|  |
| --- |
| ***HASHTAG POPULARITY PREDICTION*** |
| ***MANISHA KRISHNAMOORTHI ITHAL******CIS 787******December 12, 2016*** |

1. *INTRODUCTION*

*Given the ubiquitous nature of social media networks and the need for constant connectivity, it is not very surprising that the social media has revolutionized our communication and has become a huge part of our lives. .Twitter is one such social media which is gaining significant attention off late. According to a recent survey more than 50 % of people in the United States use either Twitter or Facebook. Twitter is a great platform to share ideas, thoughts with the followers. These ideas or thoughts are shared by “tweets” which helps in diffusing the information across the network. Given the viral nature of the social network, the ideas or thoughts shared have a huge probability of becoming popularity in a very short span of time. Because of this fast dissemination of information on twitter, we would be benefitted greatly if somehow we could predict what ideas or topics would become popular.*

*In twitter tweets are usually annotated with hashtags. Hashtags helps us, users in finding a message with a specific content. Users create a hashtags by using a “#” infront of a text or a label. Although the usage of hashtags began with Twittter, it extended to other social networks as well. Hashtags are mainly used to express the content and or context of a given message. Hashtags are widely used for sharing contents for a specific events or topics. Usage of hashtags started to trend so much so that they have come to known as “PARALANGUAGE” of Twitter. Newly created hashtags usually define emerging topics or events, these hashtags diffuse throughout the network, diffusion is mainly by the adoption of the hashtag by followers of owner of hashtag. If the hashtag is adopted by many users then that hashtag becomes a trend (a trend is a topic that’s being discussed by most of twitter users for a particular time period) on twitter.*

*The trending topics and hashtags are an effective way to boost communication to reach people beyond the normal follower circle. So it would help us greatly if we could find out or in some way predict what hashtags become popular in time as it will provide us an intuition on how topics and ideas spread through and also on trending topics in twitter. However Prediction of Hashtags’ Popularity differs from the normal text classification in a way because the features change so quickly that it is hard to cluster them. This hinders the classification when new tags emerge. And also since there is a limited information about a new hashtag it is usually onerous to predict their popularity.*

*So this report summarizes my experimentation of predicting the popularity of hashtags by extracting the information that is already available on twitter.*

# *Motivation*

*Getting ahead in a world filled with competition is quite challenging for any business to emerge. To come up with a successful business model, the most important thing is to know how to advertise effectively, as in to know what exactly the customer wants and expects. And the reality as we know, is that customers’ wants keep changing from time to time. It then becomes inevitable to know about the customer’s perspective or how they might respond to a particular type of product. This kind of intuition can be obtained by twitter: especially by knowing or predicting what might trend on twitter, so that companies are ensured to get ahead of the customers.*

# *Problem Statement*

*Usually tweets are generally associated with timestamps, Given a hashtag h in the time interval t, our task is to predict if hashtag is going to be popular if it is going to be popular for t+1, now it is difficult to come up with a discrete value for hashtags popularity. But we can relax the problem by formulating it as ranking or a classification problem into 5 classes not popular, marginally popular, popular, very popular, and extremely popular respectively.* *We define five ranges of an exponentially*

*increasing size: [0, f], [f, 2f], [2f, 4f], [4f, 8f] , (f is the popularity score.) One key note is that the ranges or classes depend on twitter data collected as in the number of tweets.*

# *RELATED WORK*

*Twitter has a huge number of users and also has huge diversity among the users today (it boasts about 317 million active users in a month). And around 500 million tweets are generated per day. Because of this humongous amount of data being generated and disseminated, research institutions and business companies are enabled with a huge opportunity of obtaining this data and exploit the underlying patterns. The area of information dissemination or diffusion has been of an area of interest among the study of social networks.*

*Most papers analyze the twitter datasets to study the properties of tweets and predict on how information spreads. [6,7,8]. People have also studied on how to predict the hashtags spread as well to [1,2] using the frequency and the data extracted by hashtag itself.*

*Topic detection and topics trending have become an important and widely researched task in the twitter. Naaman, Becker, and Gravano (2011, p. 908) defined a trend profile to be a “collection of tweets with a topical keyword” (e.g., earthquake, Obama). The authors classified trends to be “exogenous trends” and “endogenous trends “using a set of features that mainly characterized the trend. These features include “retweet fraction”, “reply fraction”, and various other components of graph. The authors concluded that both the trends are characterized by different set of features. The authors used SVM to classify tweets based on the sentiment.*

*There was also another study Romero, Meeder and Klienberg [10] that analyzed different topical hashtags had different propagation of the hashtags and concluded that some hashtags were more persistent than others based on geological locations.*

*The experimentation that is presented in this report is based on some of above mentioned papers. The main difference here is that I’m interested in predicting which hashtags will be popular in next hour (t+1) or day rather than finding out which hashtags will be popular among a set of people or in a particular location.*

## *IMPLEMENTATION DETAILS*

# *Data Collection*

*Since data extracted from Twitter is no longer publicly available I had to collect the data from the twitter API. Twitter offers a convenient API to get the data, which is quite easy to use. It allows authenticated developers to collect data from its network. The API provides with sophisticated interfaces to manipulate the twitter data, author a new tweet, read author profile, and also gives access to global stream of public tweets. However the REST API however, has a 15 minute window which allows a maximum of 180 calls according to their policy which hinders the speed of data collection.*

*I used the rest api ‘s to collect the data that i.e., tweets , I collected around 3 million tweets and only 8~ 10 % has hashtags associated with them . Also the author profile information was retrieved. After excluding non English based tweets, we ended up with more than 1 to 1.5 million tweets. Each tweet is listed in one line starting with the screen name followed by user id, timestamp, status id and the tweet context. A tweet may contain RT (indicating a retweet), mentions, hyperlinks or hashtags. Twitter allows use of two special characters “#” for marking a hashtags and “@” for marking a user. The following is an annotated sample returned by the API*.

*“*@MAUREEN\_WHITE\_ (screen name), 3225914647 (user id), 1433062773000 (time stamp), 604935234283372546 (status id) - RT @Cynthiapoet: Sunset Leopardess by Mark Dumbleton #WeAreAlive #Animals #Photography http://t.co/mu3q6UBk3t (tweet context)*”*

# Data Preprocessing

*One of the most cumbersome and time consuming process of any data mining task is cleaning and preprocessing data. It is most convoluted because there is no constraint on the structure of a hashtag. Users also use different variations of the same hastag. Ex : “DogBothering”, “DOGBothering”. Even though they converge to the same hastag, the frequencies vary greatly. So I assumed the hashtags are case insensitive*

*Another major obstacle was to remove the noise from the actual content of data. I removed all the different characters like [*;, $, %, ^, &, \*, +, }, >, ?] *so that the content of the hashtags has more importance over its structure, and the syntax of the hashtags*. *I also removed the stopwords.*

# *Feature Description and Extraction*

*In order to predict the popularity of hashtag, we must first try to define our term “popularity”. Popularity is a measure that is usually based on the number of users who adopt it. Based on some of papers [1] I have set the popularity score ∮ to be 25 , ∮ is the number of users who will atleast post one tweet using that hashtag. As we can see it , ∮ is based mainly on the user profile information as in the number of followers he has , the number of times the users have been retweeted . It is not right to say that the popularity depends only on user profile , it also depends on the content of the tweet like its sentiment polarity , the words etc. So the features that describe the hashtag can be broadly classified as Context Features and Content Features. The table is given as below.*

|  |  |
| --- | --- |
| Feature | Description |
| No of authors followers | *Can be obtained by querying the tweet and user in the twitter api* |
| No of tweets issued by the user | *Can be obtained by querying the tweet and user in the twitter api* |
| The number of users mentioned | *Can be obtained by querying the tweet and user in the twitter api* |
| Age of the tweet containing the tweet | *Can be obtained by Timestamp of tweet* |
| No of authors friends | *Can be obtained by querying the tweet and user in the twitter api* |
| Tweet Count | *Fraction of tweets containing the hashtag in the time interval t* |
| ReplyFraction | *Fraction of reply’s containing the hashtag in the time interval t(manually extracted)* |
| Retweet Fraction | *Fraction of retweets containing the hashtag in the time interval t(manually extracted)* |
| No of words in a hashtag | *Number of segment words from a hastag* |
| Polarity of the tweet that contains the hashtag | *The range is from -1 to 1 , 0 being the neutral* |
| URL Fraction | *(twitter URLS always occur with hashtags) – fraction of tweets containing the URL* |

# Methods Used.

1. CODE STRUCTURE

TWITTER API

DATA REPOSITORY

RESULTS

CLASSIFIERS

FEATURE EXTRACTION

WORD PROCESSING

*The above figure shows a very high level view of the code implemented. As we can see it has 5 modules which have functionalities decipherable from their names. The code is written in python. It makes use of object oriented concepts.It could be easily extended to add more features and make the classifiers more intelligent. The code makes use of many libraries for the computation. The main libraries that I have used are*

1. *Ntlk : for stemming and various language processing functions*
2. *Numpy : for array objects*
3. *Scipy: for scientific calculations*
4. *Scikit learn : for some classifiers*
5. *CLASSIFIERS USED*

*I have used the Naïve Bayes and the SVM classifiers to classify the data. For classification, I have used around 75% to train the classifier and 25% to test the data. In our problem setting we have set the ∮ to be 25 and also t as 1 day. This setting is subjective and we can set it to an appropriate threshold. A higher threshold might lead to overfitting and a lower threshold might lead to underfitting.*

*Once the classification is done , I did cross validation which helps in validating our results.*

*Naïve Bayes we can get better results by performing a cross validation, this also helps in obtaining the a good approximation to the parameters such as alpha.(alpha I got as 0.92).*

*However with SVM I used only five features that I found important because giving too many features makes finding the separate hyperplane is very difficult.*

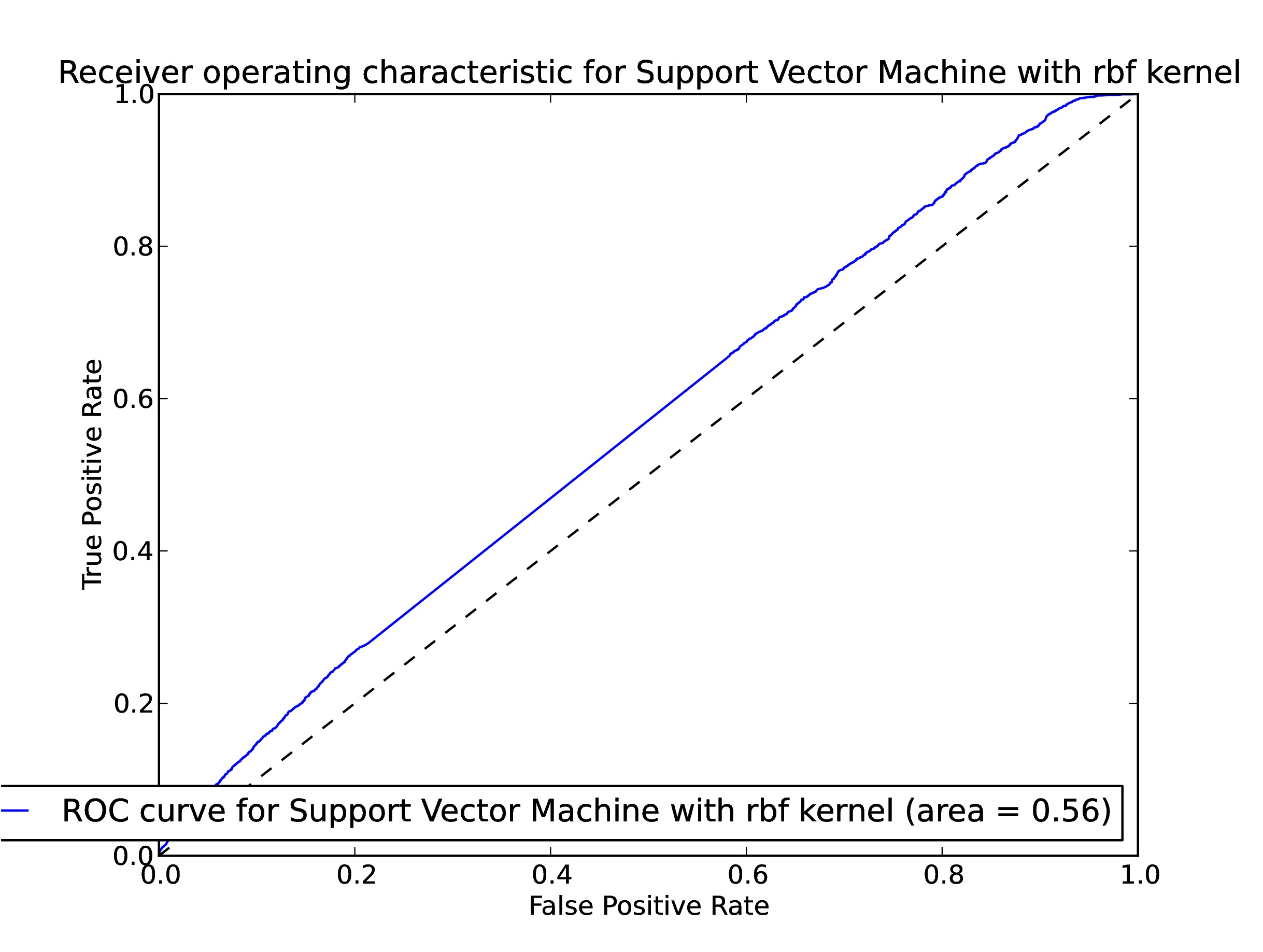
1. *RESULTS*

*The actual hashtags that fall into the five classes are shown below.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *0,∮* | *∮,2∮,* | *2∮,4∮,* | *4∮,8∮* |
| *1 day prediction* | *2000* | *300* | *60* | *10* |

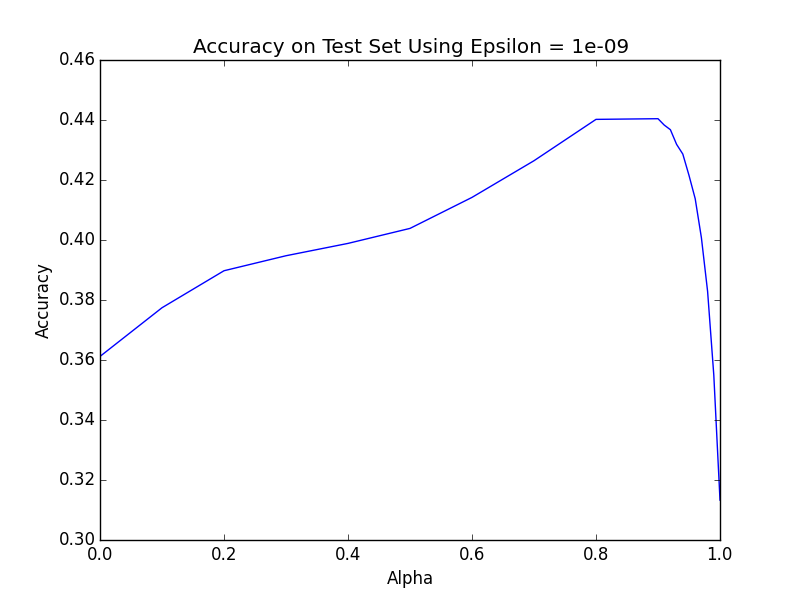
*The table shows the actual hashtags distribution for a day I have grouped them according to their popularity scores.*

*So we use NB and SVM to predict the popularity of hashtags for a time interval of a day. The following graphs show the ROC curves for the two classifiers .*

**

*The curve has been smoothened to negate the noises.*

*The alpha varying graph in Naïve bayes*

**

*Precision for the two algorithms*

*Orange – Nb*

*SVM – Blue.*

# CONCLUSION

*As we can see the SVM is a better classifier than the Naïve Bayes which is intuitive because the naïve bayes assumes that the features are independent.*

*Out of curiosity to rank which features are more important, we performed chi- square test on selected features to find out which features were important , here are some interesting results.*

|  |  |
| --- | --- |
| *No of followers* | 11343478 |
| *No of tweets by the user* | 1378000 |
| *No of words* | *988731* |
| *No of hashtags* | *199* |

# *FUTURE WORK*

*Due to limited time and scope of the project I was not able to include all the features which I mentioned in the proposal mentioned below.*

|  |  |
| --- | --- |
| Hashtag Clarity | Kullback–Leibler divergence of word distribution between tweet containing the hashtag and tweets collection T |
| Segment Clarity | *Kullback–Leibler divergence of distribution between tweets containing the any segment word in hashtag and Tweets collection* |
| AvgAuthority | *Average authority of users with hashtag among the users who used it* |
| Number of users forming triangles in the graph | *helps in estimating the ties between users* |
| Density of graph | *estimates the sparsity of graph.* |
| Ratio of connected components and number of nodes | *estimate the overall interaction of users in graph.* |

## *Citing Internet Sources*

*[1]**On Predicting the Popularity of Newly Emerging Hashtags in,* ***Zongyang Ma, Aixin Sun, and Gao Cong***

*School of Computer Engineering, Block N4, Nanyang Technological University, Nanyang Avenue, Singapore.*

*E-mail: zma4@e.ntu.edu.sg; {axsun, gaocong}@ntu.edu.sg*

*[2] Mazzia, Allie, and James Juett. ”Suggesting hashtags on twitter.” EECS 545m, Machine Learning, Computer Science and Engineering, University of Michigan (2009).*

*[3]**Mining tweets to predict their popularity, Robin Hahling, Kevin Gillerion*

*[4]. Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning:data mining, inference and prediction. The Mathematical Intelligencer 27, 2 (2005), 83{85.*

*[5.] Jurafsky, D., and James, H. Speech and Language*

*Processing An Introduction to Natural Language*

*[6] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD ’03. [10] D. Kempe, J. Kleinberg, and E ́. Tardos. Influential nodes in a diffusion model for social networks. In IN ICALP, 2005.*

*[7] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. KDD ’08.*

*[8] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. CoRR, 2010.*

*[9] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. ICDM’ 2010*

*[10] 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 Proceedings of the 13th international conference on World Wide Web, WWW ’11, 2011.*