Social Media Data and its availability in Research

Debjyoti Paul
University of Utah
deb@cs.utah.edu

OVERVIEW

The term *Social Media* gained widespread attention with the advent of Web 2.0 in the first decade of 20th century [14]. Web 2.0 is also known as *participative* or *social web* that emphasize on user interaction and user generated content encouraging participatory culture. Before we jump into more details of social media, it would be wiser to define it. The dynamic changes and continuous development of social media services makes it harder to define them, however most of the research work could be summarized it as follows.

Definition 1 (Social Media). Social media is an interactive computer mediated technological platform that facilitates the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks [15].

In contrast to the *traditional media* which operates under a monologic transmission model i.e. one source to many receivers, such as a television, newspaper or a radio station which broadcasts the same programs to an entire city; *social media* are dialogic transmission system which brings interaction, usability and a notion of individual entity in the digital world.

1 PART A

Social Media Data in Numbers

Marketing and social media experts broadly agree to classify social media with respect to media type and its usage i.e *blogs, social networks, private messaging, microblogs, photo sharing, video sharing, professional networks, enterprise social networks, forums, products/services review, social bookmarking, social gaming, collaborative projects and virtual worlds* [1]. We now present a list of relevant social media platforms according to the classification stated in Table 1.

Table 1: List of Relevant Social Media

Category	Social media sites with link
Social Networks	Facebook, Snapchat, WeChat, Quora
Private Messaging	Messenger, Whatsapp, QQ, WeChat, Skype
Microblogs	Twitter, Sina Weibo, Tumblr
Photo Sharing	Instagram, Photobucket, Flickr
Video Sharing	Youtube, Vimeo, Dailymotion
Professional Networks	LinkedIn, AngelList, Meetup
Enterprise Social Networks	Workday
Blogs	Wordpress, Medium, Buffer Blog
Forums	Reddit, Hacker News, Quora
Products/Services Review	Yelp, Foursquare, Google Places
Social Bookmarking	Pinterest, Digg, Stumble Upon Mix
Social Gaming	Pokemon Go, IGN, Gamespot [20]
Collaborative Projects	Slack, Invision, Trello, Github, Bitbucket
Social Gaming	Friendster

University of Utah, Salt Lake City, USA.

Table 2: Social media sites and number of users (in millions).

		Years						
Category	**Site**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Type**
Social Networks	Facebook	1228	1393	1591	1860	2129	2271	Total
	WeChat	355	500	697	889	989	1082	Total
Microblogs	Twitter	241	284	305	318	330	332	Active
	Weibo	140	175	237	310	340	392	Active
	Tumblr	175	–	–	–	460	550	Total
Photo Sharing	Instagram	150	300	460	600	870	1000	Active
	Snapchat	33	100	180	301	–	400	Total
Video	Youtube	700	1100	1431	1618	1767	1900	Active
Professional	LinkedIn	277	347	414	467	530	576	Total
Services	Yelp	96	135	150	158	170	178	Active
	Foursquare	33	30	50	–	–	55	Total
	Ridesharing	–	–	208	272	338	400	Active
Bookmarking	Pinterest	–	–	110	160	220	250	Total

Table 3: Social media sites and media units created per day (in millions).

		Years						
Category	**Site**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Unit**
Social	Facebook	3600	–	–	4320	–	–	posts
Microblogs	Twitter	245	399	500	–	657	682	tweets
	Tumblr	120*	205*	270*	315*	380*	448*	total blogs*
Photo Sharing	Instagram	5	31	–	–	–	67	photos
	Snapchat	–	–	–	–	760	3000	photos
Video	Youtube	69,120†	103,680†	432,000†	576,000†	720,000†	–	hours video †
Services	Yelp	40‡	55‡	75‡	95‡	135‡	171‡	total reviews‡
	Foursquare	33	–	–	–	–	12000§	total checkins§
Bookmarking	Pinterest	–	5	13	–	–	–	pins

The popularity of a social media site is primarily determined by the total number of users or monthly active users. Table 2 presents the facts about social media sites user base which gives some sense of its popularity [7, 8, 18, 21]. The attribute *type* with values *(a) Total (b) Active* represents whether the statistic is of total users or active monthly users respectively.

Other than the social media sites mentioned in Table 2 there are some significant sites where only the current user statistics are available. For example Flickr, the photo sharing platform has 90 million users. Quora, a question answer social platform has 300 million users worldwide. Reddit, a social forum has 330 million active users.

Number of users is not just an important parameter to measure the popularity of a social media site but also to estimate the amount of data storage it maintains. Another feature that will help us to estimate data storage is the amount of media units (e.g. posts, photos, microblogs, videos etc.) ingested per day. Table 3 presents all the statistics of the relevant social media from open internet [7, 8, 21, 24]. The statistics for social media sites missing in Table 3 but mentioned in Table 2 are almost impossible to find in open internet.

Social Media Storage Estimate:

Social media sites seldom reveals the amount of data they store or ingest on a daily basis. Also the ever growing social media makes

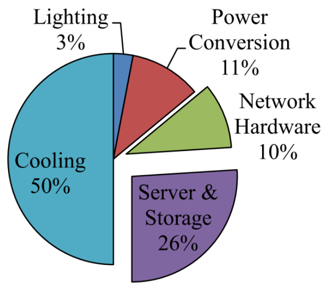


Figure 1: A typical breakdown of energy usage among components in data center [12].

it hard to estimate their storage capacity. I present few methods in the following section to estimate the social media storage.

1. Storage space estimate from media units: This method works for all the social media sites mentioned in Table 3 where the approximate storage space required by a media unit is known.

Youtube: Let us take an example of a Youtube video data. From Table 3, we find that in the year 2017 users uploaded 720,000 hours of video on Youtube per day. First, assuming the fact that Youtube fairly stores most of the videos in 1080p and also stores them in multiple resolution such as 240p, 360p, 720p and format e.g. Webm, flv, mp4, 3gp, mp3. We can determine the amount of storage space needed for a 1 minute video [13].

$$\begin{aligned} &27.71 \text{ MB (Webm)} + 17.00 \text{ MB (flv)} + 554.43 \text{ KB (3gp)} \\ &+ 45.80 \text{ MB (mp4)} + 2.81 \text{ MB (mp3)} \\ &= 93.8614355 \text{ MB} \end{aligned} \tag{1}$$

From the above Equation 1, we find that $720,000 \times 60 \times 93.8614355 \approx 4.055$ petabytes (PB) of storage space is required by Youtube everyday. We can also calculate the total amount of storage space ingested during the period of 2013 to 2017 from Table 3 by utilizing area under the curve method with interpolation. The above method reckons 3096.17 PB or 3.096 exabytes (EB) of storage. Considering videos before 2013 and new 4K video which takes more space, it can be easily assumed that Youtube uses 10-15 EB of storage space.

Twitter: Similar to the method above, we can find the space required to store a tweet. A tweet is stored in Twitter as UTF-8 format. This takes 140 characters tweets at most 560 bytes of space. However the metadata attached with a tweet is much more than the tweet itself. I personally did a random sample experiment of 100K tweets store in our databases to find the average storage space for tweet JSON object obtained from streaming API. I found that one JSON tweet object takes 3247 bytes of space in average. 682 million tweets per day will require around 2.2145 terabytes of data per day. Using the interpolation method for area under the curve, we can find that Twitter uses 3.13 petabyte of space for storing the tweet alone. It is also worth noting that 42% of tweets contains images [23]. If we assume the average image size be 100 KB then we will see $(100 * 1024)/3247 * 42\% \approx 13.2$ times increase in storage space requirement.

2. Storage space estimate from data center power usage: This section presents an approximate method to estimate space capacity of large social media companies like Facebook and Google. A typical breakdown of energy consumption by data center is given in Figure 1. The largest energy consuming component is cooling infrastructure with 50% of total energy. Rest of the energy is used by power conversion, lighting, network and server components [6, 12]. Facebook data centers use efficient data center architecture and hardware tweaks saves 8-12% of energy spent in cooling, 13-25% in power conversion, 10% in motherboard [11]. That implies at most 11% more efficient than typical data centers. Hence, it can be claimed that Facebook servers use 37% of energy. Considering Facebook's 138 MW Altoona data center equipped with 200 Watts servers each with 6×4 TB of HDD as used in their experiment for [11]. Assuming the data-center is running at peak energy $(138 \times 0.37 \times 24)/200 = 6127200$ TB = 6.1272 exabytes (EB). Taking all the data centers in consideration and diving them with replication factor we can estimate the storage capacity of Facebook. The analysis provided above supports news *Facebook Builds Exabyte Data Centers for Cold Storage* in 2013 [5].

Social Media Data for Researchers:

Regardless of the vast data in social media sites, the dataset available for researchers in public domain is very very limited. Also these datasets size are miniscule in comparison to what we mean by bigdata. The only exception is Twitter. Twitter provides $\tilde{1}\%$ of sample tweets through its streaming API. By utilizing multiple resources and some other APIs such as keyword search researchers can obtain more than 1% sample data. Also it is noteworthy to note that researchers looking for geotagged data faces a greater challenge as only 0.85% of tweets in Twitter are geotagged [19]. A study on the sample tweets and original stream (firehose) reveals that the research on sample and original can differ unless proper coverage is taken care of during data collection strategy [16]. From the previous analysis and checking our twitter streaming collection, we can estimate that $\tilde{1}\%$ sample collects 25-30 GB of uncompressed data daily.

Facebook has tightened the security and restricted access to many of its data for public research after Cambridge Analytical Scandal [4, 10]. However, Facebook launched an initiative to make a dataset available to *"The Social Science Research Council"* for assessment on impact of social data on elections [9]. That means only affiliated researchers with certain agencies will be able to access Facebook's data. I believe we will continue to see a restricted access behavior from similar social media sites in future which can affect public researchers.

To sum up, I present some of the most relevant social media dataset available for public research in Table 4. From Table 4, it is clear that there is no relationship between the amount of data social media sites possesses and the data available for researchers in public domain.

Many social media sites expose APIs for developers to access data. The free APIs of all the relevant social media sides are very restrictive. For example, facebook allows 200 API requests per hours/user. Instagram earlier had 5000 requests/hour which has been reduced to 200 request/hour. Geolocation services like foursquare 500 requests/hour on premium API end points. Hence, it is clear that

Table 4: Most relevant social media dataset.

Site	Dataset	Size	Link
Network Repository	Frienster	8 GB	link
	Twitter (1)	6 GB	link
	Twitter (2)	6 GB	link
	Twitter (3)	960 MB	link
	Orkut (1)	388 MB	link
	Orkut (2)	422 MB	link
	Sina Weibo	960 MB	link
Stanford SNAP	Facebook (ego)	4,039 nodes	link
	Google Plus	107,614 nodes	link
	Twitter Social	81,306 nodes	link
	Expinion	75,879 nodes	link
	Youtube	1,134,890 nodes	link
	Amazon Product	334,863 nodes	link
	Reddit	132,308 submissions	link
	Flickr	2,316,948 images	link
	BrightKite (Location)	58,228 Nodes	link
	Gowalla (Location)	196,591 Nodes	link
	Movies	196,591 Nodes	link
Social Computing ASU	Youtube (1)	1,138,499 nodes	link
	Youtube (2)	15088 nodes	link
	Last FM	108,493 nodes	link
	Twitter	11,316,811 tweets	link
	Flickr	80,513 nodes	link
	Foursquare	106,218 nodes	link
	Digg	116,893 nodes	link
	Delicious	103,144 nodes	link
Sentiment 140	Twitter Sentiment	160,000 tweets	link
Reddit	Reddit	1.7 billion comments	link
Yahoo	Flickr	100 million images	link
Awesome Data Github	Google Scholar	Unknown	link
	Indie Map	Unknown	link

the availability of social media data in public domain is not only subjected to efforts we invest in collecting it but also restrictive policies from companies. We will revisit about *scraping challenges* in the next section.

2 PART B.

Scraping Social Media:

Social media data is broadly divided into [3] :
1. Historic datasets: Previously accumulated and stored social/news, financial and economic data.
2. Realtime feeds: Live data feeds from streamed social media, news services, financial exchanges, telecom services, GPS devices and speech.

Historical datasets are relatively easy to collect than real-time feeds because of API limitation and limitation of scraping via crawling webpages. Social media data is mainly collected via two procedure API based or web crawling based approach. API crawling methods are easy to maintain and modifiable. Web crawling based approach can extract more information which might not be available via APIs. Also web crawling can avoid API rate limit and can crawl more data. But it needs a data cleaning procedure and a high maintenance because the web interface can change quickly which would result in a change of code and crawling procedures.

Open source projects on API libraries are available on Github and other collaborative platforms that enables researchers to collect data e.g. tweepy for Twitter, pyFacebook, python-flickr, foursquare-api, uberpy, python-vimeo, youtube-api etc. (all available in Github.) A comprehensive list of api wrappers can be obtained from here [2].

Due to the imposed restriction of APIs, crawling technolgies have evolved rapidly. Few state-of-the-art crawling libraries and utilities are scrapy [1], beatifulSoup4 [2], selenium [3], Graphene [4].

Difficulties in Scraping:

In this section, we present a comprehensive list of the challenges researchers face while scraping social media data.

Lack of free APIs: Many social media sites does not expose relevant APIs for free. They are mostly paid APIs and are costly.

API call limit: API call limit is the most discouraging element for scraping social media data. Sometimes they can even show an erratic behavior even if the scraper has not reached the documented rate limit e.g. foursquare APIs.

Javascript enabled data: Researchers resorting to web crawling face this challenge when the data is fetched via javascript calls and simple web crawling will not work. This kind of scenario needs expertise and technologies. Using libraries like scrapy with plugin splash (scriptable browser) [5] is the only best method known to researchers. But it slows down the scraping process.

IP block: Web crawling often result in IP block. Only a well resourced researcher can avoid it by charging IP addresses often through proxy services. But sometimes social media sites can even detect proxies and limit access.

Legal Terms and Conditions: Web crawling sometimes violate the legal terms and conditions from social media sites. Even if a researcher gets the data by crawling he/she might not be able to publish the data or share with other researchers.

Missing link metadata: It is often found that the data exposed either in API or in Webpages miss link metadata. Link metadata are information like followers which reduce usability of such data.

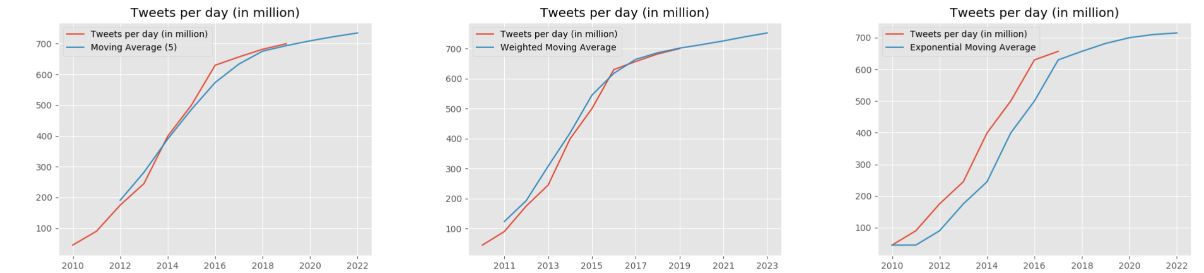
3 PART C.

Predicting future:

As mentioned in Section 1 that there is no true evident relation between the amount of social media data available with companies and data available for public research; except for Twitter. Hence, we will try to predict how much social media data will be available in 2022 within the Twitter. Also the fact that I have limited amount of information from Table 2 and Table 3, it will be wise to start with a simple forecasting methods. Also it is very common for extremely simple methods like forecast with historical average to outperform more complex methods [22]. This statement is even more likely true for short time series.

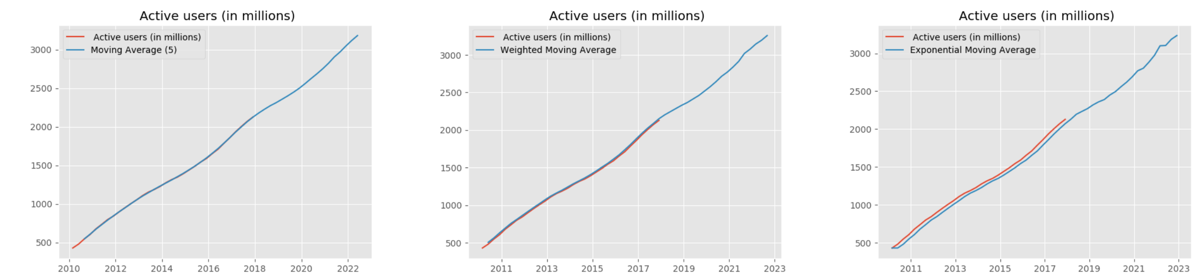
To start with the basic forecasting model we might consider historical *mean model* which assumes that the time series consists of independently and identically distributed ("i.i.d.") values, as if

[1] github.com/scrapy/scrapy
[2] github.com/il-vladislav/BeautifulSoup4
[3] pypi.org/project/selenium/
[4] graphene-python.org/
[5] github.com/scrapy-plugins/scrapy-splash



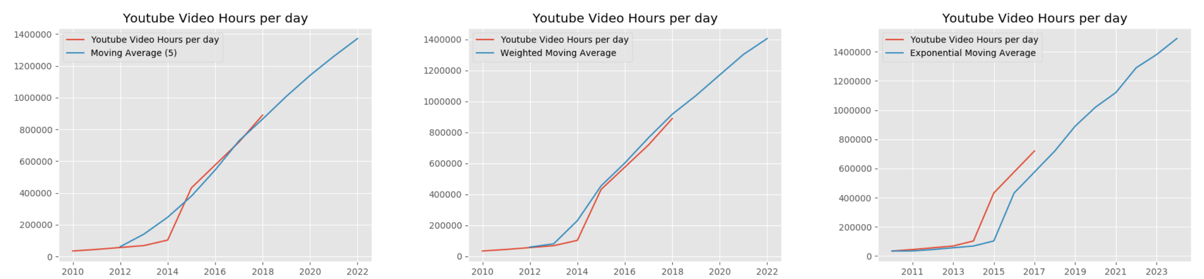
(a) Moving Average with window 5. **(b) Weighted Moving Average with window 3 (0.5, 0.3, 0.2).** **(c) Exponential Moving Average with $\alpha = 0.9$.**

Figure 2: Predicting tweets per day from 2019-2022.



(a) Moving Average with window 5. **(b) Weighted Moving Average with window 3 (0.5, 0.3, 0.2).** **(c) Exponential Moving Average with $\alpha = 0.9$.**

Figure 3: Predicting Facebook monthly average users from 2019-2022.



(a) Moving Average with window 5. **(b) Weighted Moving Average with window 4 (0.5, 0.35, 0.05,0.05).** **(c) Exponential Moving Average with $\alpha = 1.0$.**

Figure 4: Predicting Youtube video upload hours from 2019-2022.

each observation is randomly drawn from the same population. Under this assumption, the next value should be predicted to be equal to the historical sample mean if the goal is to minimize mean squared error. I tried a bunch of experiment with *mean model* and it appeared to be working well with relatively moderate mean squared error.

I avoided linear trend model as it is not a very *robust* model for time-series forecasting [17]. Since I have the priori knowledge of the series that has a positive trend or zero trend I can use a moving average model that puts more weight on the most recent values than to use a linear trend model with a *not very significant* trend estimate. With this notion I tried the *moving average model* to forecast the tweets per day metric for 2022. I tried more forecasting methods with *weighted moving average* and *exponential moving average model* . Figure 2 presents all the forecasts for tweets per day metric till 2022 [6].

Complex models like *ARMA (AutoRegressive Moving Average)* and *ARIMA (AutoRegressive Integrated Moving Average)* models did not do well (validating with small train test data).

From the prediction in Figure 2, we find that twitter will generate around 730 million tweets every day in average. That is almost 2.4 terabytes of tweet data without media.

[6] github.com/debjyoti385/MovingAverageForecasting

Similar forecasting applied on Facebook monthly active users in Figure 3 reveals that almost 3000 million people will be active in Facebook by the end of 2022. From Youtube's hourly video upload data, it is predicted that in 2022 users will upload 1.4 million hours of video everyday 4.

REFERENCES

[1] T. Aichner and F. Jacob. Measuring the degree of corporate social media use. *International Journal of Market Research*, 57(2):257–276, 2015.
[2] APIWrappers. List of api wrappers, 2018. https://github.com/realpython/list-of-python-api-wrappers, [Online; accessed 27-Nov-2018].
[3] B. Batrinca and P. C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116, 2015.
[4] BBC. Cambridge analytica facebook scandal, 2018. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html, [Online; accessed 27-Nov-2018].
[5] DataCenterKnowledge. Data center knowledge, facebook builds new exabyte data center, 2018. https://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage , Online; accessed 27-Nov-2018.
[6] M. Dayarathna, Y. Wen, and R. Fan. Data center energy consumption modeling: A survey. *IEEE Communications Surveys & Tutorials*, 18(1):732–794, 2016.
[7] Domo. Domo, 2018. [Online; accessed 27-Nov-2018].
[8] ExpandedRamblings. Expanded ramblings, 2018. https://expandedramblings.com,[Online; accessed 27-Nov-2018].
[9] FacebookNewsroom. Facebook launches new initiative to help scholars assess social mediafis impact on elections, 2018. https://newsroom.fb.com/news/2018/04/new-elections-initiative/, [Online; accessed 27-Nov-2018].
[10] FacebookNewsroom. An update on our plans to restrict data access on facebook, 2018. https://newsroom.fb.com/news/2018/04/restricting-data-access/, [Online; accessed 27-Nov-2018].
[11] E. Frachtenberg, A. Heydari, H. Li, A. Michael, J. Na, A. Nisbet, and P. Sarti. High-efficiency server design. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, page 27. ACM, 2011.
[12] I.-T. R. Group et al. Top 10 energy-saving tips for a greener data center. *Info-Tech Research Group. fi…*, 2007.
[13] jdownloader. Youtube jdownloader, 2018. http://i.imgur.com/CgX5t.png, Online; accessed 27-Nov-2018.
[14] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.
[15] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.
[16] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.
[17] R. Nau. Review of basic statistics and the simplest forecasting model: the sample mean. *Fuqua School of Business, Duke University*, 2014.
[18] NetImperative. Sina weibo overtakes twitter user numbers, 2018. http://www.netimperative.com/2017/05/chinas-weibo-overtakes-twitter-user-numbers/, [Online; accessed 27-Nov-2018].
[19] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):1–11, 2013.
[20] Statista. Most popular games in us, 2018. https://www.statista.com/statistics/580150/most-popular-us-gaming-apps-ranked-by-audience/, [Online; accessed 27-Nov-2018].
[21] Statista. Statista, 2018. https://www.statista.com,[Online; accessed 27-Nov-2018].
[22] StatsStackExchange. Is it unusual for the mean to outperform arima, 2018. https://stats.stackexchange.com/questions/124955/is-it-unusual-for-the-mean-to-outperform-arima, [Online; accessed 27-Nov-2018].
[23] Thenextweb. Next web, 2018. https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/,[Online; accessed 27-Nov-2018].
[24] Zephoria. Facebook statistics, 2018. [Online; accessed 27-Nov-2018].