively. In Section V, we briefly discuss about extraction and ranking of list items for these list hashtags. We conclude with a summary in Section VI.

II. Related Work

While none of the previous works have studied 'social list hashtags', Twitter Hashtags, including the idioms, have been well studied. Romero, Meeder and Kleinberg [1] studied the problem of categorizing Twitter hashtags into various categories such as Celebrity, Sports, Idioms, etc. and studied the difference in the information diffusion mechanism across these categories. Lee et al. [2] studied the problem of classification of Twitter trending topics/ hashtags across multiple categories. After the idiom hashtags were defined in [1], several works looked into the problem of classifying idioms from general hashtags [3], [4]. None of these works, however, have focused on 'social list hashtags'. We present an approach to identify social list hashtags from the other hashtags with significant accuracy, recall and precision.

III. Dataset

We collected a dataset of the most trending ∼4M Twitter hashtags with a minimum frequency of 20 from Jan-Jun 2015 using the Twitter Streaming API[4]. Further, since the set of SL-hashtags is a subset of idioms, we first use a state-of-the-art algorithm [3] for selecting ∼67K idioms from the ∼4M hashtags. Then we manually identified 1001 SL-hashtags from these idioms. Of these, 589 are subjective and 412 are objective. The annotation guidelines were to select those hashtags as SL-hashtags, which can lead to list type answers. To learn a classifier, we also manually identified 1001 hashtags which are not SL-hashtags to obtain a balanced dataset of a total 2002 hashtags.

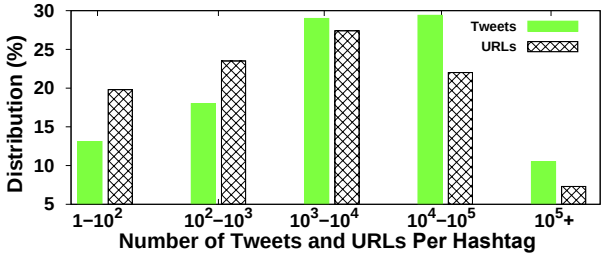

Fig. 3: Distribution of Tweets and URLs for our Dataset

We also collected English tweets corresponding to the 1001 SL-hashtags. The dataset contains ∼204.43M tweets and ∼85.07M URLs extracted from these tweets. Detailed statistics of the tweets and URLs distribution for the SL-hashtags are shown in Fig. 3. X-axis represents various ranges of tweets and URLs for any social list hashtag, and Y-axis represents percentage of total hashtags, which have tweets / URLs in this range. For example, only 13.1% of total hashtags contain tweets within range (1-100) while 19.3% of total hashtags contain URLs in range (1-100).

IV. SL-Hashtag Detection

In this section, we discuss the design of a classifier that can separate SL-hashtags from the rest. We first perform two pre-processing steps.

- Segmentation: Not all hashtags are in CamelCase style (e.g., #BeTheHope) so it is not trivial to identify constituent words from a hashtag. We segment each hashtag using a modified version of the Viterbi Algorithm [5] and the Google n-gram corpus[5].
- Parts Of Speech (POS) Tagging: We use CMU POS Tagger for tweets [6] to identify different POS tags, @-mentions, URLs, etc.

After data pre-processing, we extract the following hashtag features.

A. Features

We extract three broad categories of features: Language features, Search features, and Tweet features.

Language Features

We discuss these features in detail in the following.
(a-b) Hashtag Length in characters (a), and in words (b).
(c) Presence/absence of names of days (Sunday, Monday etc).
(d-i) Presence of some POS tags are considered as binary features: preposition (d), interjection (e); whereas some are

[3]https://goo.gl/wJfTmC
[4]https://dev.twitter.com/streaming/firehose
[5]http://books.google.com/ngrams

considered as numeric features: the count of nouns (f), pronouns (g), adjectives (h) and verbs (i).
(j) Entropy of word frequency distribution across 14 POS tags.
(k) Ratio of count of vocabulary words to count of out-of-vocabulary words.
(l) Presence/absence of numbers in the hashtag.
(m) Presence/absence of plural nouns in hashtag, e.g., 'lines' in #10breakuplines is the plural form of 'line'. Similarly, in #3thingsthatmakeyousmile, 'things' is the plural form of 'thing'.
(n) Presence of Category Match: This binary feature captures the presence/absence of any regex pattern that usually matches an SL-hashtag. Examples of patterns are as follows: (i) wh_word-*-verb: #howtobreakupin5words, #whatmakesgodsmile, #10peoplewhomeanalottome; (ii) **-in-⟨num⟩-words: #worstdayin4words, #lovestoryin5words; (iii) top-⟨num⟩*-adv*-adj*-noun: #top3favouritecomics; (iv) presence of plurals at the beginning or end: #10breakuplines, #thingsthatmakeyousmile; (v) presence of superlative adjectives (best, worst, most etc.): #bestmaleathlete, #worstbossin5words.

Search Features

We wanted to investigate how we can use the search engine results to identify social hashtags. To derive these features, we query Google with the segmented hashtag. Some hashtags are quite ambiguous. E.g., "howtoloseaguyin10days" looks like an SL-hashtag but actually is a movie name. One way to identify it, is by querying search engines. Intuitively, if the titles of returned webpages contain phrases with different cardinality of list items, it is an SL-hashtag, else it is not. E.g., all search results for "how to lose a guy in 10 days" refer to exactly 10 days. But a search for "10 valentines day gift ideas" leads to results with 14 ideas, 44 ideas, etc. in the title. Accordingly, we extract the following features.
(o) Number of times the segmented hashtag appears in the top 10 webpages.
(p-q) Number of titles in top 10/20 results which contain the hashtag but with a different number. For hashtags that do not contain the number, this feature is set to the number of titles containing the hashtag.

Tweet Features

We extract the following features from tweets related to the social hashtags.
(r) Time duration for which the hashtag was popular.
(s) Number of contiguous time chunks for which the hashtag was popular.
(t-y) While event related hashtags may co-occur frequently with other hashtags, other co-occurring tags are expected to have relatively low frequency for SL-hashtags. We encode this intuition as a set of features which provide the distribution over tweets (containing the current hashtag) with 0–5 other hashtags.

B. Classification Results

We experimented with Weka [7] implementations of various classifiers: Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Local Deep SVM (LDSVM), Binary Neural Network (BNN), Gradient Boosted Tree (GBT), Averaged Perceptron (AP), and XGBoost Binary Classification (XGBBC). Table I shows 10 fold cross validation results in terms of Precision (P) and Recall (R) for SL hashtags, Overall Accuracy (A), and Area Under Curve (AUC) and respective standard deviations (SD).

TABLE I: SL-hashtag Detection: 10-fold Cross-validation Results and respective Standard Deviations (SD) of Precision (P), Recall (R), Accuracy (A) and Area Under Curve (AUC) for Various Classifiers

Classifier	P, SD(P)	R, SD(R)	A, SD(A)	AUC, SD(AUC)
NB	78.81, 5.73	89.47, 5.13	82.67, 3.75	92.2, 1.51
LR	86.17, 4.29	85.6, 3.32	**86.12**, 2.27	**94.1**, 1.07
SVM	86.05, 8.28	74.7, 1.41	80.76, 4.29	92.2, 1.58
LDSVM	84.47, 5.2	84.3, 2.92	84.33, 2.1	90.9, 1.36
BNN	81.3, 3.85	80.1, 5.04	80.77, 1.75	90.6, 1.17
GBT	85.05, 3.81	85.81, 4.33	85.38, 2.41	93.25, 1.39
AP	83.79, 4.02	83.17, 4.01	83.44, 1.8	92.7, 1.23
XGBBC	85.27, 3.77	84.53, 5.31	84.94, 3.25	92.67, 1.77

The P-R curves for many of these classifiers are shown in Fig 4. From the figure and table, we can see that the Logistic Regression (LR) provides the best accuracy and AUC, ROC among all the classifiers. We tuned various parameters for LR and observed best AUC with regularization weights L1 = 0.5, L2 = 0.1, and initial weight = 0.5. For the proposed task, an ideal classifier would be one with very good recall and reasonable precision. A lower precision would simply amount to indexing of some extra SL-hashtags, which may never be used. Hence, we tune the logistic regression classifier to provide a high recall of 0.953 with a precision of 0.755. We perform further analysis using this recall-optimized LR classifier.

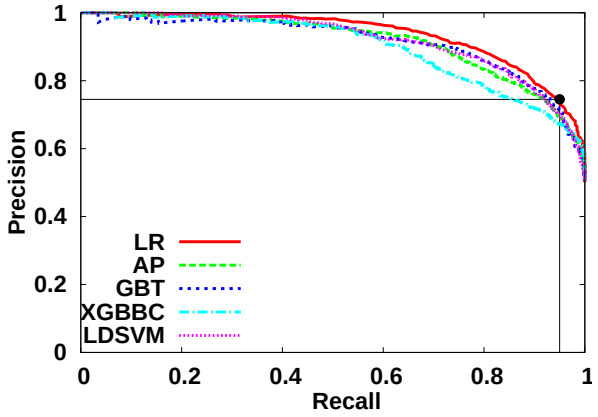

Fig. 4: Precision-Recall Curve for Multiple Classifiers

C. Feature Subset Importance

To understand the importance of various feature subsets, we explore the accuracy obtained using various feature subset combinations. The results for all the combinations of Language features, Search features and Tweet features are shown in

Table II for the Logistic Regression classifier. We fix the recall value to 0.953 for all the combinations. We see that the Language features are the most effective feature subset. However, the best result is obtained when we use all the three feature subsets together.

TABLE II: Precision (P), Recall (R) Accuracy (A) and Area Under Curve (AUC) for different Feature Subsets. Recall fixed at 95.3%. L=Language, S=Search, T=Tweet

Feature Subsets	L	S	T	L+S	T+S	L+T	L+T+S
P(%)	72.4	51.9	49.6	74.2	51.9	73.8	**75.5**
R(%)	95.3	95.3	95.3	95.3	95.3	95.3	**95.3**
A(%)	85.05	55.92	48.03	85.01	55.9	84.97	**86.12**
AUC(%)	93.4	60.7	49.8	93.7	60.2	93.5	**93.7**

D. Feature Importance

The Information Gain (IG) and One Attribute Evaluation (OAE) accuracy results for each feature are shown in Table III. OAE is the accuracy obtained when only the particular attribute (or feature) was used for classification. From Table III it is clear that "(a) hashtag length", "(f) count of nouns", "(j) POS tag entropy", "(m) presence of plurals" are very effective features with high IG and high OAE values. Interestingly, search features (o, p, q) have very low IG values, may be due to the fact that search results for the SL-hashtags are not very consistent. Time-span based features (r, s) also have very low IG values, which might be due to the fact that some idioms other than SL-hashtags also behave similar to SL-hashtags in terms of temporal behavior.

TABLE III: The Information Gain (IG) and One Attribute Evaluation (OAE) for each Feature

Feature	IG	OAE	Feature	IG	OAE
a	**0.325**	**78.1**	n	0.037	57.9
b	0.095	63.6	o	0.007	53.5
c	0.01	52.0	p	0.016	56.2
d	0.027	58.4	q	0.031	57.4
e	0.006	51.0	r	0.024	56.3
f	**0.244**	**74.9**	s	0.041	61.8
g	0.037	59.4	t	0.063	58.1
h	0.023	58.3	u	0.076	60.1
i	0.028	60.0	v	0.075	61.6
j	**0.177**	**70.2**	w	0.077	62.2
k	0.022	53.6	x	0.080	63.6
l	0.011	54.0	y	0.074	65.4
m	**0.173**	**72.7**			

We apply this classifier on the top frequent 0.1M hashtags in our dataset to get 0, 4, 47, 601, 6993 SL-hashtags from the top frequent 10, 10^2, 10^3, 10^4 and 10^5 hashtags respectively.

V. Extraction and Ranking of List Items

After identifying the SL-hashtags, we would like to extract ranked list items for these SL-hashtags.

A. Extraction of List Items

Identifying list items from tweets is quite challenging because there is no unique structure in the way the list items are mentioned. We carefully observed the patterns of list items in the tweets, and identified two types of list items: objective and subjective. Usually objective lists contain multiple list items. Hence, we can first write a regular expression to detect if the tweet contains subjective or objective list items. Further, for objective lists, we can look for typical list item delimiters, and extract individual list items. For subjective lists, unnecessary non-ASCII characters and words could be removed, and the remaining text can be extracted as a list item.

B. Ranking List Items

Several features could be used to rank list items including the following.

- Number of times this list item appears in tweets.
- For the list item, we can collect follower counts of all users who posted tweets containing this item, and use the average follower count as a feature.
- Latest timestamp of posting the list item as a feature.
- Number of co-occurring items across all tweets.

VI. Conclusions

It is generally believed that Twitter is a treasure-trove of opinionated phrases – the main contribution of this paper lies in exploiting that to lay foundation for an efficient search system. We identified that there is a special category of hashtags, called 'SL-hashtags' (e.g. #TipsforInteriorDesign), which can be leveraged to collect relevant social list answers (e.g. "Hang artwork at the right height", "Pick the paint color last", etc.). The indexed answers can then be used to return results for search query. This is a challenging task as a very small percent of hashtags in the Twitter pool are SL-hashtags. We proposed multiple hashtag- and tweet-level features and learned a logistic regression model that provides $\sim$75.5% precision at $\sim$95.3% recall. Further, we briefly discussed ways to extract and rank list items for such list hashtags.

Acknowledgments: We would like to thank Madhu Kumar Dadi and Madhu Sai Ravada for their help with data annotation.

References

[1] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," in *WWW*, 2011.

[2] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter Trending Topic Classification," in *ICDM Workshops*, 2011, pp. 251–258.

[3] S. K. Maity, A. Gupta, P. Goyal, and A. Mukherjee, "A Stratified Learning Approach for Predicting the Popularity of Twitter Idioms," in *ICWSM*, 2015.

[4] K. Rudra, A. Chakraborty, M. Sethi, S. Das, N. Ganguly, and S. Ghosh, "# FewThingsAboutIdioms: Understanding Idioms and Its Users in the Twitter Online Social Network," in *PAKDD*, 2015, pp. 108–121.

[5] G. Berardi, A. Esuli, D. Marcheggiani, and F. Sebastiani, "ISTI TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking," in *TREC*, 2011.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments," in *ACL-HLT*, 2011, pp. 42–47.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Expl.*, vol. 11, no. 1, pp. 10–18, 2009.