Improved, Scalable, and Personalized Context Recovery System: E-TweetSense

by

Tejas Mallapura Umamaheshwar

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved June 2015 by the
Graduate Supervisory Committee:

Subbarao Kambhampati, Chair
Huan Liu
Hasan Davulcu

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Browsing Twitter users, or *browsers*, often [5] find it increasingly cumbersome to attach meaning to tweets that are displayed on their timeline as they follow more and more users or pages. The tweets being browsed are created by Twitter users called *originators*, and are of some significance to the browser who has chosen to subscribe to the tweets from the originator by following the originator. Although, *hashtags* are used to tag tweets in an effort to attach context to the tweets, many tweets do not have a hashtag. Such tweets are called *orphan tweets* and they adversely affect the experience of a browser.

A *hashtag* [6] is a type of label or meta-data tag used in social networks and microblogging services which makes it easier for users to find messages with a specific theme or content. The *context* of a tweet can be defined as a set of one or more hashtags. Users often do not use hashtags to tag their tweets. This leads to the problem of missing context for tweets. To address the problem of missing hashtags, a statistical method was proposed [21] which predicts most likely hashtags based on the social circle of an originator.

In this thesis, we propose to improve on the existing context recovery system by selectively limiting the candidate set of hashtags to be derived from the intimate circle of the originator rather than from every user in the social network of the originator. This helps in reducing the computation, increasing speed of prediction, scaling the system to originators with large social networks while still preserving most of the accuracy of the predictions. We also propose to not only derive the candidate hashtags from the social network of the originator but also derive the candidate hashtags based on the content of the tweet. We further propose to learn personalized statistical models according to the adoption patterns of different originators. This helps in not only identifying the personalized candidate set of hashtags based on the social circle and content of the tweets but also in customizing the hashtag adoption pattern to the originator of the tweet.

ACKNOWLEDGEMENTS

It is my pleasure to express gratitude to my adviser, Dr. Subbarao Kambhampati for guiding me with his expertise. Dr. Rao is the perfect guide who made sure that I was always on the right course whilst gave me the intellectual freedom to explore ideas on my own. In this way, he not only helped me become a better researcher, but also instilled a confidence in me that will surely hold me in good stead in the future.

I would like to thank my committee members Dr. Huan Liu and Dr.Hasan Davulcu for their time and support towards my research work.

I would like to thank Dr. Kartik Talamadupula for making time from his schedule during his defense process to guide me. I would like to thank Manikandan Vijayakumar for his technical insights and advice that helped me in progressing quickly.

I thank Yochan team and my friends for their helpful comments on all the technical documentations and presentations. I take this opportunity to express my gratitude towards all of the department members for their help and support.

I thank my family for their continual encouragement and support. I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

Twitter allows registered users to share and read short text messages. Twitter users started tagging their tweets with hashtags to bring order to the abounding tweet messages and thus organize tweets. The # symbol, sometimes called a hash character, is used to mark keywords or topics in a tweet called a *hashtag*. Hashtags were created by Twitter users as a way to categorize messages [18]. It has been pointed out that the users engage twice as much whenever a tweet is tagged with a hashtag [12].

Twitter has 302 million monthly active users and over 500 million tweets sent per day [20]. The number of tweets that have hashtags is low. In our dataset, the percentage of tweets that have hashtags is less than 24%. Therefore, Twitter browsers often find it difficult to understand the topic of a tweet without any hashtags that appear on their time-line. Such tweets without hashtags are defined as *orphan* tweets.

In the recent past, Twitter launched a new feature to introduce random tweets into browsers' tweet feed based on an external algorithm [14] making the timeline of the browser more cluttered. This has made it more difficult for the Twitter users to derive meaning out of their time-line [3].

1.2 Problem Statement

Many users find it difficult to attach meaning to their tweet feed appearing on their time-line. Figure 1.1 [21] contrasts between orphan and non-orphan tweets. It is difficult to find the context for every orphan tweet that can appear on a browser's timeline. If a

browser has to react to a tweet, the browser needs to find the context of the tweet. Finding the context becomes difficult as the average length of a single post is about 14 words or 78 characters [9]. The number of tweets per day, the diversity of the topics of discussion, the growing social network of the browser and the originator make it even more difficult to decipher the context of a tweet.

The problem of recovering context of a tweet is different than the problem of recommending hashtags. Here, the time taken to recover hashtags given a tweet is not as crucial. However, the accuracy of recommendations is much more important.

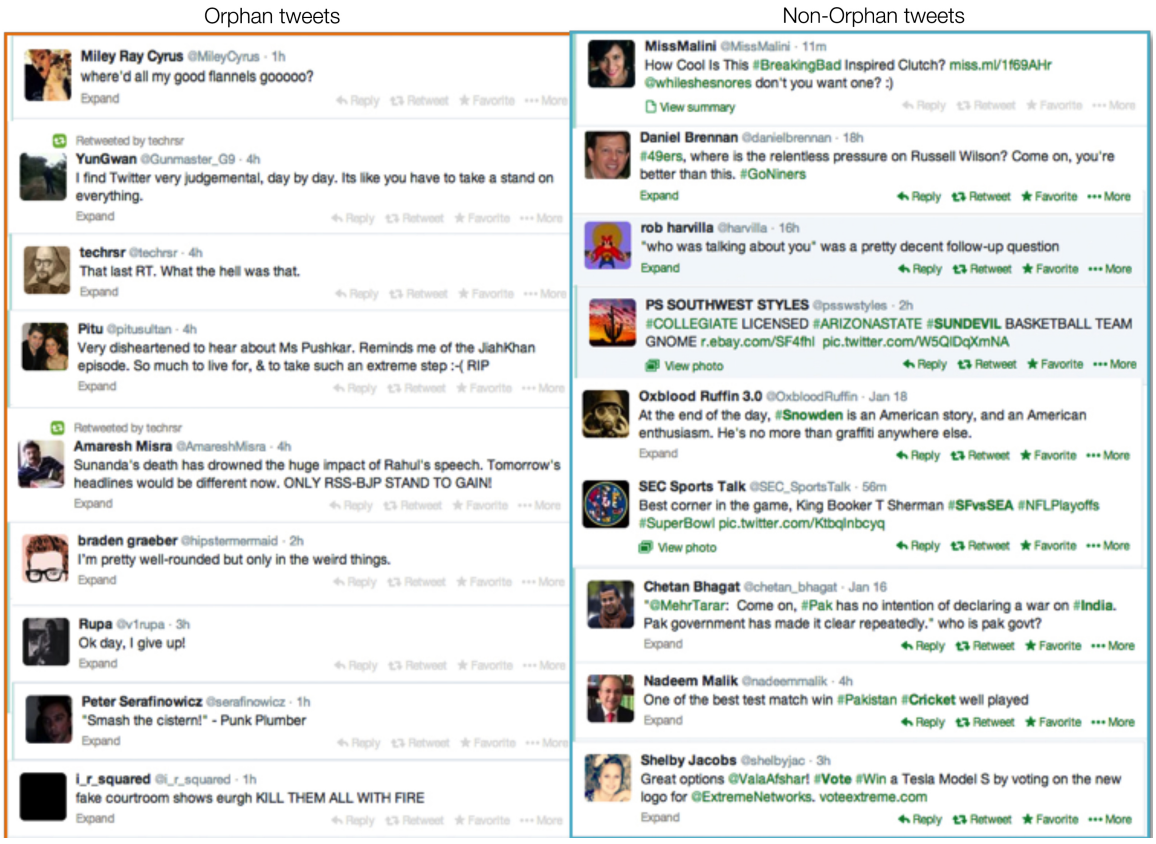

Figure 1.1: Example of Orphan and Non-orphan Tweets

1.3 Proposed Approach

Creating a tweet with a set of hashtags that attach context to it can be done in three ways.

- Originators can adopt a hashtag from their social circle.
- Originators can create tweets related to a set of one or more hashtags that may not be popular in their social circle but popular outside their respective social circles.
- Originator can create a tweet and attach new hashtags to it.

A Twitter originator can use any combination of the above model to create a tweet with a context. The existing system called TweetSense proposed by Manikandan Vijayakumar [21], accommodates the first way of tagging tweets. We are proposing to extend the existing system to incorporate the second way of tagging tweets with hashtags. To further aid this model, we have added few more features to improve predictions for the second way of attaching hashtag to tweets. Addressing the third way of tagging tweets is out of the scope of this thesis.

TweetSense system helps in recovering the context of a tweet by finding a missing hashtag for a tweet. TweetSense captures the most relevant data from a given user's social graph in order to recover hashtag(s) for a given tweet. The system however, cannot be used to recover hashtags for originators with large social circles comprising of more than 300 users as the time taken to process increases greatly with the size of the social circle making the system unusable for originators with larger social circles. We propose to remove this limitation by carefully filtering users from the social circle which also helps in reducing the computation time.

We further propose individual statistical models to learn pattern adoption by different users rather than a single model for all Twitter users. We verify the effectiveness of the

proposed system by evaluating the proposed system internally as well as externally against TweetSense.

1.4 Organization of Thesis

In the next chapter, we present the related work. In chapter 3, we present the reconstruction of the state-of-the-art system called TweetSense, and discuss the limitations of the system. In chapter 4, we present the details of the improved system. Chapter 5 and 6 explain the experimental setup and evaluation of the system, followed by the conclusion and scope for future work in chapter 7.

Chapter 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. [22] 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 within the recommendation list, and popularity of a hashtag within the underlying data set. Jieying She et al. [13] 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. [8] 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. In this case, 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.

In contrast to other existing systems, the proposed system and its predecessor TweetSense, learn the relationship between tweets and their originators that share a hashtag as against the relationship between a tweet and the corresponding hashtag and/or its originators. This allows us to reuse the same statistical model across various tweets as the model is independent of the candidate hashtags.

Chapter 3

TWEETSENSE - BACKGROUND

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, TweetSense tracks a set of candidate tweets (containing hashtags) - $\langle CT_{xi}, CH_{xj} \rangle$ extracted from the social circle of the originator O_y. If U is 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 for prediction are derived from the tweet Q_x and CT_{xi}, users O_y and U as shown in the Figure 3.1. TweetSense uses the following set of *tweet-content related* features and *user related* features:

3.1 Tweet-Content Related Features

Similarity Score: Cosine similarity between the text content of the tweet Q_x and the tweets contained in the set of candidate tweets CT_x is used as a measure to find hashtags that share similar tweet text.

Recency Score: Hashtags that are temporally close to the tweet Q_x get a higher ranking as it is more likely that a user would talk about recent affairs. The exponential decay function is used to compute the recency score of a hashtag: $e^{-\frac{CR(Q_x)-CR(CT_{xi})}{t}}$, where $t = 60 \times 10^3$ [21], to compute the recency score.

Trend Score: The trend score corresponds to the popularity of hashtags within the candidate hashtag set derived based on the originator's social circle.

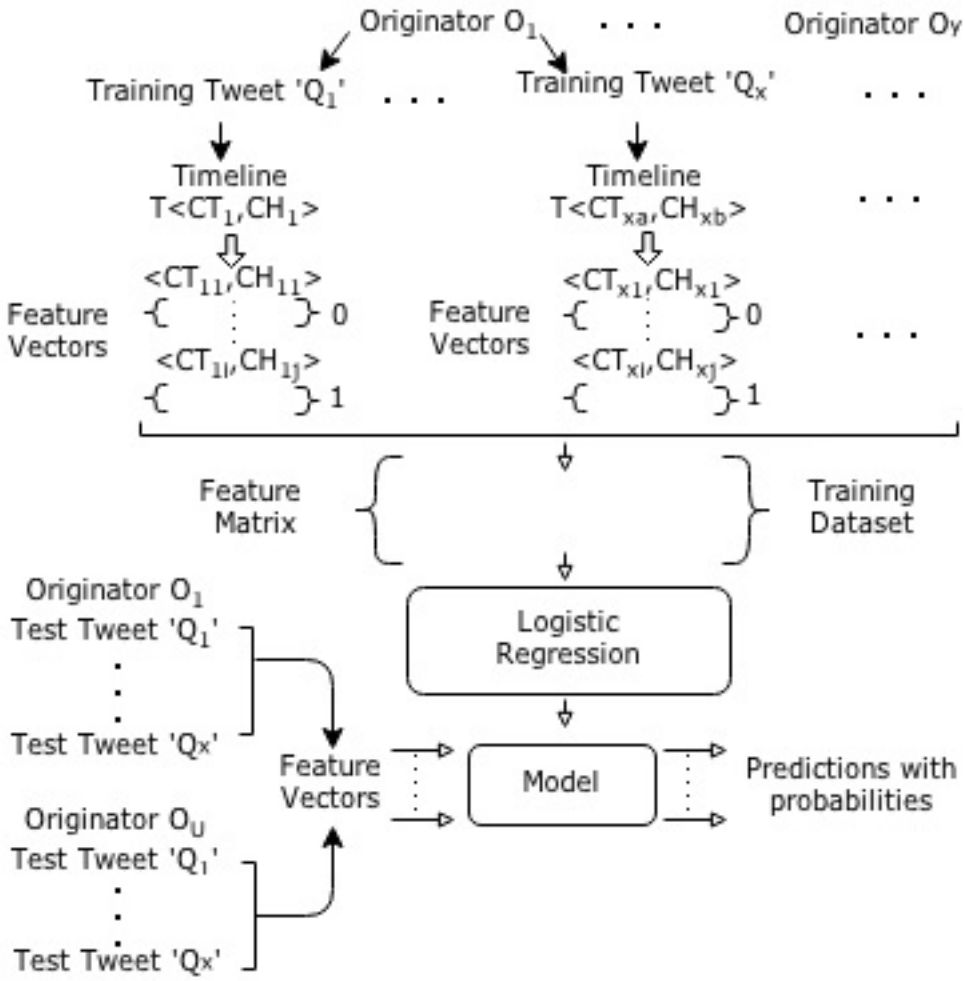

Figure 3.1: Flow Diagram for TweetSense

3.2 User Related Features

Attention Score: Attention from the user O_y to the user U is defined as the set of all tweets which has @mentions and replies [19] from O_y to U. Attention score between two users is computed as the weighted average of Attention from O_y to the user U and Attention from U to the user O_y.

Favorite Score: is the Jaccard's similarity [11] on the set of all favorite tweets of users O_y and U.

Mutual Friends Score: is computed as the Jaccard's coefficient on the set of all following/friends [16] of users O_y and U.

Mutual Followers Score: is computed as the Jaccard's coefficient on the set of all followers [16] of users O_y and U.

Common Hashtags Score: is computed as the Jaccard's coefficient on the set of all previously used hashtags of users O_y and U.

Reciprocal Score: The users who follow each other will receive a fixed score of 1.0, and 0.5 other wise.

3.3 Statistical Model

Training dataset: As shown in Figure 3.1, the training data set is constructed by considering many training tweets, Q, that belong to different users O_y. The corresponding set of candidate tweet and hashtag pairs $\langle CT_x, CH_x \rangle$ is 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 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 each training tweet Q_x. 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 under consideration, and 0 otherwise.

Classifier Learning: A 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: When the test dataset is passed to the logistic regression model, the model predicts the maximum likelihood probability for each entry of candidate hash-

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 3.1: Estimation of Odds Ratio by Feature Selection

tags CH_{xj} in tweet hashtag pairs $\langle CT_{xi}, CH_{xj} \rangle$ corresponding to the tweet Q_x. The candidate hashtags with predicated class label as 1 are then ranked using the probability of the prediction.

3.4 Results of TweetSense

TweetSense was able to recommend correct hashtags for 59% of the tweets. TweetSense also indicates that the "Mutual Friends" feature is the most important feature amongst all the features considered. This means that the predictions depend on the user in the social network to whom the candidate hashtag corresponds to.

3.5 Limitations of TweetSense

Content Based Candidate Set: Users on Twitter not only adopt the hashtags from their social circle but also adopt it from outside their social circle due to external influences or even create a new hashtag as described in the generative model. TweetSense addresses only the first aspect and derives the candidate hashtags based on the content of the tweet.

Odds ratio of Features: TweetSense has evaluated the relative importance of its features by using a odds ratio as shown in Table 5.1. TweetSense also claims that content of the tweet is not at all important according to Exp3 where a model is built without the two most important social features. Since all the candidate set of tweets are derived from the social network of the originator, and there is only one content based feature, the results are justified but may be biased towards user related features.

Note: A value of N/A in Table 5.1 shows that the feature was not considered while building the model

Scaling to Originators with Large Social Circle: TweetSense can only address users who have a small social circle of friends. This is a constraint imposed by TweetSense on the kind of originators for which it can predict hashtags, that is, the social circle or a originator should not be greater than 300 users.

Personalizing Hashtag Adoption Patterns: TweetSense learns the adoption pattern based on a randomly picked set of 18 users called the training set of users. The training instances from all these users may not be representative of all users on Twitter or their hashtag adoption patterns.

Recommending More Than One Hashtag: TweetSense outputs a list of hashtags as the most suitable context for a tweet. It does not predict multiple hashtags for the same tweet.

This thesis attempts to address the above limitations of TweetSense and to improve the efficiency of the context recovery system.

Chapter 4

IMPROVEMENTS TO TWEETSENSE

A Twitter user can be influenced by multiple sources to post a tweet. The assumption made in TweetSense is that the originator is mostly influenced by his social circle on Twitter, and rarely by any external sources [21]. However, in this thesis, the effect of external influence is approximated to the influence caused by the Twitter users who are not in the originator's social circle. To account for this external influence, we propose to use the content of the tweet to derive a set of tweets from the Twitter corpus.

We have described the generative model for creating a tweet with hashtag(s) in section 1 subsection 1.3 . The source of influence could be the social circle, therefore we gather candidate hashtags from the social circle of the originator based on the strength of ties with users in the social circle. We derive candidate set of hashtags from the tweets collection (which in our case is approximately 8 million tweets) to incorporate hashtag recommendations based on the originator's new interests which is not acquired from the social circle on Twitter. We do not address the issue of creating new hashtags as this is out of the scope of this thesis.

4.1 Scaling the System for Users with Large Social Circles

According to the table 5.1, TweetSense demonstrates that "mutual friend rank" is the most important feature in predicting hashtags based on the social circle of the user. This indicates that the originator shares interests with people in her network and tends to adopt hashtags from users in the network. We have further exploited the fact that Twitter users have many declared set of friends whilst their actual set of friends is a much smaller number [10]. We use "mutual friend" score to rank the users in the social network of the

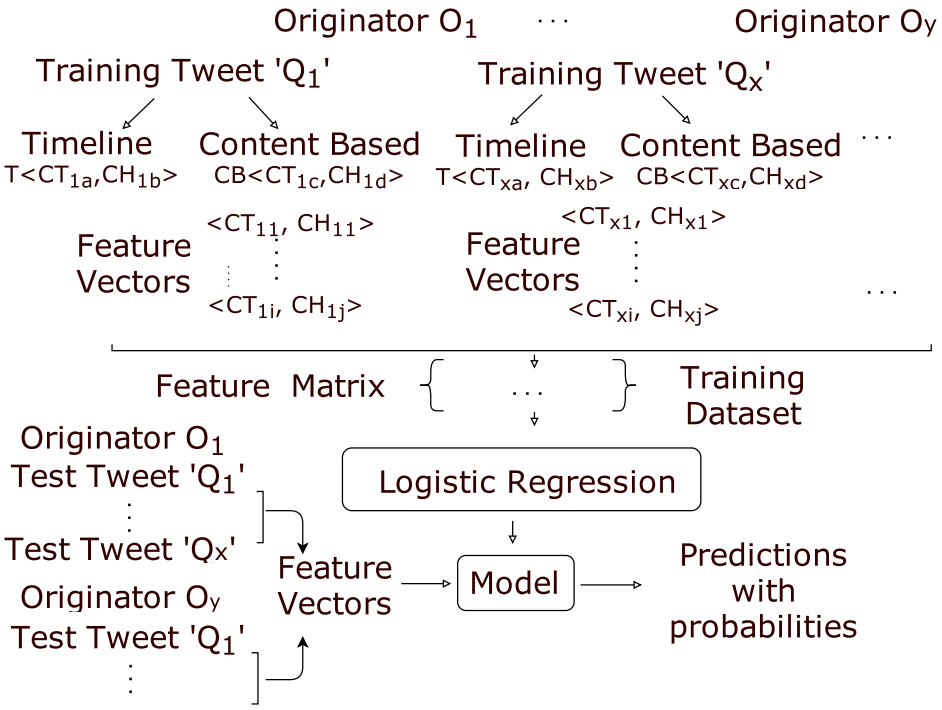

Figure 4.1: Improved TweetSense - Candidate Set Derived Based on the Content and the Social Circle of the Originator

originator and limit the collection of candidate hashtags from users that actually matter, that is, to users who share more common friends with the originator. By experimentation, we have found that limiting this number to 50 preserves the accuracy of hashtag prediction for most users (results are presented in 6.1). The other feature identified by TweetSense to be of utmost importance in predicting hashtags is the “mutual follower" score. We also rank users in the social network of the originator according to the “mutual follower” score and extract tweets from top 20 users. TweetSense collects up-to 1500 tweets per user U in the social circle of the originator as the candidate set of hashtags from that user U. In the new system, we are limiting this number to 100 tweets per user U. These 100 tweets are the most recent tweets corresponding to the user U which were created before the tweet Q_x. This helps in choosing the most influential temporal tweets.

Limiting the number of candidate set of hashtags and limiting the users from whom

All Features	Odds Ratio
Similarity Score	0.79
Recency Score	0.0133
Social Trend Score	0.0145
Attention Score	0
Favorite Score	1.7231
Mutual Friends Score	3.5792
Mutual Followers Score	0.0003
Common Hashtag Score	0
Reciprocal Score	20.5079
LDA based similarity	1

Table 4.1: Odds Ratio for LDA

to derive the candidate set from, not only allows us to scale the system to originators with larger social circles, but also helps in improving the time efficiency by reducing the computational effort of deriving candidate tweets and extracting the feature scores from them.

4.2 Improving Tweet-Content Related Feature Set

TweetSense uses cosine similarity as the tweet-content related feature that deals with the content (excluding the hashtag related features) of a tweet. To study the effectiveness of this feature, an experiment was designed to add an additional feature which is the distance between the tweet Q_x and each of the candidate tweets in the reduced dimension space. LDA is used to represent tweets in the reduced dimensions. The odds ratio of the system with an additional feature is presented in the table 4.1. Although, there was a very minor improvement in the similarity based on the reduced space of terms, this was not a sufficiently large gain.

An originator of a tweet can also be influenced by external sources and post tweets based on these influences. To address this aspect of the generative model, candidate hashtags are derived from the corpus of tweets based solely on the content of the tweet. The crawled tweets are indexed using elasticsearch [15]. The "search" API [2] is used to fetch tweets based on the content of the tweet.

TweetSense uses cosine similarity scores as a feature to measure similarity between two tweets. We propose to use three more features to strengthen predictions based on the content of the tweet. These features are defined only for the candidate hashtags derived based on the content and are defined as follows:

Similarity with bag of words corresponding to a hashtag: Tweets are very short text documents with very few words. Therefore, the cosine distance between two tweets may not be the best measure to rank candidate hashtags. We define a new measure as follows.

The bag of words corresponding to a hashtag is defined as the collection of all the words that have occurred with the hashtag in our Twitter dataset. All the stop words are eliminated and only the root of a word is preserved to get better results.

This similarity measure is based on the content of the tweet and the bag of words describing a candidate hashtag. BM25 [1] is used to find the distance between the tweet Q_x and the bag of words corresponding to the candidate hashtag $\langle CH_{xj} \rangle$.

Distance from the cluster of hashtags: Bag of words corresponding to hashtags are clustered using BM25 [1]. These clusters are used to compute the distance between a tweet and the corresponding cluster to which the candidate hashtag of the candidate tweet belongs. This measure is called the hashtag cluster distance.

Hashtag popularity: This measures the popularity of hashtags within the candidate set of hashtags derived based on the content of the tweet. The "trend" score used by TweetSense differs from this feature as the popularity is assigned only within the candidate

hashtag set derived from the content of the query tweet, Q_x and the candidate set based on the social circle is not used. This feature captures the most popular candidate hashtags based on just the content.

For the purposes of efficiency, the candidate tweets are first derived from the social circle and this set is mutually exclusive from the set of candidate hashtags derived based on the content of the tweet.

Figure 4.1 depicts the above scheme of deriving candidate set of hashtags to build a model and also depicts the testing process which is similar to that of TweetSense.

4.3 Personalization of Hashtag Adoption Pattern

The adoption pattern of hashtags varies from user to user. The pattern can also vary with time. So far, we have handled personalizing hashtags according to the user, his social circle, and content of the tweet. In this section, we propose to build personalized models based on the originator's own history of tweets and hashtag adoption pattern rather than employing a single model for all users as proposed by TweetSense. It eases the stress on finding all kinds of training instances indicative of all kinds of adoption patterns that can apply for Twitter users. Therefore, we propose to employ an individual model based on the originator's own set of tweets to predict hashtags for the originator's set of test tweets. The figure 4.2 depicts the proposed system for a single user. Here, the set of tweets belonging to a originator is divided into a set of test tweets and a set of training tweets. The logistic model is built using this set of training tweets. The model is tested with the set of test tweets of the originator.

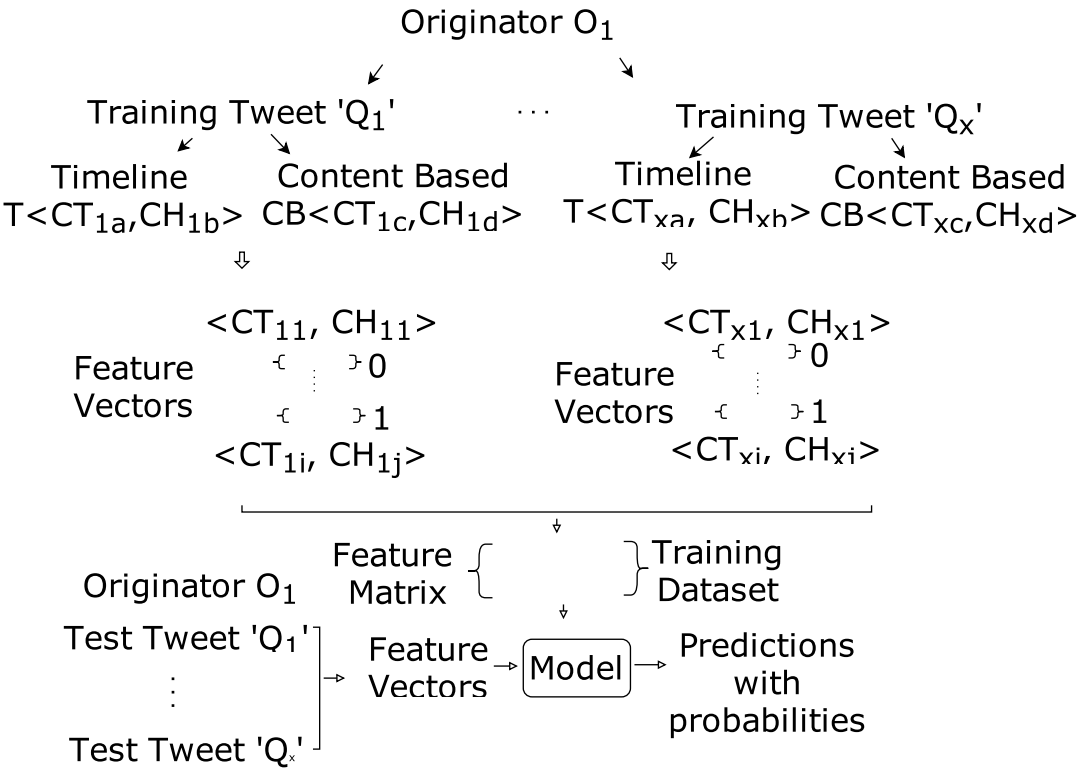

Figure 4.2: Personalized Classifier for a Single User

Chapter 5

EXPERIMENTAL SETUP

5.1 Dataset Description

The proposed approach called E-TweetSense and TweetSense are tested with the same dataset. The dataset contains up to 1,500 user timeline tweets per user. The API limits the number of friends and followers for a user that can be crawled to 5000 users. The favorite tweets are limited to the recent 200 tweets per user.

The dataset contains following 7,945,253 million tweets. Further details about the dataset can be found in Table5.1.

Characteristics	Value	Percentage
Total number of originators	63	N/A
Total Tweets Crawled	7,945,253	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 at least one Hashtag	560,849	7.06%
Total number of Favorite Tweets	716,738	9.02%
Total number of tweets with user @mentions	4,658,659	58.63%
Total number of tweets with Retweets	1,375,194	17.31%

Table 5.1: Characteristics About the Dataset Used for the Experiment

5.2 Evaluation Method

Evaluation for one model for all originators is described below:

- Divide the set of originators for training and testing.
- Build the statistical model corresponding to the set of tweets corresponding to the training set of users.
- For each originator in the test set, pick the tweets with a hashtag and deliberately remove the hashtag for evaluation.
- Run the system to get the recovered list of hashtags.
- Verify if the ground truth hashtag exists in the list.

Evaluation for personalized models for each originator is described below:

- Divide the set of tweets of originators into training and testing tweets
- For each originator, build a statistical model with the set of training tweets created by her.
- For each test tweet of the originator, deliberately remove the hashtag for evaluation.
- Run the system to get the recovered list of hashtags.
- Verify if the ground truth hashtag exist in the list.

Chapter 6

EVALUATION AND DISCUSSION

In this chapter, we present an internal and external evaluation of the proposed system.

Table 6.1 lists the different experiments conducted by limiting the social circle of the originator.

6.1 Internal Evaluation

The system is tested internally using precision at N by varying N, and also by varying the size of the social circle. One of the variations, "Limited_5020 Social Circle Only", limits the derivation of the candidate set to the social circle of the originator, and it does not derive any tweets based on the tweet content. A precision score of 1 is assigned to the prediction if the hashtag is predicted correctly at any rank position of N between and including 1 and 20. The association between an exposure and an outcome is studied using odds ratio computed on the learned statistical model.

Experiment	Limited_1	Limited_3010	Limited_5020	Limited_5020 Social Circle Only	Baseline
Number of Mutual friends	0	30	50	50	N/A
Number of Mutual followers	0	10	20	20	N/A
Social Circle Size	1	41	71	71	all

Table 6.1: Description of the Constraints on the Social Circle

6.1.1 Internal Evaluation Using Precision at N

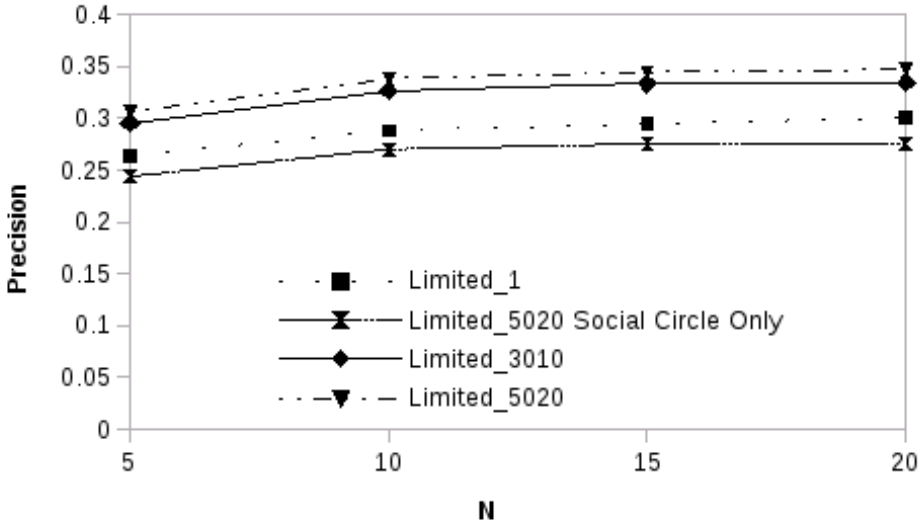

Figure 6.1: Ratio of Test Tweets Correctly Predicted with Varying Social Circle

The system is evaluated internally using precision at N = 5, 10, 15, and 20. Figure 6.1 shows the precision of the system as a ratio of the number of tweets for which a system has correctly predicted the hashtags to the total number of test tweets. The table 6.1 shows the description of each of the variations of the system in the figure 6.1.

Observations: Most of the correct predictions are achieved by considering tweets belonging to the originator. This is expected as the originator tends to reuse hashtags. The social circle provides a scope to bring in the candidate tweets from the social circle of the originator, which as shown in the graph 6.1, can help in improving the accuracy of predictions compared to Limited_1 case. It also indicates that the first set of 40 (Limited_3010) users contribute more to the gain in accuracy of predictions than the next set of 30 users (Limited_5020).

It can also be observed that removing the content based candidate set and the additional content based features (section 4.2) can reduce the accuracy further.

All Features	Limited_1	Limited_3010	Limited_5020
Similarity Score	0.0287	0.0644	0.1072
Recency Score	0.007	0.0268	0.0284
Social Trend Score	0.0023	0.0033	0.0029
Attention Score	662.2701	1024905461.3103	113323.9782
Favorite Score	0.0977	0.0184	0.1673
Mutual Friends Score	0.0128	0.4902	0.4204
Mutual Followers Score	0.0025	0.0002	0.0001
Common Hashtag Score	0	0	0
Reciprocal Score	0.1803	0.5569	0.9341
Hashtag Distance	2.7785	9.2444	1.1451
Hashtag Cluster Distance	0.2848	0.0582	0.7017
Hashtag Popularity	3.9263	5.1198	7.2272

Table 6.2: Estimation of Odds Ratio by Feature Selection

6.1.2 Feature Importance Using Odds Ratio

The Table 6.2 shows the most important features in predicting the hashtags according to the learned model. The table shows the important features by varying the social circle as well.

Observations: According to TweetSense, "Mutual Friends" score followed by "Mutual Followers" score are the two most important features in predicting the correct hashtags. The new system takes advantage of this fact to filter out the tweets from users that do not necessarily impact the hashtag adoption pattern of the originator. The new system learns from the other features like the "Attention" score followed by the "Hashtag Distance" or "Hashtag Popularity" (with the tweets derived based on the content of the tweet). It can be noted that the features "Hashtag Distance" or "Hashtag Popularity" are

Experiment	1	2	3	4	5
Baseline	0.1193	0.195	0.2359	0.2754	0.2982
Limited_5020	0.2178	0.2627	0.2868	0.3049	0.3156

Table 6.3: Precision at N = 1,2,3,4,5

better at capturing tweet-related signals than the "Similarity" feature which was being used by TweetSense.

6.2 External Evaluation

The system is evaluated by comparing the system with the baseline which is the TweetSense system [21]. The accuracy and the ranking quality of the system is compared using precision at *N*. We also provide external evaluation results with respect to the time taken for prediction by comparing the number of candidate tweet-hashtag pairs and the actual time taken to extract features from the candidate hashtags. We also compare the importance placed by the learned statistical models on the features for predictions using odds ratio as a measure.

6.2.1 Precision @ N as Compared with the Baseline

Observations:

The table 6.3 shows that E-TweetSense performs much better than TweetSense in the early predictions. This indicates that the system effectively removes noisy candidate tweets while still preserving the important candidate tweet-hashtag pairs.

Figure 6.2 is plotted by altering the definition of "Precision" score as the ratio or tweets that are predicted correctly to the ratio of tweets that have a candidate hashtag that matches the ground truth. The graph confirms that limiting the social circle is not adversely affecting the accuracy of the system. However, limiting the social circle does reduce the number of tweets for which the candidate set has a matching ground truth.

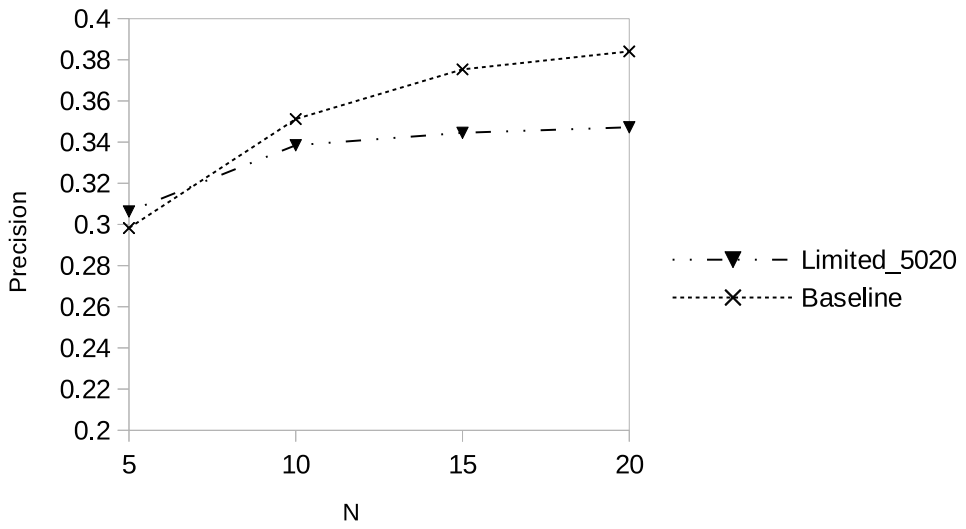

Figure 6.2: Precision E-TweetSense Versus TweetSense

Figure 6.2 shows the precision curve with respect to all test tweets. Total number of test tweets considered is 1492 tweets which corresponds to 29 different users.

6.2.2 Results for Feature Scores Comparison Using Odds Ratio

Here, we present the most important feature in predicting hashtags according to the TweetSense system and the proposed system. TweetSense placed the most importance on the "mutual friends" of the originator. The new system places most of the importance on the "attention" followed by the "hashtag distance" or "hashtag popularity" features according to the odds ratio in table 6.2

6.2.3 Size of the Candidate Set

The figure 6.3 shows the average, minimum, and maximum number of candidate tweet-hashtag pairs that are considered per tweet Q_x.

Observations: There is a trade-off between the accuracy of predictions and the time taken to extract the features. TweetSense can take a very long time to process the candidate

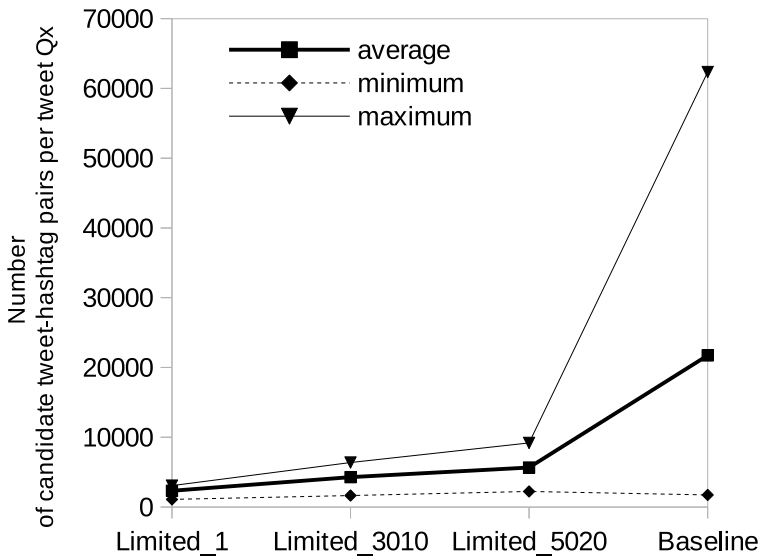

Figure 6.3: Number of Test Tweets Correctly Predicted with Varying Social Circle

Experiment	Limited_1	Limited_3010	Limited_5020	Baseline
Minimum	1090	1640	2231	1728
Maximum	3064	6374	9184	62349
Average	2323.79	4281.75	5654.44	21731.03
Standard deviation	411.00	1183.03	1841.37	15549.0216905914

Table 6.4: Description of the Constraints on the Social Circle

set of tweet-hashtag pairs depending on the size of the social circle. The improved system reduces computational effort by selectively choosing users to acquire candidate tweets from. Table **??** shows that by considering 70 users from the social circle, and 100 tweets per user, the number of candidate tweets based on the social circle, on an average, gets reduced by more than 3 times.

Note: The candidate set of tweet-hashtags pairs include the candidate set derived based on the content of the tweet (up-to 2000 tweets) for Limited_1, Limited_3010, Limited_5020.

6.2.4 Time Taken to Extract Features

The time taken to extract features from a candidate set per tweet Q_x , has reduced by 98% on an average as shown in the table 6.6.

TweetSense has a constraint imposed on the originator that the originator can have utmost 300 friends. TweetSense considers up to 1500 tweets per friend. This can potentially cause a scalability issue to extract features from a single tweet Q_x. Twitter allows every user to follow/befriend 2000 followers [17] by default. This is a potential problem in scaling TweetSense to originators who have a social circle that includes more than 150 users as the article [7] shows that average number of followers and following is approximately 350 for 82% of the users.

The proposed system limits the possible number of candidate tweets to 100 tweets per user and utmost 71 users. Thus, there is an upper limit on the time taken to process a single tweet. The feature extraction by TweetSense could take over 8 minutes for a user with the social circle size between 250 and 300 users. As the social circle size increases, the number of candidate tweets to be acquired by TweetSense increases linearly with a steep slope. As the social circle size increases the number of candidate tweets to be acquired by E-TweetSense increases linearly till the social circle size is below 71 users and then becomes constant.

On average, it takes about 0.3 minutes to derive and to extract the features for candidate tweets based on the content of the tweet Q_x. Most of the time is consumed in looking up the elasticsearch index to retrieve tweets based on the content. The percentage reduction in computation time is about 91% on average and 88% in the worst case.

Size of social circle	Baseline	Limited Social Circle	Percentage Reduction
<100	2.0059	0.0535	97.33%
100-150	3.6115	0.0492	98.63%
150-200	3.2364	0.0462	98.57%
200-250	3.7256	0.0718	98.07%
250-300	8.5415	0.0998	98.83%
Average	4.2242	0.0641	98.48%
Maximum	8.5415	0.0998	98.83%

Table 6.5: Time Taken (in minutes) for Extracting Features per Test Tweet

	Baseline	Limited Social Circle with Content Based Derivation	Percentage Reduction
Average	4.2242	0.3641	91.38%
Maximum	8.5415	0.9998	88.29%

Table 6.6: Time Taken (in minutes)

6.2.5 Personalization of Hashtag Adoption Pattern

The data from Limited_5020 is used to verify this method. The sets of tweets corresponding to a user is split into testing (25%) and training (75%) tweets. A Personalized model is built using the training set of tweets for a user as shown in figure 4.2. The built model is used for testing. The accuracy of prediction is measured using precision at N and are tabulated in the table 6.7. It can be observed that the personalized models perform better when compared with the performance of a single statistical model at lower values of N. However, with this approach, there is an overhead of maintaining and updating personalized models for every originator on Twitter.

N	1	2	3	4	5
Personalized model	0.230	0.26	0.28	0.295	0.308
Single model	0.157	0.224	0.2731	0.303	0.317

Table 6.7: Precision

Chapter 7

CONCLUSION

We have improved the time taken to construct the candidate hashtag set by carefully picking users from the social circle of the originator. The accuracy of suggestions was preserved for most users as compared to the baseline system. Further, the accuracy of the system is improved for earlier predictions as the improved system eliminates noisy candidates. To further personalize the hashtags for context recovery, we proposed to use a custom classifier for a user.

There are many ways this work can be further extended. The individual model for each user needs to be updated. The evidence for the effectiveness of personalization of models needs to established. The frequency of updation of an individual model can be learned for different users based on their activeness on Twitter can be explored. Addressing the third way of tagging according to the generative model can be explored.

REFERENCES

[1] Bm25, . URL https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-similarity.html#bm25.

[2] Search api, . URL "https://www.elastic.co/guide/en/elasticsearch/reference/current/search.html".

[3] URL http://www.newstatesman.com/sci-tech/2014/08/twitters-taking-away-your-control-over-what-tweets-you-choose-see.

[4] URL http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SGD.html.

[5] Brian S Butler. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research*, 12 (4):346–362, 2001.

[6] Wikipedia contributors. Hashtag, 2015. URL http://en.wikipedia.org/w/index.php?title=Hashtag&oldid=660837077. Page Version ID: 660837077.

[7] Twitter Count, 2010. URL http://thenextweb.com/socialmedia/2010/09/30/twitter-statistics-82-of-twitter-users-have-less-than-350-followers/.

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

[9] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[10] Bernardo A Huberman, Daniel M Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045*, 2008.

[11] Paul 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.

[12] Kevan Lee. URL https://blog.bufferapp.com/

a-scientific-guide-to-hashtags-which-ones-work-when-and-how-many.

[13] Jieying She and Lei 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*, pages 371–372. International World Wide Web Conferences Steering Committee, 2014.

[14] Ian Steadman. URL http://www.newstatesman.com/sci-tech/2014/08/twitters-taking-away-your-control-over-what-tweets-you-choose-see.

[15] Zachary Tong. Elasticsearch index, 2013. URL "https://www.elastic.co/blog/what-is-an-elasticsearch-index".

[16] Twitter, . URL https://support.twitter.com/groups/52-notifications/topics/213-following/articles/14019-faqs-about-following.

[17] Twitter, . URL https://support.twitter.com/articles/66885-why-can-t-i-follow-people.

[18] Twitter, . URL https://support.twitter.com/articles/49309-using-hashtags-on-twitter.

[19] Twitter, . URL https://support.twitter.com/articles/14023-what-are-replies-and-mentions.

[20] Twitter.Inc. URL https://about.twitter.com/company.

[21] Manikandan Vijayakumar. *TweetSense: Recommending Hashtags for Orphaned Tweets by Exploiting Social Signals in Twitter*, 2014.

[22] Eva Zangerle, Wolfgang Gassler, and Gunther Specht. On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Eva2013*, 3(4):889–898, 2013. ISSN 1869-5450. doi: 10.1007/s13278-013-0108-x. URL http://dx.doi.org/10.1007/s13278-013-0108-x.

APPENDIX A

RECOVERING MORE THAN ONE HASHTAG

TweetSense recommends many hashtags as a possible hashtag to a tweet. However, it does not attempt to recommend more than one hashtag as the context of a tweet. Formally, given a tweet Q_x and a set of hashtag(s) as the possible context to the tweet Q_x, recovering more hashtags based on this tweet and the additional information which is the recovered hashtag.

A system was developed to accept ϵ - set of recovered hashtag(s) (which was produced by TweetSense) and the corresponding tweet Q_x as input to the system to recover more hashtags. The additional information in this case is the ϵ - set of recovered hashtag(s).

Pre-processing: To address the problem of recovering more than one hashtag(s), hashtag co-occurrence frequency was learned for each hashtag in our corpus. Most popular words occur with co-occurring pairs of hashtags were also learned form the corpus.

To test the potential of this approach, a simplified version of a multiple-hashtag recovery system was built to experiment the feasibility of recommending more than one hashtag per tweet. For each hashtag in the set ϵ a set of 10 more hashtags were recovered as the possible set of hashtags for the tweet Q_x based on the co-occurrence frequency and the co-occurring word frequency. The set of 10 hashtags were determined by using a simple normalized ranking of the co-occurrence frequencies.

This method did not yield feasible results to further improve the multiple hashtag recovery system. The following were identified to be the road blocks to solving this problem:

- The limit on the number of characters per tweet is 140 characters this limits the number of words. As a result the tweets have fewer words.
- Hashtags are the important words in a tweet. When the ground truth hashtags are removed from a tweet, the important words are removed from the tweet. And the tweets is left with fewer unimportant words that may not be of much help in determining context especially when multiple ground truth hashtags are removed for the purposes of evaluation.
- The tagging of hashtags to a tweet is very sensitive to the originator.

The figure A.1 shows the example tweets that correspond to a few popular hashtags on Twitter. The figure illustrates the above listed issues.

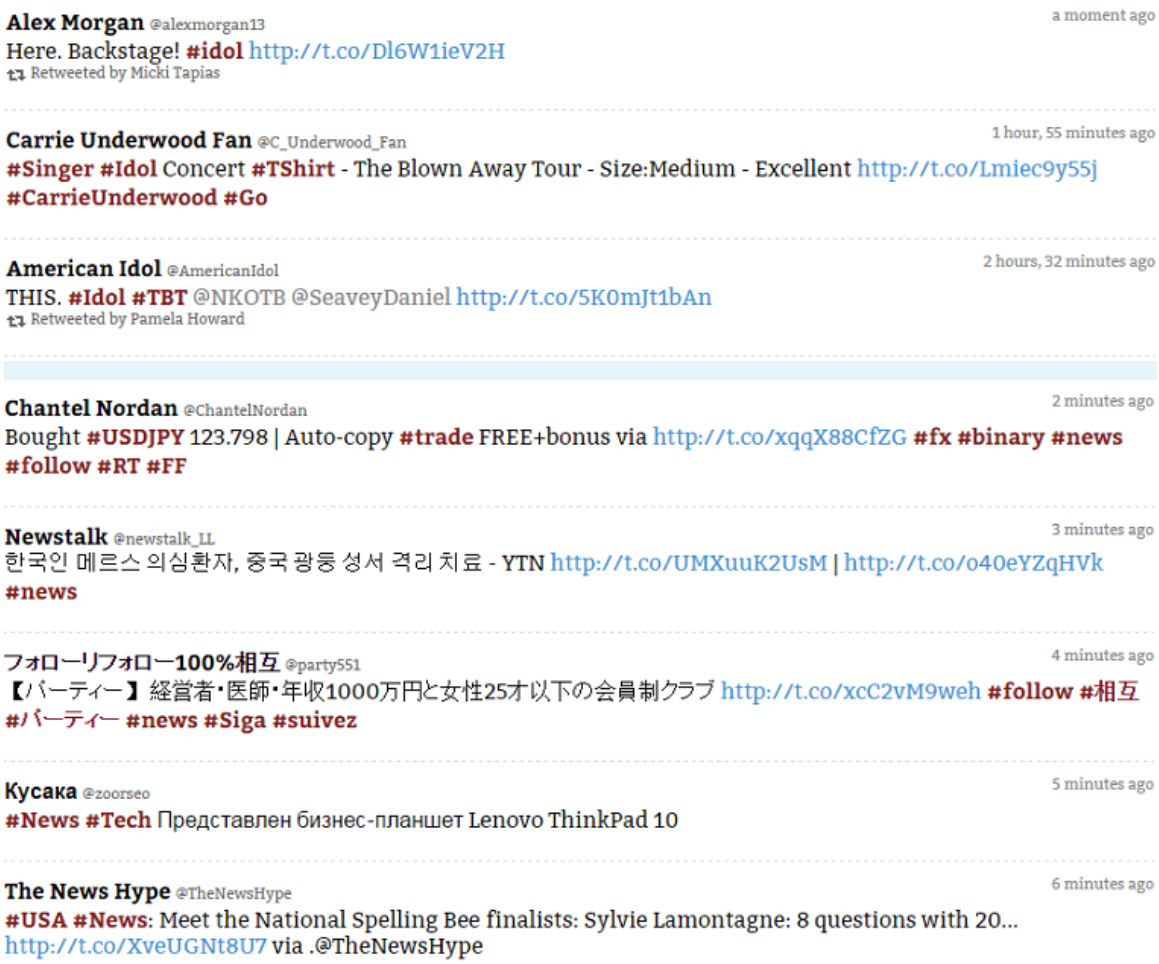

Figure A.1: Tweets with Multiple Hashtags

Shawn Nelson @Shawny_Nelson — 5 hours, 18 minutes ago
Just landed in LA! Working with @chelsiehightowr to perfect the @lovesac performance for **#DODLA #DWTS** http://t.co/iq2x5lSxsJ

Elie Elia @Elie_EOfficial — 6 hours, 49 minutes ago
@DalidaKhalil Congratulations Dalida for your wonderful achievement in **#DWTS** You really deserve the best **#ElieEliaWorld** We are proud of u!
Retweeted by ﻩ ﻩ $

GoldenYen @Golden_Yen — 8 hours, 49 minutes ago
#dwts #sick #USA #sup #NBA #GIRL #TAIWAN #FUNNY #SEXY #newyork #curry #skater http://t.co/YUhNJD6a3X

Sherri Falgout @SherriFalgout — 10 hours, 6 minutes ago
@TheRue @iamValC Y'all were my fave from Day 1 on **#DWTS**. Voted for y'all the entire competition. I really miss y'all performing. **#Fan4Ever**

team_val_and_rue @lady_ddddd — 16 hours, 22 minutes ago
I miss watching **#TeamValenRue #DWTS**. Also miss watching the family rooting them on. @Scout_Willis @buuuski @justdemi http://t.co/lDhds0JaXm
Retweeted by Demi Moore

STORMIE @StormieLynchR5 — 19 hours ago
OMSTARS I AM HAVING **#DWTS** WITHDRAWALS❤GONNA HAVE TO RE WATCH ALL THE EPISODES AGAIN & PRETEND I DONT KNOW WHO WON❤VOTE **#TeamRallison** ❤XOXOX❤
Retweeted by Fotini

Lee Runham @xLeerunham — 1 hour, 29 minutes ago
Is this Britain's Got Talent or X Factor? **#Xfactor #BritainsGotTalent** @BGT Full of singers, WE WANT VARIETY **#BGT**

Rina Nosenko @irinanosenko — 3 hours, 26 minutes ago
#AlexeySmirnov #Ksenon #AMANO #xfactor #ladygaga #BadRomance #Ukraine #Russia #rock #music #cover #singer #LTBL https://t.co/8rBqaPdizX F
Retweeted by Oksana_pohodneva

Таня @Setaa57 — 3 hours, 26 minutes ago
#AlexeySmirnov #Ksen #Ksenon #xfactor #BadRomance #cover @J_Ksen & **#AMANO #Ukraine #rock #music #love #live #LTBL** http://t.co/FceMzxqfWc
Retweeted by Oksana_pohodneva

Figure A.2: Tweets with Multiple Hashtags (continued)

APPENDIX B

TWEETSENSE WITH SUPPORT VECTOR MACHINES (SVM)

All Features	SVM Model Odds Ratio
Similarity Score	0.4144
Recency Score	3.3033
Social Trend Score	3.5399
Attention Score	4.9359
Favorite Score	-0.9085
Mutual Friends Score	-3.2752
Mutual Followers Score	3.3581
Common Hashtag Score	7.7232
Reciprocal Score	0.124

Table B.1: Estimation of Odds Ratio by Feature Selection

TweetSense used Logistic Regression to build a statistical model, we conducted an experiment to compare the performance of TweetSense by using a different method. The figure B.1 shows the flow chart of TweetSense by using Support Vector Machines as the algorithm to build the statistical model. Table B.1 shows the odds ratio for the build model

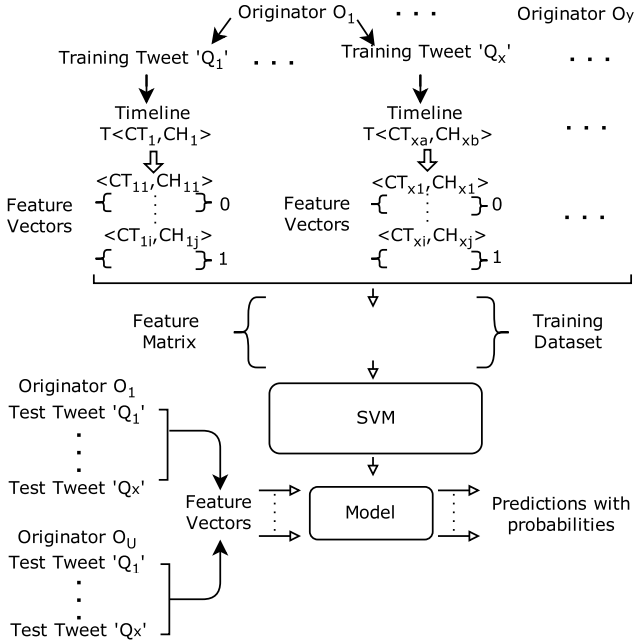

Figure B.1: TweetSense Using SVM

The prediction accuracy by using this model was lower when compared to the model build using logistic regression. The accuracy of the models are compared using precision at $N = 5, 20$ in the table B.2.

The number of input training samples is of the order 10 million instances. Therefore, batch mode variation of the stochastic gradient decent API [4] was used to build the

N	5	20
Logistic Regression	0.2982573727	0.3840482574
Support Vector Machines	0.1655495979	0.3016085791

Table B.2: Accuracy Comparison between Logistic Regression and SVM Models for TweetSense

above SVM model. Given, the number of training instances and lower accuracy of the batch mode SVM, Logistic Regression is used for experimentation in this thesis.