Twitter Food Popularity

Pietro Morichetti, Thomas Axel Deponte and Victor Plesco

All Authors took part in problem statement, solution design, solution development, data collection and writing.

January 16, 2020

1 Problem Statement

The goal of this project is to build a tool for predicting how popular will be a tweet about food.
A "tweet" is simply a post, or message, with whom users interact on Twitter, a micro-blogging and social networking service. Like Facebook's status updates, you can share media-rich links, images, and videos in a tweet as long as you keep it at 280 characters or less. Any post on Twitter is considered a tweet, but the way someone tweets can be broken down into different types, such as Regular tweet (plain text), Image tweet, Video tweet, Media-rich link tweet, Location tweet, Mention tweet, Retweet[1] and Poll tweet.
In this work we evaluate the popularity of a tweet in terms of its final retweets number, which represents the degree of diffusion of the tweet within the social network. The given input is the post's resharing history after a certain time, which includes the number of retweets, alongside their timestamp (or the time it took for the person to reshare the post), and the number of followers each retweeter has; the output is the prediction of the final number of reshares.
Our solution is well described in SEISMIC [1] paper with some exceptions we introduced to improve the final prediction.

2 Assessment and performance indexes

Since the SEISMIC model has already been compared, within the paper, with other estimators, we compare our implementation with the one proposed in the original paper, using as evaluation metrics the Absolute Percentage Error (APE) of the estimated retweet's count. We choose this metric because it's a valid instrument for comparing forecasts across different time series as long as they contain strictly positive and far from zero values.

[1] Is a repost of another Twitter user's tweet on your own profile. Currently, you can see someone's else tweets only if you're following him, i.e, you are friends on the social network.

3 Proposed solution

The SEISMIC is a statistical model that can predict the retweets number from the network topology information and from the tweet infectiousness. The core of the model is the memory kernel Φ_t that represents the probability distribution [2] of a tweet to be viewed by a follower during time, also called reaction time distribution. This distribution changes over different social networks, (e.g. Facebook, Instagram [1]). We train it once on various tweets.

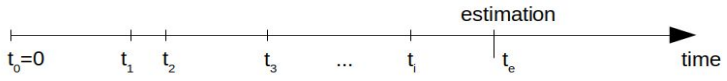

Figure 1: Tweet time line. The "*i*" represents the retweet number of the to the original tweet $i = 0$. We perform the estimation at t_e.

The second important element of the model is the infectiousness p_t, a function over time that defines the probability of a tweet to be reshared if someone sees it. This function varies among all tweets. Infectiousness of a post may depend on an intricate combination of factors, including the quality of the post's content, current time of the day, poster's geographical location as well as to the social interests and many others.

The estimation of the final reshares per tweet is done by the formula (8) [1]. It is the convergence of the numerical series that represents the summation of reshares during time among generations[2]. What influences a lot the estimation is for sure the Φ_t since it compares in the definition of N_T^e [3]. We noticed a strong assumption made by the authors of the paper [1] when training Φ_t: "Assume all retweets come from immediate followers. Under this assumption, the reaction time […] is the same as the relative retweet time" (Section 5.2 [1]). We think this limits the performance of the model cause we should consider only the retweets coming from verified immediate followers. In Section 4.2 we show the SEISMIC authors previous assumption was wrong and the dataset for training the Φ_t was wrong. This lead to a poor estimation.

4 Experimental evaluation

4.1 Data

Our data is composed of almost 272k tweets on Twitter from December 21 to December 28, 2019. The connection to the food topic has been done by the means of several popular food tags[4]. For each retweet, the data provides information on the original tweet id, original post time, and number of followers of the retweeter.

[2] In SEISMIC paper [1] the term "generation" is used to identify a set of retweets posted from immediate followers of a certain tweet user.

[3] It is the effective cumulative degree of resharers by time t, the full definition can be found in SEISMIC paper [1].

[4] In particular we used: *food*, *coffee*, *delicious*, *hungry*, *breakfast*, *recipe*, *dinner*, *beer*, *lunch* and *cooking* as anchor tags gathering tweets.

We focus on a subset of tweets with at least 50 retweets and no more than 100[5], so that our model enables the prediction as soon as sufficient number of retweets occurs. Another sampling parameter regards the followers number of the creator[6], which in our case is maximum 75000[7]. We form the training set using only the retweets made by the followers of the creator[8]. There are 239 tweets satisfying this criterion in the first 15 days. The following retweet cascades will be estimated by the SEISMIC algorithm based on the parameters derived from the training set.
The structures of the data-sets used for the implementation of the SEISMIC algorithm is so defined:

1. A *master* data-set containing the *user id*, *status id*, *original post time*, *followers* and number of aggregated *retweets* for the original tweets;
2. A series of *slave* data-sets constructed as follows:

12110205043640	*12405231999792*	*1214437830657*
...	...	...
...	...	...

Each column head represents the *status id* of an original tweet from the *master* data-set. The rows, limited to 100, represent the retweets of the *status id* in the column header and include a specific characteristic of these last. Three different data-sets with this structure were created:

- *IDs* – containing the *status id* of the retweets.
- *Followers* – containing the followers of those who retweeted.
- *Time* – containing the reaction time, expressed in seconds, to the original post.

In order to derive the first generation cascade, an ulterior *slave* dataset has been created, having, as in the previous cases, the column header's name connected to the *status id* column in the *master* data-set. The rows, limited to 75000 this time, represent the *user IDs* of the creator's followers. We define the first generation cascade as the retweeters belonging to this pool.
The connection between the *master* and *slave* data-sets is made by merging the row numbers from the *master* with the column numbers from the *slave*.

[5] The *rtweet* library enables the collection of the 100 most recent retweets of a given status, which may mislead our reaction time to the post distribution, since oldest (fastest) retweets may not be included in this sample.
[6] We refer to the *creators* as to the users who created the original tweets, whose popularity we are analyzing.
[7] The *rtweet* library enables the collection of no more than 75000 user IDs in a single call, every 15 minutes. Defining a sample with users with less than 75000 followers has enabled us into using more than one single Twitter Apps for downloading the data.
[8] Deriving the information about where the retweet comes from is time consuming. It implies the knowledge of the pool of followers of everyone who retweeted the original post. Therefore, we decided to concentrate only on the first generations of reshares, represented by the users who belong to the followers of the creator of the original tweet.

4.2 Procedure

Next, we describe the fitting of Φ_t. First we fit the memory kernel Φ_t with the training set. We take 239 sub-critical[9] tweets and we consider the retweets (~7000) coming from the first generation of followers. We fit the reaction time distribution with the same form presented in $Eq.\ 9\ [1]$ in SEISMIC paper: constant in the first 140 minutes followed by a power-law decay. After setting the constant period s_0 to 140 minutes, we estimate the power law decay parameter $\theta = -0.948$ with the complementary cumulative distribution function (ccdf), and choose $c = 1.33 \times 10^{-3}$ to make the Φ_t's integral equal to 1 on its domain. The memory kernel is a network wide parameter, hence it only needs to be estimated once. The fitted memory kernel is plotted in Figure 2.

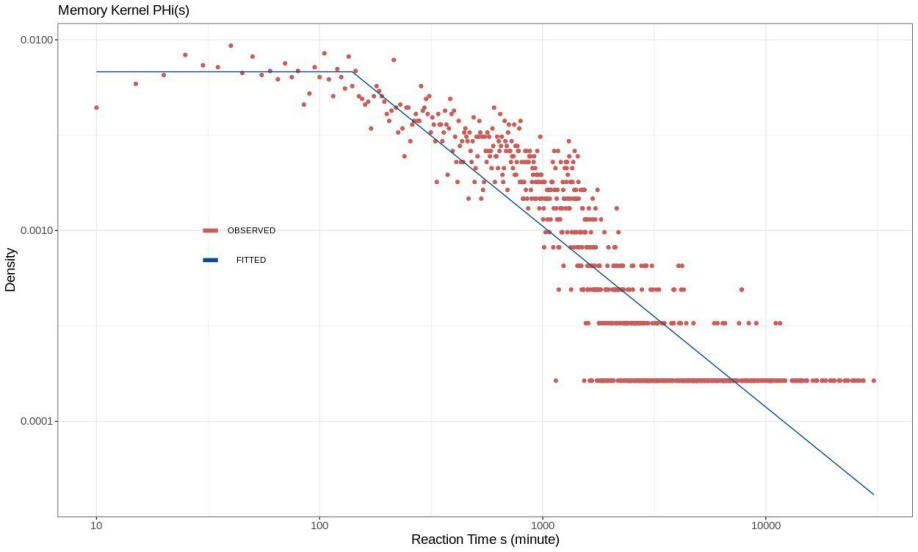

Figure 2: Plot of observed reaction time distribution and estimated memory kernel Φ_t. The reaction time is plotted on a log scale, hence a quasi-linear trend in the plot suggests a power law decay in the distribution.

4.3 Results and discussion

To empirically evaluate the performance of the SEISMIC method we plot the aggregated prediction of the retweets as a function of time (Figure 3). We denote that as the resharing history is being enriched SEISMIC is able to fastly adjust its prediction, which converges to the real final reshares at five hours of available data.

[9] A tweet is subcritical if its infectiousness parameter, after an initial explosion, starts a monotonic decay.

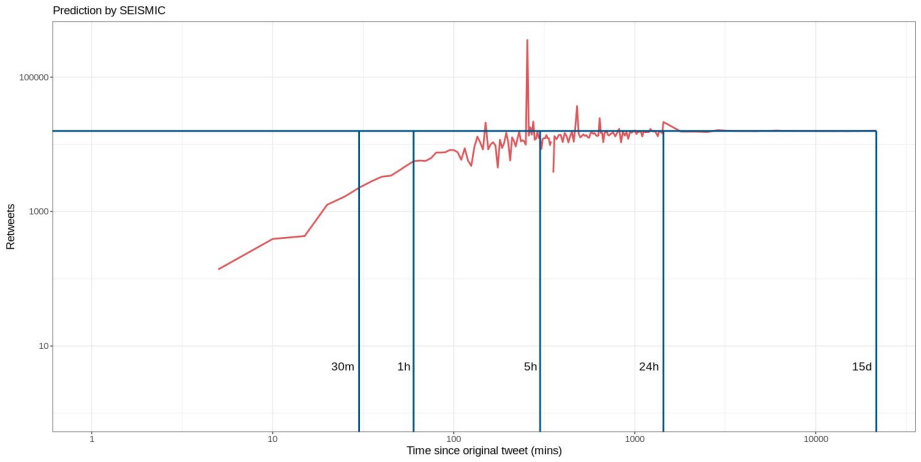

Figure 3: Predictions of the tweet's final retweet count as a function of time, on a log x-axis and y-axis scale. SEISMIC quickly finds an accurate estimate of the final retweet count represented by the horizontal blue line.

We run the SEISMIC method for each tweet and compute the *Absolute Percentage Error* metric as function of time. We plot quantiles of the distributions of APE in Figure 4. After observing the cascade for 10 minutes (t = 10min), the 90th, 75th and 50th percentiles of APE are less than 21.92%, 15.20% and 7.89%, respectively. This means that after 10 minutes, the average error is less than 7.89% for 50% of the tweets and less than 21.92% for 90% of them. After 100 minutes the error becomes unstable for the next 15 hours, while still maintaining its value around 10% for 50% the tweets and broking the 100% at the highest percentiles. The decreases until 1000+ minutes after what it maintains its value almost constant for all the percentiles.

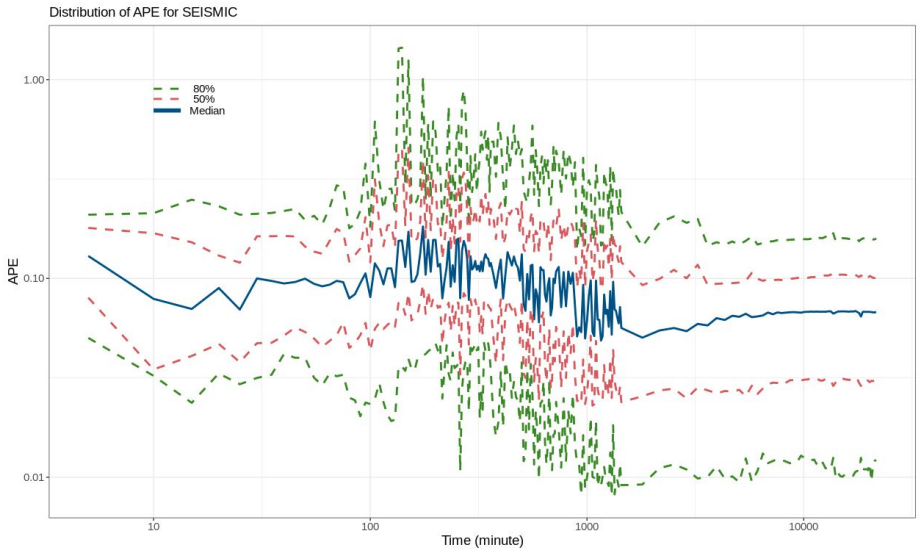

Figure 4: Absolute Percentage Error (APE) of SEISMIC. We plot the median and the middle 50%, 80% and the median percentiles of the distribution of APE across the tweets.

In this project we successfully implemented the SEISMIC model in R. It have been proven to be a useful tool for predicting the popularity of a tweet. Furthermore it provides a theoretical framework for explaining temporal patterns of information cascades [1]. Given a set of tweets and retweets is possible to measure their infectiousness, and thus analyze on which topics the social interest is oriented. In a future work could be interesting to find a predictor of the infectiousness, in order to estimate the popularity before tweet publishment.

5 References

[1], *SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity,* Qingyuan Zhao, Murat A.Erdogdu, Hera Y. He, Anand Rajaraman, Jure Leskovec, KDD '15.

[2], *Robust dynamic classes revealed by measuring the response function of a social system,* R. Crane and D. Sornette, Proceedings of the National Academy of Sciences, 105(41), 2008.

[3], Package 'rtweet':
https://cran.r-project.org/web/packages/rtweet/rtweet.pdf