# 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.

## Categories and Subject Descriptors

I.5 [**PATTERN RECOGNITION**]: Applications; H.3.3 [**Information Search and Retrieval**]

## Keywords

Twitter, Regression Model, Rectification, Hashtags, Social Network, Context

## 1. INTRODUCTION

Twitter has grown beyond the role of a platform that is used merely for sharing status updates, as it was initially envisioned. On an 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 have organically evolved to become a context providing feature for the tweets. The *context* of a tweet can then be described as a set of one or more hashtags. 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 shown in Table 1. In this paper, we present the TweetSense system that helps in recovering the context of a tweet. The underlying hypothesis is that when the creator of a tweet, called the *originator*, uses a hashtag, they are likely to reuse one or more hashtags that they see on their own timeline.

To reflect this generative model, in *TweetSense*, a statistical model is built to capture a set of social signals, temporal and tweet content related features of the tweet and the originator of the tweet. *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 $k$ most promising hashtags.

**Related Work:** A problem that is superficially similar to the context recovery problem is the hashtag recommendation problem, which involves suggesting possible hashtags to the originating user at the time of tweet creation (c.f. [1, 2, 3]. In contrast the recovery problem involves figuring out the missing hashtag *after* the tweet appears in the timeline of a browsing user. As such, the recovery problems poses more stringent demands on the prediction accuracy (while loosening the demands on running time).

## 2. OVERVIEW OF TWEETSENSE

The orphan tweet context recovery problem addressed by *TweetSense* can be stated as follows: Given an orphan tweet $Q_x$ found in the browsing users timeline, where $Q_x$ is created by a user $O_y$, *TweetSense* aims to recover the missing hashtag for $Q_x$. It does this by starting with a candidate set of tweets from the time-line and social circles of the originating and browsing users, and computing the probability that the hastag of a candidate tweet is in fact the missing hashtag of $Q_x$. *TweetSense* estimates this probability discriminatively using a logistic regression framework that uses a variety of features drawn from the timeline and social circles of the originating and browsing users. Figure 1 provides an overview of *TweetSense* approach.

**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 identified, 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, the feature scores are computed with respect to the $Q_x$, and user $O_y$. 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 $Q_x$, and 0 otherwise.

*WebSci ’15 June 28 - July 01, 2015, Oxford, United Kingdom*
© 2015 ACM. ISBN 978-1-4503-3672-7/15/06...$15.00
DOI: http://dx.doi.org/10.1145/2786451.2790157

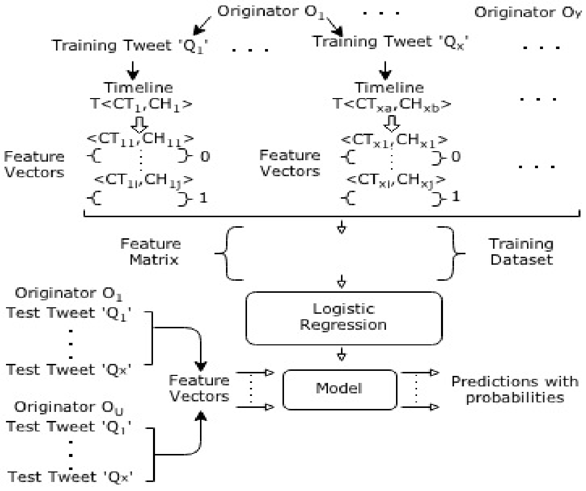

**Figure 1:** ***Training the Model from Tweets With Hashtags to Predict the Hashtags for Tweets Without Hashtag***

| 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**

**Handling unbalanced training set:** The training dataset has a class distribution of 95% negative samples and 5% positive samples. We use the Synthetic Minority Oversampling Technique (SMOTE) [**?** ] to re-sample the unbalanced dataset to a balanced dataset with 50% positive samples and 50% negative samples to achieve better precision.
**(Discriminative) Model 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.
**Using the Learned Model:** When the test dataset is passed to the learned model, it predicts the probability for each of the candidate hashtags $CH_{xj}$ in tweet hashtag pairs corresponding to the test tweet. The candidate hashtags with predicated class label as 1 are then ranked using the probabilities.

## 2.1 Features used in Model Learning

**Similarity Score:** is the cosine similarity between the text content of the tweet $Q_x$ and each tweet in the candidate set of tweets $CT_x$. We only consider the tweets in English and ignore query tweets in other languages, special characters, emoticons, URLs, and 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. We adapt the exponential decay function to compute the recency score of a hashtag. We use the expression $e^{-\frac{CR(Q_x)-CR(CT_{xi})}{60\times10^3}}$, for computation.
**Social Trend Score:** is the normalized frequency of hashtags within the candidate set.
**Attention and Favorite Score:** If a particular user was @mentioned recently, it is more likely that they share topics of interest and When he/she favorites a tweet, the user is consciously letting his friend know that he shares interest that specific topic.
**Mutual Friends, Mutual Followers and Common Hashtags Score:** are defined as the Jaccard's coefficient [**?** ] on the sets of friends, followers, and hashtags of any two users.
**Reciprocal Score:** The users who follow each other will receive a fixed score of 1.0, and 0.5 other wise.

# 3. EMPIRICAL EVALUATION

We present an internal and external evaluation of TweetSense. The testing dataset comprised of tweets that had exactly one hashtag which is used as a ground truth for the test tweet. Characteristics of the dataset is shown in the Table 1.

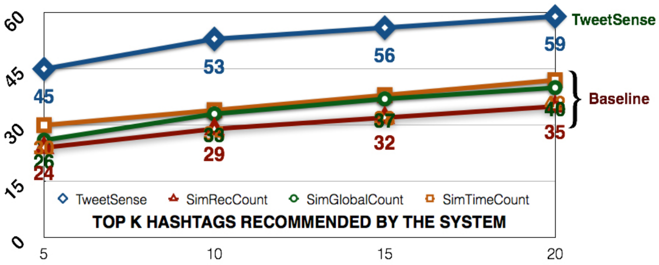

**Figure 2:** ***External evaluation againt state-of-the-art system for Precison @ $N$***

**External Evaluation Of TweetSense Based On Precision at** $N$**:** 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. [3] as our baseline. At precision at 20, our system was able to recommend 59% of correct hashtags over only 35% by baseline and dominates for all values on $N$.
**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, 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.

# 4. 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-

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

**Table 2: Estimation of Odds Ratio by Feature Selection**

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] 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.

[2] 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.

[3] 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.