Report: Social Trends Prediction

Yuqing Zhu

April, 2021

1 Problem

We aim to find the social trends that are potential to go viral.

2 Method

Given a list of hashtags related to a specific field, e.g., *#bitcoin* and *#ether* in the cryptocurrency world, we first retrieve all the related tweets T in the period from the given start date sd to the given end date ed. We can then obtain a much smaller space of tweets regarding only the tweets we need in the specific domain to fetch the potential trends. Let $I(t)$ be the influence of tweet t, e.g., $I(t)$ = [retweets, likes, replies] of the tweet t. If $I(t)$ satisfies our criteria, i.e., when $I(t) \geq$[10,10,2] (a set of constants), we treat t as an influential tweet and the corresponding user u_t as an influential user. Then, we retrieve u_t's past N (set as 100) tweets and use its historic average influence $\tilde{I}(u_t)$ in the past N tweets to serve as a predictor for the future influence. Meanwhile, in the content of past N tweets, we record the hashtags h over the period from sd to ed and record u_t's maximal influence over each h (denoted as $I(h)$). Next, we use $p(h) = \tilde{I}(u_t) - I(h)$ to be u_t's potential influence on hashtag h, where $I(h)$ is u_t's current influence on hashtag h. In the end, we accumulate all the influencers' potential influence on the retrieved hashtags.

Algorithm 1: Potential Influence

Input: StartDate sd, EndDate ed, Hashtag list hl;
Output: Potential Hashtags and their potential influence in the future $\{PH|\{h : p(h)\}\}$;

```
Retrieve the tweets set T based on hl and [sd, ed];
foreach tweet t ∈ T do
    if I(t) satisfies the criteria ∧ Ĩ(u_t) = 0 then
        Retrieve u_t's past N tweets T_N;
        foreach t_N ∈ T_N do
            Initialize a set of hashtags H ← ∅ and I(·);
            if t_N.time ∈ [sd, ed] then
                foreach mentioned hashtag h in t_N do
                    I(h) ← max(I(h), I(t_N));
                    Push h into H;
        Record u_t's average influence over T_N as Ĩ(u_t);
        foreach h ∈ H do
            p(h) ← p(h) + Ĩ(u_t) − I(h);
return {PH|{h : p(h)}};
```

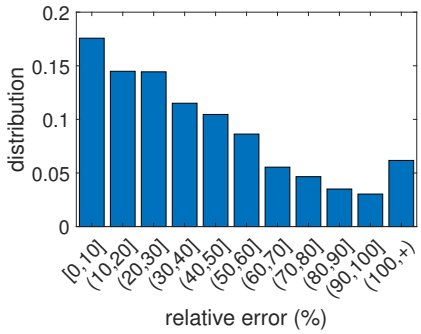

Figure 1: Distribution of relative error.

3 Estimator Evaluation

Note that we use the past N tweets' average influence for each influencer as an estimator, i.e., $[N \rightarrow \text{now}]$ tweets. Let U be the retrieved influential users. For each influencer $u \in U$, to evaluate the quality of the estimator $\tilde{I}(u)$, we get the recent $[2N \rightarrow N]$ tweets and evaluate the average influence on $[2N \rightarrow N]$ tweets over the recent $[N \rightarrow \text{now}]$ tweets. In other words, we use the average influence on $[2N \rightarrow N]$ tweets as an estimator of the average influence on $[N \rightarrow \text{now}]$ tweets.

In total, when $hl = [\#bitcoin, \#ether]$, we get 640280 related tweets from 2021-04-14 to 2020-04-20. Among them, 2177 are influential tweets satisfying our set criteria and 2177 influential users are analyzed. For these influences U, we get $\tilde{I}(u)$ on past $[N \rightarrow \text{now}]$ tweets and $\tilde{I}(u)'$ on past $[2N \rightarrow N]$ tweets. The relative error of $\big(\tilde{I}(u)' - \tilde{I}(u)\big)/\tilde{I}(u)$ is reported in Figure 1. Most influencers' influence are relatively stable with the relative error smaller than 50%.

4 Two Strategies

Let $hl = [\#bitcoin, \#ether]$. By Algorithm 1, we use $\tilde{I}(u)$ to predict u's influence on each hashtag h that he mentioned. Meanwhile, $I(h)$ represents u's current influence on hashtag h. If we accumulate $I(h)$ and $\tilde{I}(u)$ on each hashtag over all the influencers, we can get the current spread of h, denoted as $S(h)$ and our prediction on h, denoted as $\tilde{S}(h)$. In addition, we also record the number of influencers $N(h)$ that help to propagate the specific hashtags. To verify the two strategies of **Mean Reversion** and **Trend Following**, we perform the prediction from 2021-04-07 to 2021-04-13 and check the trend from 2021-04-14 to 2020-04-20. A total number of 478957 tweets are crawled from 2021-04-07 to 2021-04-13 with 1582 influential ones and 640280 are crawled from from 2021-04-14 to 2020-04-20 with 2177 influential ones. Note that the influential users we retrieve are different on these two time windows as the influencers are retrieved depending on whether their tweets in the specific time window are counted as

influential tweets. Let $S(h)_1$ and $S(h)_2$ be the influence spread of a hashtag propagating in the first time window of 2021-04-07 to 2021-04-13 and in the second time window of 2021-04-14 to 2021-04-20 respectively.

4.1 Mean Reversion

Among the set PH returned by Algorithm 1 in the first time window of 2021-04-07 to 2021-04-13, when $\tilde{S}(h) > S(h)_1$ and $N(h) \geq 5$, the hashtag h is considered to potentially revert to the mean spread and we treat the hashtag as a candidate **Mean Reversion** hashtag. We detect a total number of **490** candidate hashtags from 2021-04-07 to 2021-04-13. In the second time window, if we detect a hashtag previously thought as a candidate **Mean Reversion** hashtag with the current influence over the second time window greater than 10% of the predicted value in the first time window, i.e., $S(h)_2 > 0.1 \cdot \tilde{S}(h)$, we consider it to be a **Mean Reversion** hashtag. Overall, we find **449** such trends. In all, this strategy has a **91.6%**(449/490) accuracy.

4.2 Trend Following

When $1.1 \cdot \tilde{S}(h) \leq S(h)_1$ and $N(h) \geq 5$, i.e., current spread of the hashtag h is 10% higher than the prediction, we count h as a hashtag that is already viral and is likely to become a **Trend Following** hashtag. We find a total number of **93** candidate hashtags from 2021-04-07 to 2021-04-13. Next, if we find that the current influence over the second time window of a candidate **Trend Following** hashtag is greater than the predicted value in the first, i.e., $S(h)_2 > \tilde{S}(h)$, we count it to be a successful **Trend Following** trend. A total of **63** such trends are found from 2021-04-14 to 2020-04-20. In all, this strategy has a **64.3%** (63/93) accuracy.