Fall 2013

Determining spread of diseases using Social Networking data

Jiten P. Oswal
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the [Computer Sciences Commons](#)

Recommended Citation

Oswal, Jiten P., "Determining spread of diseases using Social Networking data" (2013). *Master's Projects*. 335.

DOI: <https://doi.org/10.31979/etd.tke5-h7kh>
https://scholarworks.sjsu.edu/etd_projects/335

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Determining spread of diseases using Social Networking data

A Thesis
Presented to
The Faculty of the Department of Computer Science
San José State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
Jiten P Oswal
December 2013

© 2013

Jiten P Oswal

ALL RIGHTS RESERVED

SAN JOSE STATE UNIVERSITY

The Designated Thesis Committee Approves the Thesis Titled

Determining the spread of diseases using social networking data

by
Jiten P Oswal

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

December 2013

Dr. T. Y. Lin, Department of Computer Science	Date
---	------

Dr. Robert Chun, Department of Computer Science	Date
---	------

Grinish Engineer, Senior Software Engineer, Cisco	Date
---	------

ABSTRACT

“Hi, in a weeks, there will be high possibility you may get infected by flu. Please go to take a flu shot.” – This is a sample warning to some users in Twitter about the spreading of flu to people from specific location. Twitter may already be used to plan social lives, interact with celebrities and communicate with friends. But now data from the social networking sites could have a far more serious use like tracking diseases and learning about their spread. Using this kind of data, I will help Health Agencies to personalize the prediction and warn the general public about spreading of certain communicable disease so that any individual can take appropriate actions to avoid getting infected from such disease.

ACKNOWLEDGEMENTS

I am thankful to each and every person who has contributed his plethora for the completion of this project. Working on this project was unique experience. Knowledge and experience gained from this project will remain with me as an ingratiating memory.

I would like to express my sincere appreciation to my project advisor Dr. T. Y. Lin for suggesting this topic. Also, his guidance, cooperation and time in helping me out carry out the research to success. Thanks for the benevolent support and kind attention. I would also like to thank my committee members Dr. Chun and Grinish for their support and patience.

I would also like to thank our department for providing us with the necessary software required in our project. I'm also thankful to the library for providing necessary books and materials required to learn different concepts for our project.

Last but not the least, sincere thanks to my parents, family and friends for constant moral support, inspiration, encouragement and blessings without which project completion would have been next to impossible.

Table of contents

1. Introduction.....	9
2. Literature Review.....	12
2.1 Current Method.....	12
2.2 My Approach.....	13
2.3 Data Flow.....	14
2.3 Clustering.....	16
2.3 Prediction using curve fitting function technique.....	26
3. Verification.....	36
4. Limitation.....	37
5. Future Work.....	38
6. Conclusion.....	38
7. Reference.....	39

List of Tables

Table 1: Real-time relevant data collected from Twitter.....	16
Table 2: Real-time relevant data collected from Twitter after aggregation.....	17
Table 3: Historical data collected by scraping health agency websites.....	19
Table 4: Average of all 10yrs of historical data region-wise.....	25
Table 5: Polynomial Regression results for example dataset with order 2.....	29
Table 6: Polynomial Regression results for historical data with order 10 (North Cal).....	32
Table 7: Polynomial Regression results for historical data with order 10 (South Cal).....	34

List of Figures

Figure 1: Historical data representation of San Francisco (2003-2004).....	20
Figure 2: Historical data representation of San Jose (2003-2004).....	21
Figure 3: Historical data representation of Santa Clara (2003-2004).....	21
Figure 4: Historical data representation of Sunnyvale (2003-2004).....	22
Figure 5: Historical data representation of whole North Cal cluster (2003-2004).....	22
Figure 6: Historical data representation of whole South Cal cluster (2003-2004).....	23
Figure 7: Average of all 10 years of historical data for both North Cal and South Cal.....	25
Figure 8: Function plot we got as result of applying polynomial regression.....	30
Figure 9: Function plot by applying polynomial regression of order 10 (North Cal).....	32
Figure 10: Function plot by applying polynomial regression of order 10 (South Cal).....	35

Introduction

Twitter users send around 500 million tweets a day, an endless amount of information about how people feel, what they're doing, what they know and where they are, almost everything and anything.

For epidemiologists and public health officials, it's a potential gold mine of data, which can be used in so many ways. Like been able to track where disease is breaking out and how it spreads and how many people it is affecting and then how to best help them and prevent others in those affected places. This all day comes with a lot of noise and finding the right information from this endless gold mine is a big task for them.

The question is: How do you take these millions of messages, find the useful information from it and then get it to right people who can react to it?

This is a very big question and difficult one to tackle as well and so many researchers have always push away from the idea of using Twitter data, which they say is too messy and too uncontrolled compared with traditional methods of collecting health data, such as surveys and analyses of hospital visits. Some argue that, its about time we once learn how to effectively massage this data, Twitter's very messiness of data will be what makes it an invaluable resource.

Twitter currently use twitter to just tweet whatever, whenever. It's like a status board where people just broadcast their thoughts and wellbeing.

Using Twitter data has other advantages like it's faster: It can take the Centers for Disease Control and Prevention about two weeks to publish findings. Also, these numbers can additionally be delayed because these agencies get data only when a person goes to hospital or to some place where he gets reported with the agency database.

Twitter, on the other hand, might reflect it the first morning someone wakes up with a sore throat or any other kind of sickness. This property of twitter data being real-time is very advantages when want to track emerging diseases or viral fevers. So understanding how important it would be of being notified about a disease the instance it is been tweeted about would be so powerful.

Twitter also provides you additional metadata like location which most of the tweets have with it. That, coupled with faster data, could help keep hospitals and clinics from getting overwhelmed in the middle of an outbreak into getting their facilities straight and up to the mark everyone is treated and human life is saved. Also, getting detailed location-specific data would be useful in identifying clumps of non-communicable diseases like cardiovascular disease etc., and allowing health officials to focus education efforts in the areas that need it the most and spread awareness against it.

All these advantages, coupled with the fact that we are getting better at tracking and analyzing useful information from big data servers, means that we should soon start using this amount of data with all its merits to help our public health agencies to better

serve human life. To start with, in this project of mine I have modeled one of the many ways people could use social media as a complementary tool to better get us started using all this information.

2. Literature Review

In this section we discuss the current system of predicting spread of disease used by government and other health agencies.

Current Method

Government and these health agencies collect data from all the hospital on biweekly basis and from all this data and number of people infected lets out warning signals to general public when this number meets certain threshold.

- Collect data on biweekly basis from hospitals and other medical institutions
- From all this data from various institution calculating number of people infected by certain disease in a certain region/zone.
- Let out warning to general public when number hits certain threshold.

Limitation of this approach

- Missing out on daily trends as data is collected in bulk after every 2 weeks throughout the year.
- Also, the analysis is simply based on hitting the threshold curve; no real analysis is actually done.
- Grouping of data on biweekly basis is done zone-wise and hence key data points are missed.

- It doesn't determine speed or rate of disease spread and predict when it would reach a certain place. It just provides medical warning until certain area hits a threshold.

My approach

- Collect real-time region specific data from twitter w.r.t our disease tracking library. (Cache it on nightly basis or as required like hourly depends on infrastructure as well as our requirements)
- Clean all that data to remove all the noise and outlier cases. (Natural Language Processing = Sentimental Analysis + Subject Predicate Analysis + Context Analysis)
- Segmenting all this data region wise and analyzing it against historical data from biweekly government and health agency databases.
- Saving this analysis and utilizing it in next days'/weeks' prediction to get the rate of spread of disease (approximate number of people expected to suffer from the disease in next few days) in between two regions.
- As you would understand, I'm learning from real-time self-generated data so that with maturity we will get more accurate results. Also, the data would show the direction in which the disease is travelling and how soon would it reach.

Data Flow in my approach

- Twitter data is scraped for region specific tweets using twitter v1.1 api's.
- All these tweets are then parsed one by one through data cleaning algorithm where tweets are accepted or removed on basis of positive-negative notion, subject-predicate notion, interrogative analysis. All these cleaned accepted tweets are only considered for our analysis.
- After cleaning, segmentation of these live tweets is done where we aggregate all newly formed data region wise.
- Then from this newly formed data we get direction of spread for the disease causing infection.
- Now, for predicting number of people who will be affected we use polynomial regression where we will find 2 curved function which pass through both the historical data and the real-time clean data to predict the rate of spread of disease in coming days.

Analysis of Tweets for finding noise and outlier scenarios -

There are different checks that are required for verifying clean or dirty data in our context.

Following are the checks that I'm performing:

Check 1: Positive-Negative Notion check

This can be detected from the type of words used in a tweet. And then following cases exist:

1. Tweets with positive notion are discarded.
2. Tweets with negative notion are considered as person feeling sick or having flu should be sad.

Check 2: Subject-Predicate Notion check

This can be detected using Stanford library of words used to under first person, second person speech. And then following cases exist:

1. Tweets with first person context are only considered; as we only want an individual's health details neither his views nor know-hows of anyone else's health details.
2. Tweets with any other subject-predicate context are discarded to discard noise from our system.

Check 3: Interrogation check

This can be detected from the type of words used in a tweet. And then following cases exist:

1. Tweets with interrogative words are discarded.

Clustering –

Clustering of historical and real-time data is done by grouping data based on distance between cities to form regions.

Clustering of newly cleaned real-time data –

Real-time Data collected using twitter crawler and after removing all the outlier cases looks something like this –

City	Timestamp	Username	Disease
San Francisco, CA	Mon Nov 19 07:08:34	halezyay	flu%20%3A%28
San Francisco, CA	Mon Nov 19 05:48:31	AquaEK4i	flu%20%3A%28
San Francisco, CA	Mon Nov 19 03:28:31	SwoffordLegacy	flu%20%3A%28
San Francisco, CA	Mon Nov 19 01:13:45	MACHETEMAMBO	flu%20%3A%28
San Francisco, CA	Mon Nov 19 01:11:03	hitmanAdam	flu%20%3A%28
San Francisco, CA	Sun Nov 18 23:24:57	ESM_JonathanL	flu%20%3A%28
San Francisco, CA	Sun Nov 18 15:42:34	adeodhar	flu%20%3A%28
San Francisco, CA	Sun Nov 18 07:43:19	abjornsen	flu%20%3A%28
San Francisco, CA	Sun Nov 18 06:56:59	adylene1009	flu%20%3A%28
San Francisco, CA	Sun Nov 18 05:23:01	Michaelasvelez	flu%20%3A%28
San Francisco, CA	Mon Nov 19 13:02:40	proulx_cam	flu%20%3A%28

Table 1 – Real-time relevant data collected from Twitter

And after aggregating individual data points in above table we will finally get something like this.

City	Week	Disease	Total No. of Users
Los Angeles, CA	3 rd Nov – 10 th Nov	Flu ☹	6026
San Francisco, CA	26 th Oct – 3 rd Nov	Flu ☹	7267
Los Angeles, CA	10 th Nov – 17 th Nov	Flu ☹	6546
San Francisco, CA	3 rd Nov – 10 th Nov	Flu ☹	7765
San Francisco, CA	10 th Nov – 17 th Nov	Flu ☹	7610
Los Angeles, CA	17 th Nov – 23 rd Nov	Flu ☹	6179
San Francisco, CA	17 th Nov – 23 rd Nov	Flu ☹	8023

Table 2 – Real-time relevant data collected from Twitter after aggregation

Clustering of historical data –

Historical Data from health agencies – (collected using web-crawler parsing health agency websites) I have collected data from 2003-2013. Below is sample format of data from my database.

Date	SF	SJ	SC	Sunny vale	North Cal	LA	SD	LV	South Cal
2/12/12	3285	2609	2201	2686	10781	2968	2900	2199	8067
2/19/12	3893	2792	2166	2734	11585	3529	3205	2477	9211
2/26/12	3786	2575	2219	2576	11156	2840	2640	2495	7975
3/4/12	3643	2341	2406	2484	10874	2969	2315	2134	7418
3/11/12	4232	2674	2291	2525	11722	3096	2629	2662	8387
3/18/12	3932	2733	2376	2587	11628	2752	2148	2738	7638
3/25/12	3733	2829	2473	2496	11531	2734	1878	2464	7076
4/1/12	3395	2688	2453	2486	11022	2280	1875	2159	6314
4/8/12	2964	2337	2328	2497	10126	2112	1821	2116	6049
4/15/12	3042	2227	2209	2404	9882	1933	1788	2268	5989
4/22/12	2582	1965	2108	2328	8983	2130	1381	1891	5402
4/29/12	2364	1677	1941	2110	8092	1934	1233	1729	4896
5/6/12	2341	1682	1927	2073	8023	1954	1065	2083	5102
5/13/12	2418	1672	1826	1946	7862	2026	1150	1930	5106
5/20/12	2207	1757	1744	1902	7610	1832	1237	1807	4876
5/27/12	1974	1854	1648	1760	7236	1858	1567	1791	5216
6/3/12	1849	1629	1516	1596	6590	1655	1607	1500	4762
6/10/12	1783	1444	1458	1481	6166	1598	1391	1454	4443

6/17/12	1467	1176	1299	1399	5341	1499	1181	1284	3964
6/24/12	1369	1035	1221	1309	4934	1221	1019	1154	3394
7/1/12	1196	935	1172	1250	4553	1206	809	968	2983
7/8/12	1195	859	1130	1178	4362	1190	802	837	2829
7/15/12	1270	832	1083	1133	4318	1160	859	819	2838
7/22/12	1217	847	1002	1040	4106	1085	808	754	2647
7/29/12	1095	790	959	1010	3854	1026	818	760	2604
8/5/12	1065	790	911	964	3730	981	820	705	2506
8/12/12	1234	788	894	901	3817	1064	768	684	2516
8/19/12	1278	831	847	872	3828	1283	810	690	2783
8/26/12	1615	993	851	847	4306	1390	984	789	3163
9/2/12	1823	1147	858	821	4649	1613	1170	907	3690
9/9/12	2076	1394	890	801	5161	1698	1257	1075	4030
9/16/12	2290	1573	940	824	5627	1957	1240	1242	4439
9/23/12	2513	1532	993	855	5893	2005	1571	1417	4993
9/30/12	2679	1535	1034	931	6179	2391	1919	1406	5716
10/7/12	2836	1782	1209	1087	6914	2696	2157	1327	6180
10/14/12	3002	2103	1301	1305	7711	2759	2357	1468	6584
10/21/12	3181	2019	1503	1502	8205	2929	2083	1809	6821
10/28/12	3551	2158	1563	1958	9230	3158	2226	1980	7364
11/4/12	3142	2232	1601	1975	8950	3098	2167	1901	7166
11/11/12	3786	2488	1683	1912	9869	3667	2621	2255	8543
11/18/12	3424	2375	1727	1943	9469	3540	2765	2200	8505
11/25/12	3651	2585	1871	2032	10139	3940	3064	2998	10002
12/2/12	3712	2872	1980	2056	10620	4174	2809	3429	10412
12/9/12	3509	2609	2060	2059	10237	3904	3220	2926	10050
12/16/12	3519	2673	2388	2168	10748	3901	3673	3973	11547
12/23/12	4961	4037	2461	2334	13793	5370	6288	5073	16731
12/30/12	6673	5500	3269	2655	18097	7490	8434	6613	22537
1/6/13	10207	7859	5350	4187	27603	9540	10487	8220	28247
1/13/13	15027	10424	10530	8794	44775	12759	13997	9419	36175
1/20/13	16039	13317	14285	11375	55016	18474	19053	12882	50409
1/27/13	12094	11223	12163	8766	44246	14647	15885	11141	41673
2/3/13	9482	8540	9332	7224	34578	9877	10641	7190	27708
2/10/13	7459	7133	7103	6077	27772	5873	7229	4926	18028
2/17/13	6578	4833	5994	5206	22611	5768	6345	4447	16560
2/24/13	5454	3769	4682	4660	18565	5290	5280	3357	13927
3/3/13	4540	3279	3318	3863	15000	4510	4530	3098	12138
3/10/13	3745	3438	2458	3334	12975	3581	3172	2522	9275

3/17/13	3542	3306	1910	2963	11721	2562	3030	2230	7822
3/24/13	3292	2935	1571	2765	10563	2466	2586	2207	7259
3/31/13	2994	2568	1252	2330	9144	2117	2121	2145	6383
4/7/13	2677	1978	989	2031	7675	2016	2258	2078	6352
4/14/13	2258	1892	800	1939	6889	1901	1853	1644	5398
4/21/13	2401	1971	694	1837	6903	1671	1688	1605	4964
4/28/13	2090	1612	613	1732	6047	1643	1709	1477	4829
5/5/13	1933	1562	532	1505	5532	1499	1663	1435	4597
5/12/13	1671	1434	474	1375	4954	1314	1338	1448	4100
5/19/13	1607	1190	448	1291	4536	1390	1250	1343	3983
5/26/13	1531	1106	419	1271	4327	1247	1121	1313	3681
6/2/13	1317	1033	387	1257	3994	1241	1151	1278	3670
6/9/13	1321	1007	368	1254	3950	1162	1215	1118	3495
6/16/13	1372	1075	355	1281	4083	1216	1216	992	3424
6/23/13	1447	1035	337	1246	4065	1157	1181	960	3298
6/30/13	1171	967	317	1219	3674	1027	961	932	2920
7/7/13	1150	856	313	1203	3522	1106	964	929	2999
7/14/13	1069	840	314	1227	3450	1028	854	889	2771
7/21/13	1240	902	320	1229	3691	1007	1100	932	3039
7/28/13	1122	1371	456	1621	4570	1176	1181	1061	3418
8/4/13	1019	1091	450	1502	4062	1133	1023	1021	3177
8/11/13	1442	1018	465	1492	4417	1049	1214	1078	3341
8/18/13	1409	799	484	1382	4074	1015	1593	1256	3864
8/25/13	1390	662	468	1390	3910	1208	1398	1206	3812
9/1/13	1679	702	512	1321	4214	1281	1338	1451	4070
9/8/13	1491	888	522	1403	4304	1285	1801	1631	4717
9/15/13	1375	1346	502	1369	4592	1413	1650	1852	4915
9/22/13	1738	1302	519	1455	5014	1611	1453	1844	4908
9/29/13	1657	1472	533	1404	5066	1294	1494	1728	4516
10/6/13	1917	1459	538	1514	5428	1592	1746	1597	4935
10/13/13	2158	1315	553	1424	5450	1592	1416	1487	4495
10/20/13	2269	1700	649	1438	6056	1509	1590	1860	4959
10/27/13	2665	2242	679	1455	7041	1431	1624	1476	4531
11/3/13	2235	2056	879	1642	6812	1879	2220	1647	5746
11/10/13	2459	2105	984	1673	7221	1813	2216	1629	5658
11/17/13	2589	1967	1000	1593	7149	1556	1766	2661	5983

Table 3 – Historical data collected by scraping health agency website

Region-wise grouping of all data is done according to the regions defined by Health agencies and also available data from metropolitan cities.

Graphical representation of historical data –

Figure 1 – Historical data representation of San Francisco (2003-2004)

Figure 2 – Historical data representation of San Jose (2003-2004)

Figure 3 – Historical data representation of Santa Clara (2003-2004)

Figure 4 – Historical data representation of Sunnyvale (2003-2004)

Figure 5 – Historical data representation of whole North Cal cluster (2003-2004)

Figure 6 – Historical data representation of whole South Cal cluster (2003-2004)

Historical Data Charts for 53 weeks of all 10 years to get for 2 regions to get fitting function -

Averaging all 53 weeks of all past data (union with real-time data till last successfully completed week) to determine fit function.

Week No.	Avg of 10 yrs (NorthCal)	Avg of 10 yrs (SouthCal)
01	8222.6	6125
02	8038.6	5965
03	8280.6	6346.6
04	8428.7	6241.5
05	8419.5	6247.3
06	8886.4	6884.2
07	9537.6	7485.9
08	10488.8	8579.9
09	13068	10021.8
10	16158	12745.9

11	18306.6	14313.3
12	18760.2	13984.3
13	16857	12056.1
14	14686.1	10284.2
15	13399.4	9264.9
16	12615.6	8729.5
17	13135	8182.3
18	13125.3	7695.3
19	12389.9	6938.2
20	11559.9	6607.5
21	10509.6	5988.7
22	9499	5331.1
23	8821.6	4995.7
24	7961.3	4563.6
25	7277.1	4197.7
26	6817.6	4004.9
27	7340.1	4622.4
28	7035.7	4162.1
29	6222.2	3656.1
30	5662.2	3390.8
31	5389.6	3204.5
32	5046.3	2911.2
33	4773.9	2762.8
34	4760.5	2732.9
35	4633.5	2613.1
36	4538.6	2616.7
37	4334.3	2566.7
38	4081.8	2457.1
39	3951.4	2432
40	3873.7	2459.8
41	3795.1	2409.3
42	3861	2474.7
43	3878.9	2613.7
44	3991.8	2778.2
45	4129.5	2911.6
46	4362.9	3133.9
47	4590.7	3538.4

48	4878.1	3900.9
49	5436.7	4376.7
50	6248.2	4677.4
51	7038.9	5205.2
52	7661.9	5584
53	8177.2	6217.1

Table 4 – Average of all 10yrs of historical data region-wise

The above data can drawn on a graph in this way. You can see in below graph there is a relation between data for these two cities and that's what we are trying to use to understand and learn from to make future predictions.

Figure 7 – Average of all 10 years of historical data for both North Cal and South Cal

Prediction using curve fitting function technique –

The prediction of next day data is done using the current real-time data and all historical data. This prediction is done by using fitting function statistics from historical data and applied onto the real-time data to get future data points.

Regression Technique -

Polynomial regression is similar to linear regression but over here the relationship between the independent variable x and the dependent variable y is pictured to be of an order of n th polynomial. It fits appropriately a nonlinear relationship between the value of x and the corresponding conditional mean values of y . It has been used variedly to describe various nonlinear behaviors, which includes my study of interest here progression of disease epidemics.

The goal of using regression analysis is to evaluate the expected value of y (dependent variable) in terms of the value of x (an independent variable).

Simplest form of linear regression is given by -

$$Y = a_0 + a_1X + \epsilon$$

And it is used, where ϵ is some random error. In this equation, for each unit increase in the value of x , the conditional expectation increases by a_1 units of y as well. As this being the simplest case this, relationship form is rarely used.

In general, we can model the expected value of y as an n th order polynomial, which gives us the general polynomial regression model as follows –

$$Y = a_0 + a_1X + a_2X^2 + \dots + a_nX^n + \epsilon$$

Also, in this regression method, the choice of degree and the evaluation of the fit's quality depend on judgments that are left up to the user.

Finding the best model: The coefficient of determination R^2 tells you how well your model describes the data. It takes values between 0 and 1 and the closer it is to 1, the better your data fits the regression curve. It is important to keep in mind that the value of R^2 alone is not the only factor that you should consider when choosing the best model.

Other factors include:

- The validity of the model.
- The effectiveness of the model.

And in my case, I have found results with regression model of order 10 be the most efficient with only 10-15% error.

How polynomial regression works –

Polynomial regression depends on calculating least square residual value. The aim of “least-squares fit” is to define a function $f(x)$ such that the sum of the square of the distance between $f(x_i)$ and y_i for each data point (x_i, y_i) is minimized (minimum error).

For example - To fit a second-order polynomial fit, you want to find the coefficients A, B, and C for the equation $f(x) = Ax^2 + Bx + C$. You find the formula for the sum of the squares of the difference and from that generate 3 new formulas: the partial derivative with respect to A, B, and C and set all those 3 equal to zero. You now have 3 equations and 3 unknowns, which you can then solve.

Assume example –

x	y
1	100
2	50
3	33
4	25
5	20

Using this you have to minimize the equation –

$$(A + B + C - 100)^2 + (4A + 2B + C - 50)^2 + (9A + 3B + C - 33)^2 + (16A + 4B + C - 25)^2 + (25A + 5B + C - 20)^2$$

Taking the derivative of that equation with respect to A, B, and C gives you the three equations:

$$2(A + B + C - 100) + 8(4A + 2B + C - 50) + 18(9A + 3B + C - 33) + 32(16A + 4B + C - 25) + 50(25A + 5B + C - 20) = 0$$

$$2(A + B + C - 100) + 4(4A + 2B + C - 50) + 6(9A + 3B + C - 33) + 8(16A + 4B + C - 25) + 10(25A + 5B + C - 20) = 0$$

$$2(A + B + C - 100) + 2(4A + 2B + C - 50) + 2(9A + 3B + C - 33) + 2(16A + 4B + C - 25) + 2(25A + 5B + C - 20) = 0$$

Which simplify to:

$$1958A + 450B + 110C = 2994$$

$$450A + 110B + 30C = 998$$

$$110A + 30B + 10C = 456$$

And solving those 3 equations for A, B, and C give us:

$$A = 7.0714285714287$$

$$B = -60.9285714285719$$

$$C = 150.60000000000008$$

Making our curve function –

$$Y = 7.0714285714287x^2 + -60.9285714285719x + 150.60000000000008$$

X	Y	Calculated y	Error
1	100	96.74285714	3.257142857
2	50	57.02857143	7.028571429
3	33	31.45714286	1.542857143
4	25	20.02857143	4.971428571
5	20	22.74285714	2.742857143

Table 5 – Polynomial Regression results for example dataset with order 2

Figure 8 – Function plot we got as result of applying polynomial regression

Similarly, following above technique we get fitting function for the historical data we have collected by scraping the health agency websites.

Curve fitting function for North Cal –

Week No.	Avg. of 10 yrs. (NorthCal)	Calculated Y	Error
01	8222.6	8278.680714	56.0807138
02	8038.6	8385.096519	346.4965186
03	8280.6	7917.158879	363.4411208
04	8428.7	7731.85647	696.8435303
05	8419.5	8145.702231	273.797769
06	8886.4	9151.587557	265.1875568
07	9537.6	10573.77562	1036.175625
08	10488.8	12173.84717	1685.047169

09	13068	13718.54458	650.544577
10	16158	15018.76647	1139.233531
11	18306.6	15947.43699	2359.163007
12	18760.2	16442.60417	2317.595828
13	16857	16500.90188	356.0981194
14	14686.1	16165.43263	1479.332626
15	13399.4	15511.18539	2111.785386
16	12615.6	14630.28652	2014.686515
17	13135	13618.68432	483.6843157
18	13125.3	12565.28147	560.0185317
19	12389.9	11544.04627	845.8537337
20	11559.9	10609.24571	950.6542928
21	10509.6	9793.643092	715.956908
22	9499	9109.282059	389.7179407
23	8821.6	8550.330097	271.269903
24	7961.3	8097.369696	136.0696955
25	7277.1	7722.496602	445.3966024
26	6817.6	7394.604279	577.0042788
27	7340.1	7084.293804	255.806196
28	7035.7	6767.941301	267.7586994
29	6222.2	6430.572621	208.3726212
30	5662.2	6067.32972	405.1297199
31	5389.6	5683.45699	293.8569905
32	5046.3	5292.881062	246.5810616
33	4773.9	4915.59626	141.6962601
34	4760.5	4574.192356	186.3076436
35	4633.5	4289.963458	343.536542
36	4538.6	4079.109184	459.4908157
37	4334.3	3949.573712	384.7262879
38	4081.8	3899.057079	182.7429209
39	3951.4	3914.668462	36.73153757
40	3873.7	3974.565151	100.8651509
41	3795.1	4051.7258	256.6258001
42	3861	4119.734437	258.7344371
43	3878.9	4160.094758	281.1947578
44	3991.8	4170.144689	178.3446893
45	4129.5	4170.09114	40.59113958

46	4362.9	4207.026502	155.8734977
47	4590.7	4353.013986	237.6860138
48	4878.1	4693.430367	184.6696325
49	5436.7	5300.724482	135.9755178
50	6248.2	6187.57986	60.62014027
51	7038.9	7232.152516	193.2525156
52	7661.9	8066.582209	404.6822092
53	8177.2	7918.339665	258.8603345

Table 6 – Polynomial Regression results for historical dataset with order 10 (North Cal)

$$Y = -2.158440256 \cdot 10^{-9} x^{10} + 5.972822644 \cdot 10^{-7} x^9 - 7.069923942 \cdot 10^{-5} x^8 + 4.665172946 \cdot 10^{-3} x^7 - 1.872535528 \cdot 10^{-1} x^6 + 4.672330251 x^5 - 70.82114664 x^4 + 605.4740074 x^3 - 2517.936707 x^2 + 4350.605057 x + 5906.869831$$

Figure 9 – Function plot by applying polynomial regression of order 10 (North Cal)

The function smoothens the curve and ignores all the outlier cases and gives us good enough approximation of data to make better predictions from this data.

Curve fitting function for South Cal –

Week No.	Avg. of 10 yrs (SouthCal)	Calculated Y	Error
01	6125	6221.080253	96.08025308
02	5965	6228.1274	263.1273997
03	6346.6	5845.033889	501.5661113
04	6241.5	5761.083936	480.4160643
05	6247.3	6201.227086	46.0729144
06	6884.2	7119.887877	235.6878775
07	7485.9	8338.064975	852.1649755
08	8579.9	9635.657883	1055.757883
09	10021.8	10809.2026	787.4025977
10	12745.9	11703.60498	1042.295017
11	14313.3	12225.02484	2088.275157
12	13984.3	12340.77722	1643.522783
13	12056.1	12070.97291	14.87291114
14	10284.2	11475.6103	1191.410299
15	9264.9	10639.94751	1375.047506
16	8729.5	9660.220927	930.7209273
17	8182.3	8631.125063	448.8250625
18	7695.3	7635.922603	59.37739749
19	6938.2	6739.60503	198.5949697
20	6607.5	5985.165398	622.3346021
21	5988.7	5392.768907	595.9310931
22	5331.1	4961.406088	369.6939119
23	4995.7	4672.480308	323.2196925
24	4563.6	4494.708612	68.89138768
25	4197.7	4389.695147	191.9951469
26	4004.9	4317.562109	312.6621091
27	4622.4	4242.087055	380.3129453
28	4162.1	4134.889882	27.21011835
29	3656.1	3978.330616	322.2306161
30	3390.8	3766.912764	376.112764

31	3204.5	3507.129068	302.6290677
32	2911.2	3215.829599	304.6295992
33	2762.8	2917.328813	154.5288133
34	2732.9	2639.591062	93.3089383
35	2613.1	2409.935708	203.1642919
36	2616.7	2250.775981	365.9240191
37	2566.7	2175.942618	390.7573816
38	2457.1	2188.136809	268.9631914
39	2432	2277.999458	154.0005422
40	2459.8	2425.168051	34.63194949
41	2409.3	2601.510845	192.2108453
42	2474.7	2776.473488	301.7734882
43	2613.7	2924.137892	310.4378922
44	2778.2	3031.170011	252.970011
45	2911.6	3104.31452	192.7145199
46	3133.9	3175.472969	41.57296883
47	3538.4	3301.670303	236.7296973
48	3900.9	3556.365321	344.5346795
49	4376.7	4007.586231	369.1137693
50	4677.4	4677.265598	1.344020903·10 ⁻¹
51	5205.2	5474.902173	269.7021726
52	5584	6097.282993	513.2829934
53	6217.1	5884.450304	332.6496962

Table 7 – Polynomial Regression results for historical dataset with order 10 (South Cal)

$$Y = -2.043886443 \cdot 10^{-9} x^{10} + 5.6487132 \cdot 10^{-7} x^9 - 6.662902465 \cdot 10^{-5} x^8 + 4.367947075 \cdot 10^{-3} x^7 - 1.734752015 \cdot 10^{-1} x^6 + 4.259958788 x^5 - 63.1068545 x^4 + 522.8371095 x^3 - 2089.624974 x^2 + 3440.997311 x + 4405.886876$$

Figure 10 – Function plot by applying polynomial regression of order 10 (South Cal)

The function quite smoothens the curve and ignore all the outlier cases and gives us good even approximate to make better predictions in the future.

Verification -

For verification purpose, I ran my analysis for the week of 17th Nov to 24th Nov. During this week as you could see from the historical data, I don't have any results. So, what I do is for this particular week, I find out the people who are infected and talk about it on twitter. These people are likely the people who will show up in health agency database in coming week's result. So, from the historical data that I have and from the data I analyzed from Twitter would be use my curve fitting function and judge how many more people are really going to be affected for the infection. So for confirmation, I compared this calculated results from all data (past and present real-time data) with actual data at the end of the week proves my thesis. There is an error of 5-10%, which is quite acceptable for this type of regressive data intensive problem.

Limitation

There are certain limitations to my proposed approach.

- The clustering of regions that I have used is based on the data I have from health agency databases. So in future as we go on collecting real-time daily data we can store and use it later to get much better trend analysis on daily basis rather using weekly data for analysis.
- For prediction, I used hybrid polynomial regression method, which may not be the best one, but according to my research it is one of the most intuitive and effective one for our data points. But there are many other approaches for this type of interpolation using time series etc.
- Also, with respect to my proposed approach, I did only consider Twitter data as it gives us much better freedom to access their data using their APIs. But if we were to get special accesses to other social media data as well, we could get more data and hopefully much better results.

Future Work

- In-depth research of all different algorithms better than my approximation approach can be used for prediction of future data points.
- Even developing better algorithm for data cleaning could be considered so that outlier cases are better dealt.
- Also, as I mentioned before more data from different social networking mediums could also be considered but upon getting better access to the data within them.

Conclusion

In this project, I presented a way to determine rate of spread disease using social networking data, which is more or less instantaneous. I did some test runs of my approach with historical datasets from reliable sources like government health agency websites and google trends against datasets retrieved from Twitter. Results generated with my approach are significantly accurate as compared to the actual results with an error of 5-15%.

References

[1] Centers for Disease Control and Prevention

<http://www.cdc.gov/flu/weekly/>

[2] Stefan Waner, Steven R. Costenoble, "Finite Mathematics and Calculus Applied to the Real World"

[3] Twitter real-time data

<https://twitter.com/search?q=flu%20%3A%28&src=typd&f=realtime>

[4] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis (Foundations and Trends in Information Retrieval)"

[5] Google Trend Data <http://www.google.org/flutrends/us/#1014226>

[6] Bing Liu, Morgan & Claypool Publishers, "Sentiment Analysis and Opinion Mining"

[7] Text Mining

http://en.wikipedia.org/wiki/Text_mining

[8] Polynomial Regression

http://en.wikipedia.org/wiki/Polynomial_regression

[9] LA County Public Health Department

<http://www.publichealth.lacounty.gov/acd/Flu.htm>

[10] California Department of Health Care Services,

<http://www.dhcs.ca.gov/dataandstats/reports/Pages/default.aspx>

[11] State and Local Health Departments and Services

<http://www.statelocalgov.net/50states-health.cfm>

[12] San Francisco Department of Public Health

<http://www.sfcddcp.org/healthalerts.html>