Flu Trend Prediction Using Social Media Network Data

Ali Al Essa, Dr. Miad Faezipour, Dr. Jeongkyu Lee, and Gopala Duggina
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Introduction

Flu and Seasonal Influenza are serious problems that may cause fatalities. About 250,000 to 500,000 deaths worldwide each year are because of flu [1]. Public health providers are in favour of knowing about the seasonal flu as soon as possible in order to take the required actions for their communities. Getting an early warning will help to prevent the spread of flu in the population.

Nowadays a very large number of people use social media networks on a daily basis to share their news, events, and even their health status [2]. This leads to the idea of using commonly used social media networks for flu detection and providing early warnings to public health providers to take the right action at the perfect time. Users of social media network can be used as sensors to predict the flu trend in a specific area and time. The social media network used for this research project is Twitter. It is one of the most widely used social network. It has over 271 million monthly active users [3]. The retrieved data from Twitter is enormous and contains large number of attributes. However very few attributes will only be required in this project analysis. Since the Twitter data is very huge, Big Data Hadoop systems and MapReduce programming can be used for analysis and to create a good prediction model for flu trend and then to provide health care providers with timely warnings. This can help health care industry to provide high quality services at the right time.

Data Set

The data set is collected from Twitter web site using a developed crawler which works together with the Twitter API to stream tweets. The crawler is designed to filter the tweets based on flu-related keywords such as flu and influenza. Also the developed crawler cleans the data set to include only the following attributes:

- Created_at: tweet posting date
- Text: the posted tweet
- Loc: location of the user.
- Time Zone: the time zone used by the user

The data set consists of 135,160 tweets collected during the month of November 2015.

Problem Definition

Most of health care providers take the required actions only after getting the reports of flu from the Center for Disease Control and Prevention (CDC). This center collects data from health care providers to monitor Influenza –Like illnesses. It then publishes the reports. This generally takes one to two weeks delay resulting in the fact that the required warnings come late to the provider's attention [1]. This is while health-care providers need to be warned early enough in order to do the right actions to prevent the spread of flu.

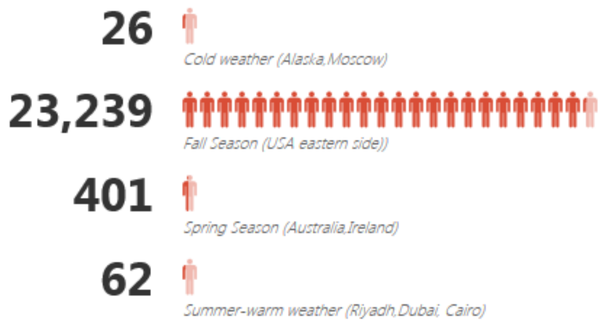

Figure 1: Flu trend and seasons relationship

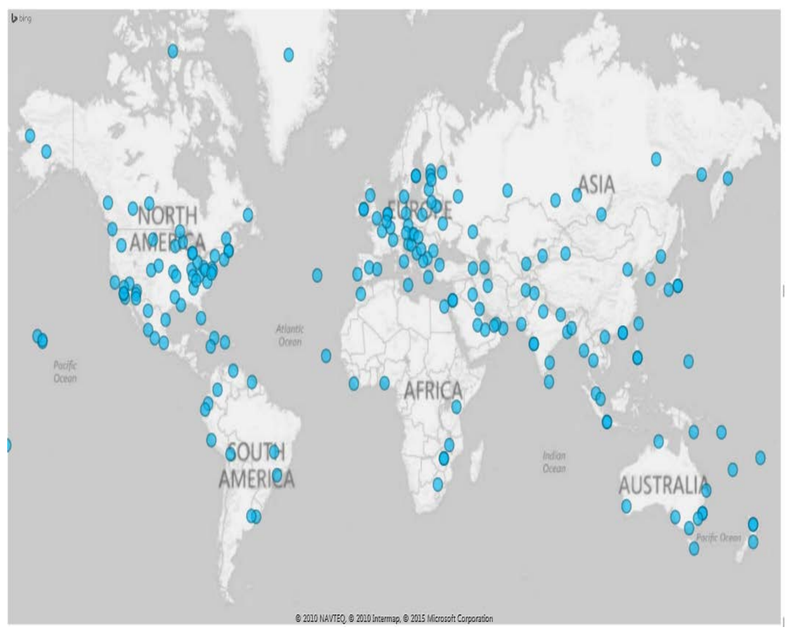

Figure 2: Flu trend based on location

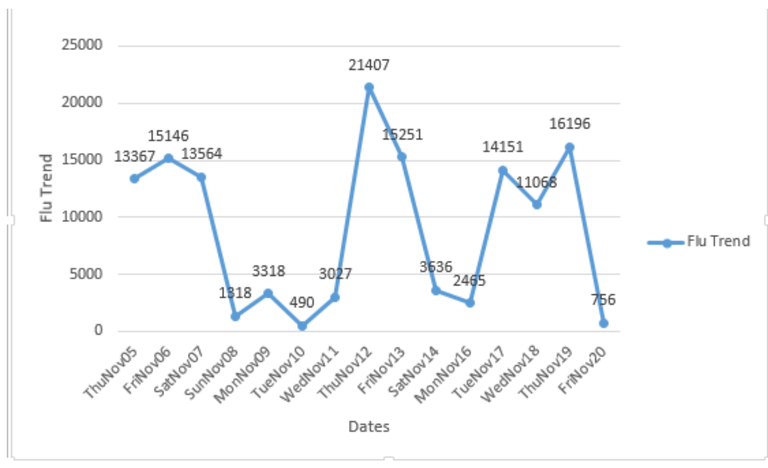

Figure 3: Flu trend based on time

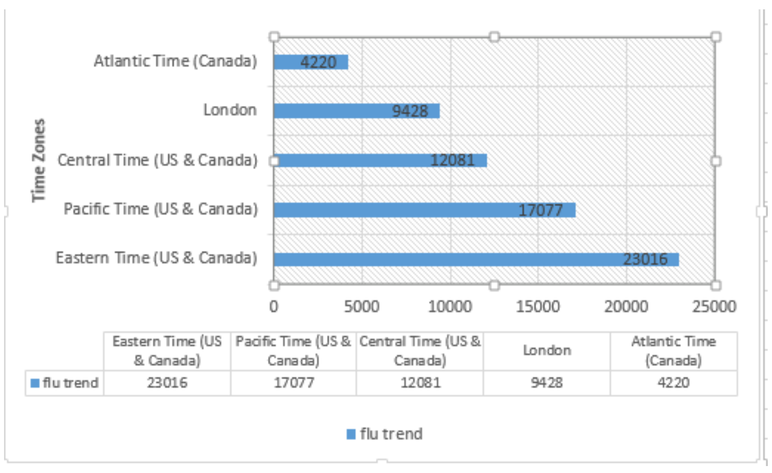

Figure 4: Top 5 affected areas based on time zones

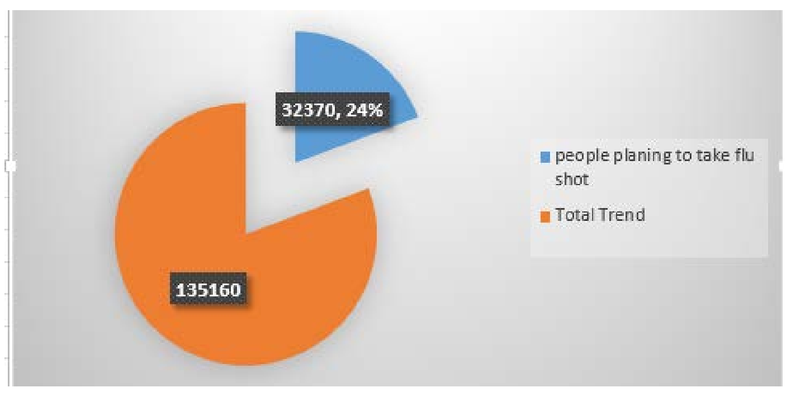

Figure 5: Tweets on flu and flu shot

Implementation, Analysis, and Results

The prediction of flu spread could be much more efficient by taking advantage of Hadoop, MapReduce programming and Hadoop Eco Systems. These tools and techniques can use the available twitter data and predict the flu trend at the earliest time. In the MapReduce programming we used one mapper and one reducer to do all the analysis scenarios. The mapper's input is just the collection of tweets about flu. Mapper takes the input and then does the required mapping process as a key and value pairs. Finally the reducer takes the output from the mappers and sums up the values part.

In this study we implemented six different flu prediction analysis scenarios.

1.FLU TREND AND SEASONS RELATIONSHIP:

Figure 1 describes our first analysis based on flu trend and seasons relationships.

We picked four areas with different seasons during November. This analysis shows that the eastern side is the most affected area in November. We conclude that the flu virus became more active during fall season.

2. FLU TREND BASED ON LOCATION:

This analysis shows the flu trend in different locations. Figure 2 shows users' flu-tweets around the world.

From this analysis we can say that people in the USA and EUROPE are most affected when compared to people in the other locations.

3. FLU TREND BASED ON TIME:

This analysis shows the number of flu-tweets in different dates. Figure 3 demonstrates the results of this analysis

From the figure we can observe that the maximum flu trend was on Thursday November 12th and least flu trend on Tuesday November 10th. From this analysis we can say that people are affected by flu after long working days.

4.TOP 5 AFFECTED AREAS BASED ON TIME ZONES:

From Figure 4 we can say that people in the areas using the Eastern Time zone are the most affected by flu. Based on this analysis people in this area have to take necessary actions to prevent the flu.

5. FLU AND FLU SHOT:

This analysis provides the information on how many tweets have been posted about flu and gives the number of users who took or are planning to take a flu shot or vaccine. Figure 5 shows surprisingly that 32,370 people out of 135,160 tweet about flu shot. This constitutes only 24% of whole dataset.

Conclusion and Future Work

Most of health care provides need to get an early warning of the flu season in order to take right actions. Our results show that using the data of social media network together with BigData tools and techniques can provide an early warning about flu trend, which might help to prevent the spread of flu in the population.

For the evaluation, we are planning to use CDC reports and Google Flu Trend as a ground truth. This will help measure the quality of our proposed solution output.

References

[1].Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on (pp. 702-707). IEEE.
[2].Murthy, D., Gross, A., & Longwell, S. (2011, June). Twitter and e-health: a case study of visualizing cancer networks on twitter. In Information Society (i-Society), 2011 International Conference on (pp. 110-113). IEEE.
[3].Nambisan, P., Luo, Z., Kapoor, A., Patrick, T. B., & Cisler, R. (2015, January). Social Media, Big Data, and Public Health Informatics: Ruminating Behavior of Depression Revealed through Twitter. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 2906-2913). IEEE.
[4].Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media.International journal of environmental research and public health, 7(2), 596-615.