Machine Learning Tells You How to Get More Retweets (Project Proposal)

Bowei Zhang
Courant Institute of Mathematical Sciences
New York University
Email: bz553@nyu.edu

Jingxin Zhu
Courant Institute of Mathematical Sciences
New York University
Email: jz1371@nyu.edu

I. Introduction

Twitter has become one of the most popular social media platforms. Its users nomally post millions of tweets every day [1]. Among this huge number of tweets, whether a new tweet will be attractive and influntial seems to deserve a careful study. Therefore, this project will build prediction models and analyze the possibility of a tweet attracting relatively more attention.

In this project, the number of retweets is treated as evaluation metric, measuing influence of a single tweet. Firstly, we will collect and preprocess raw tweets from Twitter. Then, various algorithms will be implemented to build machine learning models, which predict the possibality of a tweet obtaining a lot of retweets. The last part will be a brief comparison of different algorithms.

II. Data

The data utilized to train models is Twitter text messages, namely tweets, from celebrities on Twitter website. The reason of choosing their tweets is that those accounts tweet more frequently, have more followers, and more importantly, their tweets usually get large numbers of retweets.

The tools used for crawling data include tweepy and twitaholic. Tweepy is a Python language wrapper for official Twitter API [2]. It enables users to make function calls to retrieve tweets from designated users. Twitaholic is a third-party website, where it lists top 1000 users who have the most followers. These users are celebrities whom the program retrieves tweets from. The project team plans to get 100 most recent tweets from each of them and therefore form a dataset of 100,000 tweets.

The crawled tweets are in JavaScript Object Notation (JSON) format. Each one of them has various fields of information and they will be used to generate feature vectors. Among all of them, the number of retweets is used as the label for prediction. Initially, only a subset of fields are utilized to generate vectors, including number of followers, number of favorites, length of the tweets, does it include hashtags, etc. Other features such as text content and/or timestamp will be used in later experiments.

III. Methodology

Machine learning methodology is utilized to find a predictive model. Different algorithms will be tested using evaluation methods and the candidate with lowest training error will be chosen as the optimal one.

A. Algorithms

Following algorithms will be implemented to find the optimal model:

1) classification
    a) linear SVM
    b) SVM with kernel
    c) naive Bayes classifier
2) clustering
    a) K-means

If time allows, following algorithms will also be experimented: PCA and SVM, recursive partitioning trees, and regression models.

The label of data point will vary based on different algorithms. For example, when performing regression algorithms, the original label (number of retweets) will be used whereas when utilizing classification algorithms, a threshold will be decided and any label greater than the threshold will be marked as 1, otherwise -1.

B. Evaluation

Before plugging the data into learning model, 30% of it will be held out as test set. It will not be used until the very end to evaluate the model. All the training tasks will be performed on the rest 70% of the dataset.

For supervised algorithms, cross validation will be performed in order to select the parameters and choose an optimal model. Also, feature vectors will be tuned in order to achieve better performance. During the experiments, some features may be treated as noise and removed while others may be introduced.

Other evaluation may be come up in later stage of the project.

C. Tools

The program will mainly be developed in Python language. Following are supplementary tools that will be utilized for specific purposes:

1) tweepy: Twitter API for collecting data
2) twitaholic: website listing top 1000 Twitter users
3) scikit-learn: library of machine learning algorithms
4) NLTK: library used to process content of tweets

IV. Expected Results

After implementing different algorithms, key factors making a tweet popular will be found. They will be applied to predict whether a new tweet would attract large number of retweets. The team expects to see that a tweet with a photo, more hashtags, explicitly requesting retweets will probably receive more retweets.

Meanwhile, various algorithms involved in this course are covered, which provides us an opportunity to implement them to solve problems in real life.

V. Timeline

1) Mar. 27, finish project proposal.
2) Before Apr. 15, generate datasets of different features.
3) Before Apr. 22, implement algorithms on datasets.
4) Before Apr. 27, build and compare models.
5) Before May 4 , finish draft of project report.
6) Before May 12, further experiment and finish poster.

References

[1] Twitter. (2013) New tweets per second record, and how! [Online]. Available: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

[2] Tweepy. (2013) An easy-to-use python library for accessing the twitter api. [Online]. Available: https://www.tweepy.org