

# Hashtag Popularity Prediction

**Instructor: Dr. Huan Liu**

**Arizona State University**

[huanliu@asu.edu](mailto:huanliu@asu.edu)

**Mukthadir Choudhury**

**Arizona State University**

[mhchoudh@asu.edu](mailto:mhchoudh@asu.edu)

**1207597049**

**Pooja Thakur**

**Arizona State University**

[pthakur2@asu.edu](mailto:pthakur2@asu.edu)

**1207664649**

## Abstract

In this paper, we attempt to tackle the problem of predicting hashtag popularity, i.e. for the given dataset of tweets, we aim to find out how many times the hashtags contained in them will appear in the near future. In our proposal, we look at tweet data of some days and try to find out the number of times the hashtags appearing in them will re-occur in the next hour. These are the hashtags that are trending in the present and have a possibility of being used again by people of the world. We develop a Hashtag Prediction Weight Model (HPWM), which is simple in nature but efficiently takes into account the past frequency of the hashtags with time, and uses it to determine the frequency of hashtags in the coming hour for a Twitter dataset of more than 10 million tweets. We then evaluate our algorithm by dividing the original dataset into partitions, and find the root mean square error to calculate its accuracy.

**Keywords:** twitter, tweets, hashtag, data, timestamp, frequency.

## 1. Introduction

These days, almost every person in the world is using social media and interacts with his/her friends through some social networking sites like Twitter, facebook etc. Research shows that more than 50 percent of online adults now use at least two social media sites [1, 2]. As of September 2014, 23% of online adults used Twitter [2, 3]. Social media is a platform where one can express his/her thoughts, ideas and let the rest of the world know what they are feeling. Twitter is such a social networking site where people share the ideas, feelings, thoughts with their followers. Tweets also contain a lot of hashtags which generally represent what the tweet is about. These can be considered as the most important words of the tweet. Using these hashtags, we can very well understand the theme of the tweets that contain them.

However, hash tag prediction is different from normal texts classification. This is because the tags have unique features because they change so frequently that it is hard to identify the number of clusters. This hinders the process of classification when new tags can come out at any time. Thus it is important to have some idea of the way in which hashtags are gaining popularity, to in turn gain information on trending tweets. The content of this paper will assist researchers who are studying the navigation of upcoming hypertext structures, and the engineers who are seeking to enhance the navigability of social tagging systems. This can be seen in the current research being carried out in determination of topics from hypertext content [4] and to find out new trends on topics based on the existing tags [5, 6]. The idea behind our paper is to use the existing data to extract information of the hashtags in tweets relative to the timestamps i.e. to analyze how hashtags are trending with time. We use this data to calculate the frequency with which these hashtags will appear in the next hour. For this, we see which tag has occurred in which timestamp and also in which hour.

In this paper, we have tried to come up with a solution where, by utilising the recently used hashtags in tweets, we will be able to predict which hashtags will be used in the next one hour and thus will be able to predict the theme of the tweets which will allow us to know what topics are trending. Our paper is divided in the following sections: Section 2 deals with defining the problem statement. In Section 3 we present related work. In Section 4 we describe our Tweet corpus. In Section 5, we introduce our Hashtag Prediction Weight Model (HPWM) approach to predict hashtag episode frequency. In Section 6, we present experimental results and error, which is the deviation from the actual values. Section 7 concludes the paper and Section 8 presents the future work.

## **2. Motivation**

In today's world, competition is very high. From classrooms to multi national companies, everyone is competing with each other to secure the top position. For a successful business model, the most important thing to know is to know what the customer needs and what are the customer's expectations. And customer needs keep on changing time to time. To know what the customers need at a particular time, it then becomes imperative to know what is flowing through their mind. A clue to that can be obtained by finding what is trending on Twitter. But the businesses need to be ahead of the customers, to ensure that they make the required products at the required times. This would lead to increase of trust of the customers on their dealers. Thus stems the need to predict what will be in the minds of the users and to model their processes accordingly.

For example, say we have a football match in the next two hours and suddenly football lovers find out that their favorite heroes favorite is a red hat. So, everyone would like to come to the football ground with a red hat to show their support for their team. In this case, say, a firm called GetItNow.com is able to provide them the required red hat just outside the football stadium, then not only would the sales of GetItNow.com's would increase at that moment but also its customer's trust for them would also increase and soon this single business model can make them super rich in very months of time. Hence, hashtag prediction can lead to serious benefits if executed properly.

## **3. Problem Statement**

Given set of tweets,  $T = t_1, t_2, t_3, \dots, t_n$  where  $n$  is number of tweets in the given dataset. Let  $H$  be the set of hashtags contained in the tweets. We find the timestamps of the tweet, which are extracted from the json data. Our problem is one to find out those hashtags that will be seen the most in the next one hour.

For this we compute a Weight Value for each hashtag. We use the hour of the hashtag and assign it a logical weight to find this weight value for each hashtag, over the last 20 days. This acts as a measure to predict the number of times that hashtag will re-occur in the next hour.

## **4. Related Work**

There are more than 271 million active users in Twitter per month [6]. And more than 500 million tweets are generated on a daily basis[6]. This vast amount of information enables a large number of research organizations and universities to obtain this data and perform rigorous analysis to predict something that would lead to breakthrough and boom their economy. And the area of information diffusion has indeed been well studied though most of the work focused on study of diffusion through social networking and information websites, e.g. [7, 8, 9, 10, 11, 12]. Most of the papers analysing Twitter datasets have focused on understanding the properties of the tweets and predicting the information diffusion [11, 12, 13, 14]. People have also studied different approaches to predict the popularity of a particular hashtag at a particular instant of time in [14].

We have also found out that diffusion related studies generally fall into two different categories. One of those includes analysing social networks as graphs of connected nodes. Here, users are considered as nodes and their

friendships as edges and then how information is diffused from one node to many is analyzed. Arruda et. al [15] proposed that network metrics plays a vital role in finding out the most influential nodes in a network and from where the information starts to spread [6]. In their paper, they have examined the effect of few centrality measures on epidemic models (like SIR) and concluded that some centrality measures such as closeness centrality is strongly correlated with the result of the spreading rumors model [6].

The next category includes study of the diffusion model through content analysis by utilizing natural language processing. For example, a study predicted that some specific set of words is very highly likely to be present in few viral tweets. Li et al. analyzed few tweets and concluded that highly interactive tweets tend to contain more negative emotions than normal tweets [6], [15].

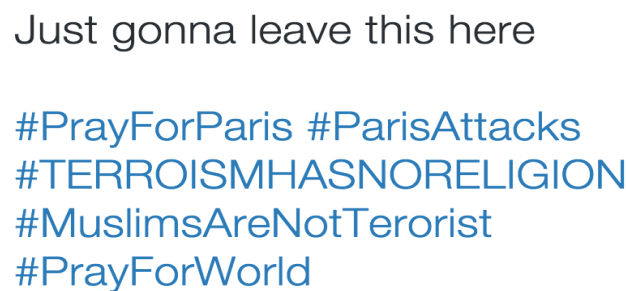
In another study, Romero, Meeder and Klienbergl [16] observed that different topical categories of hashtags had different pattern of propagation of hashtags. They added that some hashtags are more ‘persistent’ than others meaning were appearing more in a particular place rather than appearing everywhere.

Our work builds upon some of these. The main difference is that we are interested in finding the hashtags that would be used the most in the next one hour unlike the above works where the authors are trying to find out which of the hashtags and tweets are popular in a particular area or with a particular set of people. Thus we are use a weight value model to find out which of the hashtags would be most popular in the next one hour.

## 5. Proposed Method

### 5.1. Twitter data

Twitter allows the usage of two meta characters: @ marking a user name (e.g. @BarackObama), and # marking a hashtag: a sequence of non whitespace characters preceded by the hash character. Our dataset of more than 10 million tweets also contains these two kinds of keywords. The use of hashtags is a popular way to denote the context of a tweet [17]. For example: The hashtags #PrayForParis and #ParisAttacks give the background and ideology for the tweet:



Just gonna leave this here

#PrayForParis #ParisAttacks  
#TERROISMHASNORELIGION  
#MuslimsAreNotTerorist  
#PrayForWorld

**Figure 1: A Sample Tweet**

As is visible, #Paris is made use of as a hashtag and it can also be used as a critical part of the tweet.

### 5.2 Preprocessing

Since timestamps hold the most importance for our problem definition as well as its solution, we used the sampled data for the first 25 days out of the whole data. The reasons for this are manifold. We used the data of the first 25 days to predict results for the 26th day. Thus data had to be sequential, as we aim to show how tweets on a new day depend on the days continuously before it. Therefore it was not logical to take data from the first few days and use it to hypothesize values for the last few days. Similarly, we could not correlate data from the first few days along with data later in the dataset for finding future values. We could also have taken data from some other part of the dataset

to start off. However, as Albert Einstein said, “Most of the fundamental ideas of science are essentially simple”, we focused on finding an efficient solution and not a complex one.

Since there are no constraints on the structure of the hashtag, analysis of its content poses some technical problems. A hashtag can be a lexical word (#paris), or a compound of lexical words (#parisattacks) [17]. Users also use variations of the same hashtag content like #parisattacks, #ParisAttacks, #Parisattacks, #PARISATTACKS and #PARISattacks. Although semantically identical, it has been found that their usage frequencies differ greatly. Thus we had to make the hashtags case-insensitive.

We also had to remove noise to clean the data. For this, we removed different characters such as [;, \$, %, ^, &, \*, +, }, >, ?] so that the content of the hashtags gains importance over its structure, and the syntax of the hashtags does not affect our results. Thus we obtained clean case insensitive data for the first 25 days.

### 5.3 Prediction Method and Algorithm

Here we propose the working of the model that we have used to predict the hashtags which will occur the most in the next one hour. We start with extracting information about the timestamps from the given json file which contains data of millions of tweets. Getting the maximum and the minimum timestamp gives an interval in which all the timestamps of the data lie. In the dataset, which we are using, there are tweets which span over twenty five days. To obtain the window of a single day, we divide the total time into 25 fragments. We now carefully parse the json dataset so as to obtain all the hashtags along with their corresponding timestamp.

Based on the windows obtained before, these hashtags are categorized. Hence, the dataset is divided in 25 different parts and the count of each hashtag in those parts is maintained. This count is called the hashtag volume of each fragment. Now an importance factor for each day is calculated by assignment of weights by using a logical probability distribution. These are obtained by using formula (1). Each hashtag volume is then multiplied by this importance factor to obtain the weighted hashtag volume, as shown in formula (2).

$$y = x * 0.02 \quad (1)$$

where x represents the day and y denotes the importance factor obtained for each day

$$z = v * y \quad (2)$$

where v is the hashtag volume, and z is the weighted hashtag volume for each day

These weighted hashtag volumes are iterated to obtain a summation called Total Weighted Hashtag Volume (TWHV) for each hashtag. Now this value is scaled using a scale factor to get the predicted frequency of occurrence in the next hour. To scale the value, we divided it by  $2^5 = 32$ . The reasons for choosing this scaling factor is twofold. The first aim here is obviously to reduce any error induced upon scaling. The second reason is to prevent the pattern of bits from changing. This emerges from the fact that dividing a number by a power of two is equivalent to shifting all the bits to the right once for each power of two [18]. These predicted frequencies are now sorted and saved in a csv file along with the corresponding hashtag.

For example, consider a hashtag named PrayForParis which appears 500 times in day 1, 300 times in day 2, 400 times in day 5, 200 times in day 6 and 300 times in day 10. Importance factor y is obtained as:

$$\begin{aligned} y(\text{day } 1) &= 0.02 \\ y(\text{day } 2) &= 0.04 \\ y(\text{day } 5) &= 0.10 \\ y(\text{day } 6) &= 0.12 \\ y(\text{day } 10) &= 0.20 \end{aligned}$$

Weighted hashtag volumes  $z$  are obtained as:

$$\begin{aligned}z(\text{day } 1) &= 500 * y_1 = 10 \\z(\text{day } 2) &= 300 * y_2 = 12 \\z(\text{day } 5) &= 400 * y_5 = 40 \\z(\text{day } 6) &= 200 * y_6 = 24 \\z(\text{day } 10) &= 300 * y_{10} = 60\end{aligned}$$

Total Weighted Hashtag Volume (TWHV) for the hashtag is:  $\Sigma z = 146$

The predicted frequency in the next hour is :  $z/32 = 4.56$

---

#### Algorithm 1

---

**Input:** given json file

**Output:** dictionary of unique hashtags and its frequency of occurrence in each day

```
1: parse input json file in proper format for processing
2: repeat until last tweet
3:     save the timestamp
4:     save the keywords
5:     repeat until last keyword in keywords
6:         check if keyword is hashtag
7:         if hashtag
8:             compare to find if hashtag is repeated or not
9:             if repeated
10:                 check the day hashtag was used
11:                 append the timestamp to the specific day value of key as hashtag in
dictionary
12:                 increment the hashtag count of that day by 1
13: return dictionary containing unique hashtags and its frequency of occurrence in each day
```

---

---

#### Algorithm 2

---

**Input:** dictionary of unique hashtags and its frequency of occurrence in each day

**Output:** list of hashtags with their popularity value

```
1: repeat until last key in dictionary (till last unique hashtag)
2:     save the value of key in an array
3:     repeat until last element in the array (till last day)
4:         calculate importance factor of the day
5:         calculate weighted hashtag volume of the day
6:         calculate total weighted hashtag volume of entire span
7:         calculate frequency (popularity) of hashtag
8: return list of hashtags with their frequency (popularity) value
```

---

## 6. Results and Findings

### 6.1 Results

On implementing our Hashtag Prediction Weight Model on the clean data we obtained, we analyzed tweets that spanned over 25 days to predict the hashtags that would be used the most in the 1st hour of the 26th day. And the hashtags that would be used the most according to our experiment were found to be:

**Table 1: Hashtag with its predicted frequency**

Hashtag	Predicted frequency of occurrence
occupywallstreet	10777
Ows	2701
P	486
occupywallst	268
Nypd	236
Tcot	210
Occupy	169
occupytogether	161
Occupyla	140
takewallstreet	134

After this process, it can be easily observed that the theme of most of the tweets during those 4 weeks was related to wallstreet.

### 6.2 Evaluation

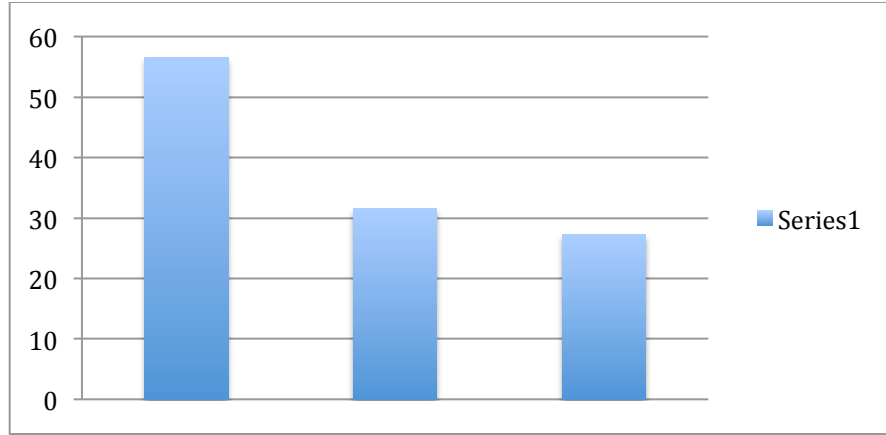
To evaluate how efficient and how accurate our algorithm is, we found the Root Mean Square Error (RMSE) for our predicted frequency by comparing it with the actual occurrence of hashtags for a specific duration of time.

For this, we decided to carry out our experiment on the dataset of (a) the first 10 days to predict the frequency of hashtags that would show up on the 1st hour of 11th day, (b) the first 15 days to predict the frequency on the 1st hour of 16th day and (c) the first 20 days to predict the frequency on the 1st hour of 21st day. Then, we compared the obtained values with the actual values in the dataset for the same time intervals to calculate the Root Mean Square Error for each pair of values.

The values of RMSE for the time intervals are as follows:

**Table 2: Error values corresponding to different data sizes**

Time Interval	RMSE value
1st hour of 11th Day	56.5857718851
1st hour of 16th Day	31.613494642
1st hour of 21st Day	<b>27.240159981</b>



**Figure 2: Graph of RMSE values**

From the above evaluation, we can deduce the following:

- Since the maximum occurrence of a particular hashtag is in thousands, our RMSE value shows that our algorithm is able to predict the frequency of occurrence of a hashtag with high accuracy.
- RMSE value keeps on decreasing with increase in analyzed data i.e. the more we increase the training data, the lower error our model produces. This is in line with the established theories of machine learning[19].
- Due to limitation of system resources, we were restricted to running our algorithm in a limited data set, which contained roughly 1 million tweets. However, we can conclude from the above observation about error reduction that when relatively large amount of twitter tweets are analyzed using our hashtag prediction weight model (HPWM), the model will be able to predict the hashtags that would be used in the near future with much more accuracy.

The following table shows the predicted hashtags versus the actual hashtags that were used the most on 1st hour of 21st day.

**Table 3: Comparison of Predicted and Actual Hashtags**

Predicted Hashtag (in decreasing order of occurrence)	Actual Hashtag (in decreasing order of occurrence)
occupywallstreet	occupywallstreet
ows	ows
p	p
nypd	tcot
takewallstreet	occupyla
occupywallst	occupy
usdor	occupytogether
occupyla	nypd
occupytogether	globalrevolution
anonymous	anonymous

We can see that in Table 3, 7 out of hashtags are same which means that the algorithm predicts the future hashtags with high accuracy. Thus it is clear that our algorithm is able to find out accurately the hashtags that would be used the most in the next one hour.

### **6.3 Runtime analysis:**

Our algorithm analyzes the given 1gb of data within 3 minutes and predicts the results in less than 5 minutes. Thus, we can say that our algorithm will be able to analyze bigger data sets within 10-12 minutes and provide results with reasonably high accuracy.

## **7. Conclusion**

In this paper, we proposed a very simple yet powerful algorithm to predict the hashtags that would be popular in the near future all over the world, and also their predicted frequency. We also evaluated the validity of our algorithm by running it on three different known sets of data and then comparing the obtained values with actual data. We found that our algorithm was able to predict the hashtags with fairly high accuracy by comparing the top few hashtag values in both types of data. We further proved the accuracy of our findings by calculating the root mean square error value, which was found to be fairly less. The RMSE value was also found to be decreasing with increase in the amount of data analyzed. With this we concluded that our algorithm will be able to perform even better with increase in amount of training data.

## **8. Future Work**

Right now, our algorithm efficiently predicts the hashtags that would be used the most in the next one hour. Our algorithm can be easily extended to find the hashtags that would become popular in the next day or the next week and so on. Our algorithm can also be coupled with linear regression model to increase the accuracy of the predicted hashtags. To find more accurate predictions for large intervals of time like weeks and months, use of location and retweet information can also be made.

## **9. Acknowledgements**

We would like to thank Professor Dr. Huan Liu for teaching us such a great course and also allowing us to work on the projects which helped us learn cutting edge technologies and more about how large scale social data is analyzed for patterns and predictions. We would also like to thank Fred Morstatter and Ghazaleh Beigi for all the help throughout the semester. Last but not the least, we would also like to thank Stackoverflow community where we found solutions to most of our issues.

## **References**

- [1] <http://www.huffingtonpost.com/kyle-mccarthy/five-facts-about-social-media-usage-today/>
- [2] <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>
- [3] <http://www.doit.wisc.edu/news/social-media-usage-today/>
- [4] Murfi H and Obermayer K, A Two-Level Learning Hierarchy of Concept Based Keyword Extraction for Tag Recommendation, ECML PKDD Discovery Challenge, (497): pp. 201–214, 2009.



- [5] Bundschuh M, Yu S, Tresp V, Rettinger A, Dejori D, and Kriegel HP, Hierarchical Bayesian Models for Collaborative Tagging Systems, ICDM, pp. 728–733, 2009.
- [6] Heymann P, Ramage D, and Garcia-Molina H, Social Tag Prediction, SIGIR, pp. 531–538, 2008.
- [7] Hasan Davulcu, Sultan Alzahrani, Saud Alashri, Anvesh Reddy Koppela, and Ismail Toroslu. A Network-Based Model for Predicting Hashtag Breakouts in Twitter.
- [8] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10.
- [9] 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.
- [11] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. KDD '08.
- [12] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. CoRR, 2010.
- [13] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. ICDM' 2010.
- [14] B. A. Huberman, D. M. Romero, and F. Wu. Social Networks that Matter: Twitter Under the Microscope. Social Science Research Network Working Paper Series, Dec. 2008.
- [14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? WWW '10.
- [15] Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted?. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 925–936. International World Wide Web Conferences Steering Committee 2014
- [16] 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.
- [17] What's in a Hashtag? Content based Prediction of the Spread of Ideas in Microblogging Communities Oren Tsur Ari Rappoport
- [18] [https://en.wikipedia.org/wiki/Scale\\_factor\\_\(computer\\_science\)](https://en.wikipedia.org/wiki/Scale_factor_(computer_science))
- [19] <http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models/>