TweetSense: Context Recovery for Orphan Tweets by Exploiting Social Signals in Twitter

Manikandan Vijayakumar,
Tejas Mallapura Umamaheshwar
Subbarao Kambhampati
Arizona State University,Tempe, AZ 85281
{manikandan.v,tejas.m.u,rao}@asu.edu

Kartik Talamadupula
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598
krtalamad@us.ibm.com

ABSTRACT

As the popularity of Twitter, and the volume of tweets increased dramatically, hashtags have naturally evolved to become a *de facto* context providing/categorizing mechanism on Twitter. Despite their wide-spread adoption, fueled in part by hashtag recommendation systems, lay users continue to generate tweets without hashtags. When such "orphan" tweets show up in a (browsing) user's time-line, it is hard to make sense of their context. In this paper, we present a system called *TweetSense* which aims to rectify such orphan tweeets by recovering their context in terms of their missing hashtags. *TweetSense* enables this context recovery by using both the content and social network features of the orphan tweet. We characterize the context recovery problem, present the details of *TweetSense* and present a systematic evaluation of its effectiveness over a 7 million tweet corpus.

1. INTRODUCTION

Twitter has grown beyond the role of a platform that is used merely for sharing status updates, as it was initially envisioned. Recent work including that of Java et. al. [5] has identified daily chatter, conversations, information sharing, and news reporting as some of the motivations for users that actively participate in the Twitter network. On average, a user's feed gets a few hundred new tweets every ten minutes It is hard to make sense out of such a feed unassisted, especially when many tweets appear without a *hashtag.*

Hashtags are one of the major features of tweets (and Twitter); they are either a single word or an unspaced phrase prefixed with a pound sign #. The *context* of a tweet can then be described as a set of one or more hashtags. Twitter provides hashtags, partly in an attempt to organize the stream of tweets. However, using hashtags as a method to find the topic of a tweet does not always work, mainly because users do not always tag their tweets with hashtags. As an illustration, in the data that we crawled for our experimentation (all from the year 2014), 76.30% of tweets are orphan tweets. In this paper, we present the TweetSense system that helps in recovering the context of a tweet by tagging the tweet with a suitable hashtag. TweetSense captures the most relevant data from a given user's social graph in order to recover hashtag(s) for a given tweet. The underlying hypothesis is that when the creator of a tweet, called the *originator*, uses a hashtag (to define the context for a tweet), they are likely to reuse one or more hashtags that they see on their own timeline. This includes both tweets posted by the originator, as well as tweets created by the people that the originator follows (friends). Originators are most likely to use hashtags which are temporally close, and are also more likely to reuse hashtags from other users whose tweets they have favorited, retweeted, and @mentioned (a tweet that conatins "@username").

To reflect this generative model, in TweetSense, a statistical model is built to capture a set of social signals, temporal signals related to the tweet and the originator of the tweet. These features measure the tie strength between users, temporal locality and trendiness of the hashtags within the users' social graph. TweetSense learns a model to predict whether a hashtag is applicable to a tweet or not. Given a test tweet lacking a hashtag (context), the model is used to predict hashtags from a set of hashtags collected from the timeline of the creator of the test tweet.

2. RELATED WORK

A problem that is related to the context recovery problem is that of recommending a hashtag for a tweet that the originator is about to post. There has been some previous work on the hashtag recommendation problem. Eva et al. [7] present a recommender system that aims at creating a more homogeneous set of hashtags by considering similarity of tweet text. This candidate recommendation list is later refined using recently used hashtags, popularity of hashtags with in the recommendation list, and popularity of a hashtag within the underlying data set. Jieying She et al. [6] propose a TOpic MOdel-based HAshtag recommendation (TOMOHA) solution. The model learns whether the topic of a tweet is related to a topic which is local to the user or to a global background topic of the corpus. The trained model is used to recommend the most probable hashtags for a tweet. Wei Fang et al. [3] propose a Personalized Hashtag Recommendation system which suggests both content-relevant and user-relevant hashtags when users are composing tweets. The hashtag-relevant features are also used to create hybrid versions of the two systems.

In the hashtag recovery problem, the time taken to predict a hashtag is not as critical as compared to a recommender system. The accuracy of prediction is more important in the problem of context recovery as we are aiding in finding the topic of the tweet rather than suggesting possible topics for the tweet being composed. The temporal information corresponding to the orphan tweet and its creator becomes very important.The problem of recovering a hashtag for tweets on a user's timeline has so far not been addressed.

3. OVERVIEW OF TWEETSENSE

To set up the model for the problem of context recovery, given a tweet Q_x created by a user O_y, we track down the most promising hashtags for it. The candidate set of tweets CT_x is derived based on the generative model of our system by tracing down tweets T_x on the user O_y's timeline. The candidate set of tweets CT_x contains only the tweets that are created before the tweet Q_x was created.

Given a query tweet Q_x, without a context created by an originator O_y appearing on the time-line of a browsing user on Twitter, a set of candidate tweets (containing hashtags) - $\langle CT_{xi}, CH_{xj} \rangle$ extracted from the social circle of the user O_y, and U - the creator of $\langle CT_{xi}, CH_{xj} \rangle$, we want to compute $P(CH_{xj}|Q_x, CT_{xi}, O_y, U)$ - which is the probability that hashtag CH_{xj} of tweet CT_{xi} from the candidate set CT_x is actually the context of Q_x. We estimate the probability discriminatively using a Logistic Regression model. The features are derived from tweets Q_x and CT_{xi}, users O_y and U.

The *tweet-content related* features include similarity between tweet text, hashtag popularity and temporal information of the tweet. The *user related* features include mutual friends, mutual followers, and social signals like @mentions, favorites and common hashtags between the user who created tweet O_y, and the user U who is a part of O_y's social network and created the tweet CT_{xi}. The scoring methods for each feature is described in the following section.

3.1 Tweet-Content Related Features

Similarity Score: is based on the cosine similarity between the text content of the tweet Q_x and the tweets contained in the set of candidate tweets CT_x. We assume that the tweets in CT_x that share the text content with Q_x is more likely to share the hashtag with Q_x. $cos\Theta_{xi} = \frac{\vec{Q}.\vec{CT_{xi}}}{\|\vec{Q}\|\|\vec{CT_{xi}}\|}$

We only consider the tweets in English and ignore query tweets in other languages, special characters, emoticons, URLs, and HTTP links. We also remove stop words.

Recency Score: Hashtags that are temporally close to the query tweet get a higher ranking. We determine the time window for the tweet, hashtag pair, $\langle CT_{xi}, CH_{xj} \rangle$, using the "created at" timestamp, $CR(CT_{xi})$, associated with the tweet CT_{xi}. We adapt the exponential decay function to compute the recency score of a hashtag. We use the expression $e^{-\frac{CR(Q_x)-CR(CT_{xi})}{t}}$, where $t = 60 \times 10^3$, to compute the recency score. By varying the sensitivity of the time window from 1 minute to 170 hours, we found that the results are more promising when the time window is set to 17 hours. This corresponds to a value of t equal to 60×10^3.

Social Trend Score: corresponds to the popularity of hashtags within the candidate hashtag set, CH_x. As the candidate hashtag set CH_x is derived from the timeline of the user U who posted the tweet Q_x, it is intuitive that a hashtag with high frequency is popular in the user's social network. The social trend score is computed based on the "One person, One vote" approach. It is used to get the count of frequently used hashtags in CH_x.

3.2 User Related Features

Attention Score: If a particular user was @mentioned recently, it is more likely that they share topics of interest. This also means that they might use similar hashtags. We compute a user's attention score by a weighted average sum on the conversations between two users. Let $AT(T_{O_y})$ be the set of all tweets of user O_y and $AT(T_U)$ be the set of all tweets of user U, let $AT(T_{O_y,U})$ be the set of all tweets which has @mentions and replies of O_y with U and let $AT(T_{U,O_y})$ be the set of all tweets which has @mentions and replies of U with O_y, where U is a user who belongs to the list of friends of O_y. We compute the weighted average of @mention and replies between the users as: $a_{i,j} = \frac{|AT(T_{O_y,U})|}{|AT(T_{O_y})|}$
$a_{j,i} = \frac{|AT(T_{O_y,U})|}{|AT(T_U)|}$

Final Score $= (\alpha)\ a_{i,j} + (1-\alpha)\ a_{j,i}\ where\ \alpha = 0.5$ We have set $\alpha = 0.5$.

Favorite Score: When a user favorites a tweet posted by his friend, the user is consciously letting his friend know that he shares interest with the friend on that specific topic. Higher the number of times a user favorites a tweet of another user, higher is the favorite score. Let $FV(T_{O_y,U})$ be the set of all tweets which has favorites of O_y with U and let $FV(T_{U,O_y})$ be the set of all tweets which has favorites of U with O_y, where U is a user who belongs to the set of friends of O_y. Favorite score can be computed by using the expression: $a_{i,j} = \frac{|FV(T_{O_y,U})|}{|FV(T_{O_y})|}$ $a_{j,i} = \frac{|FV(T_{O_y,U})|}{|FV(T_U)|}$
Final Score $= (\alpha)\ a_{i,j} + (1-\alpha)\ a_{j,i}\ where\ \alpha = 0.5$ We have set $\alpha = 0.5$.

Mutual Friends Score: Mutual friends score is computed to rank the friends based on their number of common friends that they share in their social network. If F_{O_y} contains set of users that are friends with user O_y and F_U contains set of users that are friends with user U. We use the same Jaccard's coefficient [4] on the two set as the measure of the "mutual friends" feature.

Mutual Followers Score: Mutual followers score is computed to rank friends based on the number of followers they share in their network. If FW_{O_y} contains set of users following user O_y and FW_U contains set of users following user U. We use the same Jaccard's coefficient [4] on the two set as the measure of the "mutual followers" feature.

Common Hashtags Score: Common hashtags score is computed between any two users based on the hashtags that are shared between them. If two users O_y and U use the same set of hashtags for a particular time window, then both the users are talking about the same topic. To compute this, we first collect the unique set of hashtags used by each user, and then use Jaccard's coefficient [4] on the hashtag sets H_{O_y} and H_U.

Reciprocal Score: The user might follow his friend but also follow a topic of his interest such as a news channel or a celebrity. To give more importance to a user's friends over others, the reciprocal rank assigns fixed values to classify the user's followers as a "friend", or as "not a friend". The users who follow each other will receive a fixed score of 1.0, and 0.5 other wise.

3.3 Statistical Model

The problem is modeled as shown in Figure 1. We build a Logistic Regression model based on the feature matrix extracted based on the tweets corresponding to the set of training users.

Training dataset: The training data set is constructed by considering many training tweets Q. The corresponding set of candidate tweet and hashtag pairs $\langle CT_x, CH_x \rangle$ is

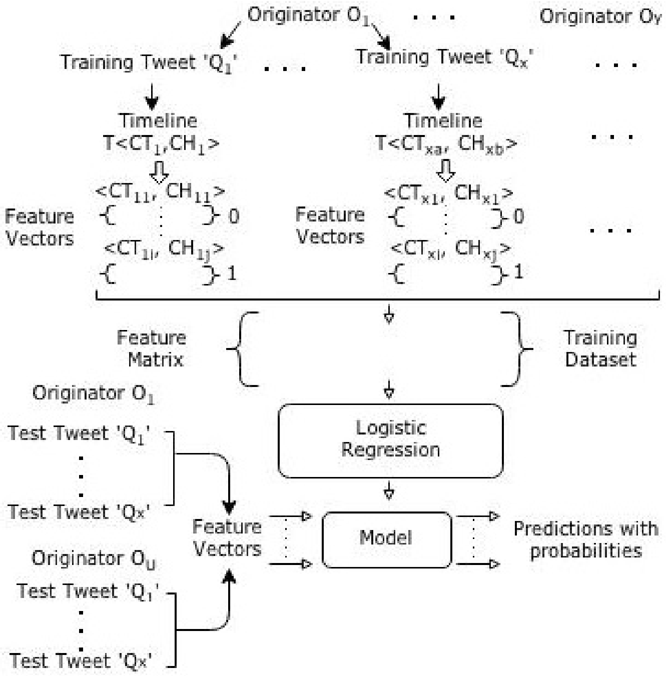

Figure 1: Training the Model from Tweets With Hashtags to Predict the Hashtags for Tweets Without Hashtag

identified. Here, the candidate set of tweets are the tweets from the timeline of the user O_y who posted the tweet Q_x containing the hashtag CH_x . For each candidate tweet, and candidate hashtag pair $\langle CT_{xi}, CH_{xj}\rangle$ created by user U in the candidate tweet set, the feature scores are computed with respect to the Q_x, and user O_y.

The training dataset is a feature matrix containing the feature vectors of all $\langle CT_{xi}, CH_{xj}\rangle$ pair corresponding to all training tweets Q. The class label for a feature vector is 1 if the hashtag CH_{xj} in the candidate set of tweets is equal to the hashtag in Q_x, the tweet at consideration, and 0 otherwise.

Handling unbalanced training set: The training dataset has a class distribution of 95% negative samples and 5% positive samples. Learning the model from an unbalanced dataset will cause very low precision. We use the Synthetic Minority Oversampling Technique (SMOTE) [2] to re-sample the unbalanced dataset to a balanced dataset with 50% positive samples and 50% negative samples.

Classifier learning: We apply the Logistic regression to learn a statistical model from the training dataset to predict the probabilities of the top K most promising hashtags for a given test tweet. Logistic regression assumes that all data points share the same parameter vector with the test tweet.

Using the Classifier: For each test tweet, its candidate set of tweet-hashtag pairs are tracked down and feature vectors are computed. When the test dataset is passed to the learned model, it predicts the maximum likelihood probability for each of the candidate hashtags CH_{xj} in tweet hashtag pairs $\langle CT_{xi}, CH_{xj}\rangle$ corresponding to the test tweet. The candidate hashtags with predicated class label as 1 are then ranked using the probabilities.

4. EXPERIMENTAL SETUP

Dataset: We use Sprintze [1] to crawl Twitter data through the Twitter Streaming API. In order to crawl a user's timeline, the method can only return up to 3,200 of a user's most recent tweets from his timeline. The favorite tweets that can be crawled are limited to 200 most recent tweets per user.

Characteristics	Value	Percentage
Total number of users	8,949	N/A
Total number of originator users	63	N/A
Total Tweets Crawled	7,212,855	100%
Tweets with Hashtags	1,883,086	23.70%
Tweets without Hashtags	6,062,167	76.30%
Tweets with exactly one Hashtag	1,322,237	16.64%
Tweets with more than one Hashtag	560,849	7.06%
Tweets with Favorites	716,738	9.02%
Tweets with @mentions	4,658,659	58.63%

Table 1: Characteristics of the dataset used for the experiments

We randomly picked users by navigating through the trending hashtags during a fixed time interval. For each of the selected users, we crawled the most recent 1500 tweets, and further crawled recent 1500 tweets for each friend (followee) of the selected user. Since, the number of tweets crawled to build a user's social graph is directly proportional to the number of friends, we randomly constrained the user selection process to choose users with at most 300 friends. We crawled 7,212,855 tweets for 8,949 users. Further details about the characteristics of the dataset can be found in Table2.

5. EMPIRICAL EVALUATION

We present an internal and external evaluation of TweetSense. The testing dataset comprised of tweets that had exactly one hashtag associated with it. The hashtag was removed for the purpose of testing, and this served as the ground truth for the test tweet.

5.1 External Evaluation

The closest related work for the problem of context recovery is the problem of recommending hashtags. Therefore, we choose the system proposed by Eva et al. [7] as our baseline. Their system aims at creating a more homogeneous set of hashtags by considering the similarity of tweet text to create a candidate recommendation list. This candidate recommendation list is later refined using recently used hashtags, and popularity of hashtags with in the candidate recommendation list.

External Evaluation Of TweetSense Based On Precision at N:

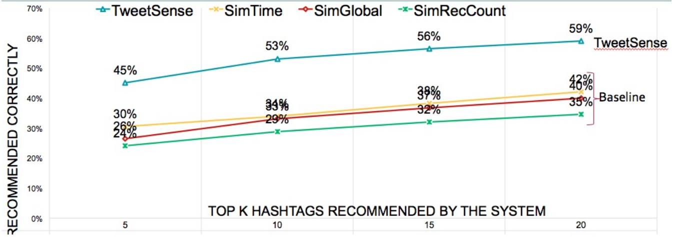

Figure 2: External evaluation againt state-of-the-art system for Precison @ N

Our system was able to recommend correct hashtags for precision at 20 for 59% of the tweets, which in general is above 50%. Also compared to the best possible ranking method of the baseline model which could recommend correct hashtags for 35% of the tweets.On an average, TweetSense dominates the baselines for different values of N.

A user tweets about his interests and also about what he is exposed to on his timeline. A user would rarely use a hashtag, which he has never seen. There are many indicators that indicate how a user adapts hashtags and most of these are related to user's social network. TweetSense picks the most suitable set of hashtags as candidate hashtag set by looking at the user's timeline rather than at global Twitter ecosystem. We have identified different features that can further help in determining the most important indicator of all the indicator by assuming the user's environment at the time of the creation of a tweet. These indicators change with time, and we also model this by considering different set of candidate hashtags for the same user for different tweets.

5.2 Internal Evaluation

The correctness of the system is evaluated using Precision at N. We compare the importance of different features in the model by using odds ratio.

Results of Internal Evaluation Of Precision at N by Varying the Training Dataset: We compare the precision at N at 5,10,15, and 20 of the proposed system. Our approach gets better precision as the size of N is increased. For a total sample size of 1599 random tweets with hashtags whose hashtags are deliberately removed for evaluation. At the value of $N = 5$, 720/1599 sample tweets are recommended with the correct hashtags. Similarly, 849/1599 at $N = 10$, 901/1599 at $N = 15$ and 944/1599 at $N = 20$ are predicted correctly.

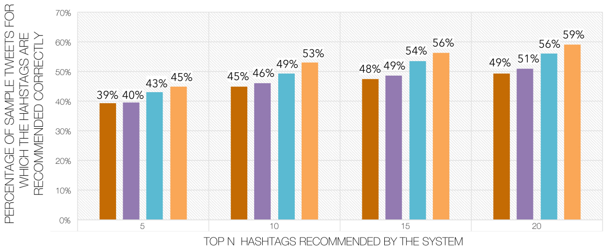

Figure 3: Precision at N = 5, 10, 15, and 20 on Varying the Size of the Training Dataset.

Results for Estimation of Odds Ratio by Feature Selection: We measure the association between an exposure and an outcome using odds ratio. In the Table 2, *Exp1* column indicates that "Mutual Friends" feature is contributing the most to the odds of the outcome when compared to the other features. This reinforces the hypothesis that the social signals are more important than the tweet-content related features while predicting hashtags.

In order to validate whether the prediction capability of the model is based solely on a single feature, we created a model by ignoring the "Mutual Friends" feature during the training phase. In this case - *Exp2*, as shown in the Table 2, we can see that the "Mutual Followers" feature becomes very important. There could be a correlation between the two features because of feature redundancy. In *Exp3*, we remove most of the social features - "Mutual Friends", "Mutual Followers", and "Reciprocal" features to build a model. We can observe that the odds ratio of the features being considered do not improve significantly. In *Exp4*, we ignore all the features, but the "Mutual Friends" feature to build a model to further verify the importance of the "Mutual Friends" feature. As we could see, the score is low in this case while it is higher when the model was built with all other features. This indicates that we require all the other features along with the "Mutual Friends" feature to make better predictions.

All Features	Exp1	Exp2	Exp3	Exp4
Similarity Score	0.0942	0.1123	0.1134	N/A
Recency Score	0.0022	0.0024	0.0026	N/A
Social Trend Score	0.0017	0.0017	0.0016	N/A
Attention Score	0	0	0	N/A
Favorite Score	0.2837	0.24	0.2112	N/A
Mutual Friends Score	13538.65	N/A	N/A	0.2081
Mutual Followers Score	0.0923	3.115	N/A	N/A
Common Hashtag Score	0	0	0	N/A
Reciprocal Score	0.7144	0.7717	N/A	N/A

Table 2: Estimation of Odds Ratio by Feature Selection

All these experiments emphasize the fact that social features rather than the tweet-content related features are the most important features in recovering context of an orphan tweet.

Results on Accuracy of Ranking based on Rank Position: The accuracy on hashtags recommended by the system is shown by determining the ranking positions of the top 10 recommended hashtags for the test dataset.Results for each ranking position as follows: Rank1-27.75%,Rank2-21.53%,Rank3-13.56%,Rank4-12.28%, Rank5-6.70%,Rank6-5.10%, Rank7-4.31%,Rank8-2.07%, Rank9-3.51%, Rank10-3.19%.The consistent performance by the system for the top four positions of the top K ranking positions imply that the system is more accurate.

6. CONCLUSION

In this paper, we defined and motivated the context recovery problem from orphan tweets. We then described *TweetSense* a discriminative learning approach for recovering the context of the orphan tweets in terms of their missing hashtags. *TweetSense* uses a variety of features drawn from the timeline, content and social network. Our experiments on a large tweet corpus demonstrate the effectiveness of *TweetSense.*

Acknowledgments: We gratefully acknowledge the significant help from Sushovan De in this research. This research is supported in part by the ARO grant W911NF-13-1-0023, and the ONR grants N00014-13- 1-0176, N00014-13-1-0519 and N00014-15-1-2027, and a Google faculty research award.

References

[1] Twitter"s streaming api ,https://blog.gnip.com/tag/spritzer/.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 2011.

[3] W. Feng and J. Wang. We can learn your hashtags: Connecting tweets to explicit topics. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 856–867, March 2014.

[4] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[6] J. She and L. Chen. Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 371–372, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[7] E. Zangerle, W. Gassler, and G. Specht. On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Eva2013*, 3(4):889–898, 2013.