**Predicting the Popularity of Newly Emerging Hashtags in Twitter**

**Introduction**

In recent days, twitter (a social networking), almost 100 million active users are tweeting about their thoughts/suggestions/helps/status in a short message of about 140 characters. Because of Twitter’s popularity and the viral nature of information dissemination on Twitter, predicting which Twitter topics will become popular in the near future becomes a task of considerable economic importance. Many Twitter topics are annotated by hashtags. we propose methods to predict the popularity of new hashtags on Twitter by formulating the problem as a classification task, probably the logistic regression model performs the best in terms of the Micro-F1 measure. We also observe that contextual features are more effective than content features.

Note: Twitter is a popular microblogging service that allows users to post short messages called “tweets.” Twitter also provides social networking features that allow users to follow other users, to retweet (or repost) their received tweets, and to reply to other users’ tweets. According to a Twitter blog post on March 21, 2012, more than 340 million tweets were posted daily by 140 million active Twitter users.

our work is the first to use both content and contextual features to predict the popularity of hashtags on a daily basis. This distinguishes our work from existing studies that merely consider one type of feature or predict hashtag popularity at much coarser time granularity

List of Content Features,

|  |  |
| --- | --- |
| Fc1 ContainingDigits | Binary attribute checking whether a hashtag contains digits |
| Fc2 SegWordNum | Number of segment words from a hashtag |
| Fc3 URLFrac | Fraction of tweets containing URL in Tt h |
| Fc4 SentimentVector | 3-Dimension vector: ratio of neutral, positive, and negative tweets in Tt h |
| Fc5 TopicVector | 20-Dimension topic distribution vector derived from Tt h using topic model |
| Fc6 HashtagClarity | Kullback–Leibler divergence of word distribution between Tt h and tweets collection T |
| Fc7 SegWordClarity | Kullback–Leibler divergence of word distribution between tweets containing any segment word in h and tweet collection T |

List of Contextual Features,

|  |  |  |
| --- | --- | --- |
| Fx1 | UserCount | Number of users Ut h |
| Fx2 | TweetsNum | Number of tweets Tt h |
| Fx3 | ReplyFrac | Fraction of tweets containing mention |
| Fx4 | RetweetFrac | @ Fraction of tweets containing RT |
| Fx5 | AveAuthority | Average authority of users in Gt h |
| Fx6 | TriangleFrac | Fraction of users forming triangles in Gt h |
| Fx7 | GraphDensity | Density of Gt h |
| Fx8 | ComponentRatio | Ratio between number of connected components and number of nodes in Gt h |
| Fx9 | AveEdgeStrength | Average edge weights in Gt h |
| Fx10 | BorderUserCount | Number of border users |
| Fx11 | ExposureVector | 15-Dimension vector of exposure probability P(k) |

**Algorithm and Methodology,**

Datasets we have in desk, is literally an unstructured data in a sequential format, therefore we formulate our problem as a classification task, so 7 content features and 11 contextual features extracted hashtag popularity prediction.

We use five commonly used classification models

1. Naïve Bayes
2. K-nearesh neighbour
3. Decision Tree
4. Support Vector Machine
5. Logistic Regression + three baseline method.

Our experimental results show that contextual features are more effective than content features for the prediction task, and that LR and KNN outperform the other three classification models. We also conducted experiments to evaluate the effectiveness of the features for popularity prediction for hashtags that have been popular for the past 2 days instead of 1 day

**Problem we may face?**

All tweets received from a Twitter stream are partitioned into consecutive fixed-time intervals by their time stamps. The time interval could be an hour, a few hours, or a day, depending on the number of tweets received, as well as the time criticality of the prediction. We define the popularity of a hashtag h in time interval t to be the number of users who post at least one tweet annotated by h within the time interval t, and we denote this by Φt h . Given a new hashtag at time t, our task is to predict its popularity at time t + 1, or Φt h +1. Note that predicting the exact value of Φt h +1 is extremely difficult and is often not necessary. Therefore, we relax the problem and predict the range of its popularity

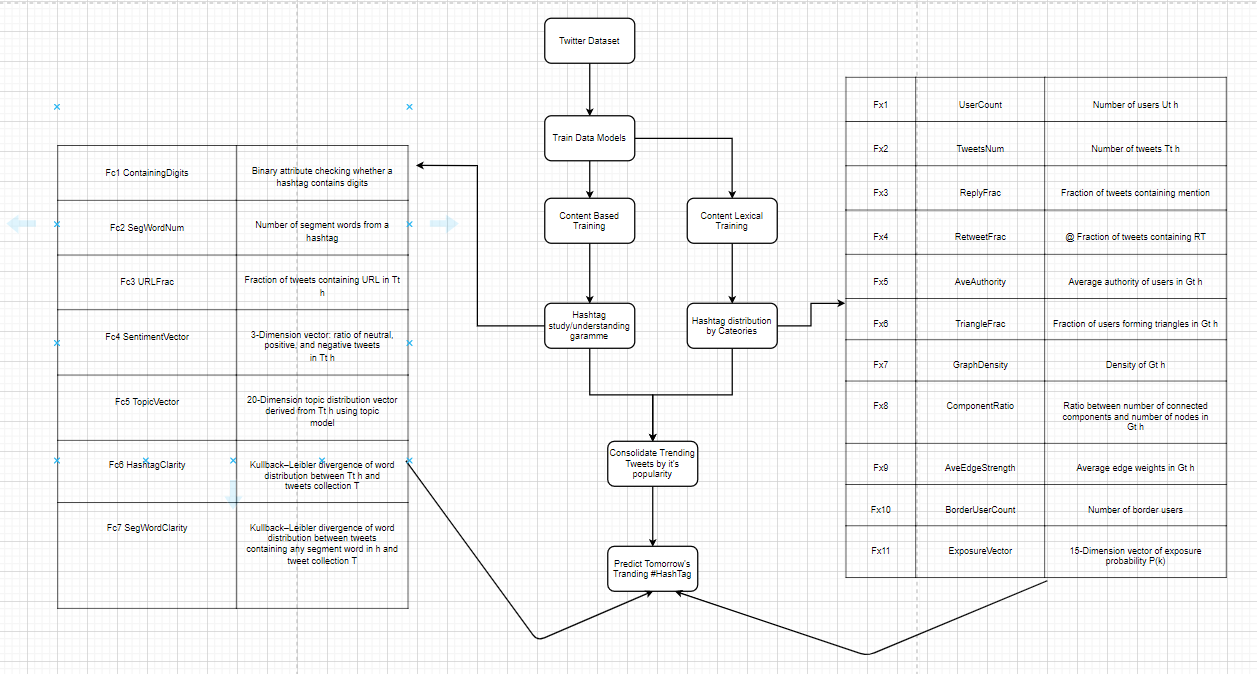
|  |  |
| --- | --- |
| [0, f] | not popular |
| [f, 2f], | marginally popular |
| [2f, 4f], | popular |
| [4f, 8f], | very popular |
| [8f, +•] | extremely popular |

Note that, depending on the number of tweets received from the Twitter stream and the requirements of a specific prediction application, a different number of ranges may be defined. The value of f controls the relative sizes of the ranges defined.

**Implementation:**

1. Train the available datasets which are available with us
2. Training should be done via CMLP – Common machine learning Platform via Logistic Regression and Classification
3. Other way, we can push these data into a DB which is supported with relational mapping.
4. Query of RM, should retrieve all data which has same #hashtag, which bring repeating under timestamp.
5. Each tweet record can be mapped with required popularity.

**Flow Diagram**



Comparing results of content based and contextual base results, and showcase how superior is contextual results ends when compare to content based results.

**Conclusion**

To predict the hashtag popularity of new topics on Twitter by formulating the problem as a classification task and evaluating three baseline methods and five classification methods. The main focus of our work was to identify and evaluate the effectiveness of content and contextual features derived from tweets annotated with candidate hashtags. Our experiments demonstrated that contextual features are more effective than content features.