

INTRODUCTION

Health data is important because they measure a wide range of health indicators for a community. CDC collects health data using regular surveys which requires population surveillance with the help of Health Professionals. Since people are active on social media, a low cost alternative would be to collect and analyze those records from social network. In this work we consider broader range of health applications using twitter data. here we try to classify and analyze twitter data using supervised learning methods. We incorporate prior knowledge as training data into this learning process and apply it to track illnesses over different time zones of USA. The result shows a strong correspondence between tweets and public health data.

OBJECTIVES

Public health researchers dedicate considerable resources to population surveillance, which requires the use of public surveys and health professionals. The data collected from social network can be considered as a low-cost alternative to that. Several Twitter studies have demonstrated that aggregating millions of messages can provide valuable insights into a society. These studies were actually used in tracking the influenza rate in United states and United Kingdom. Using this we can broaden the spectrum by analyzing the tweets for overall health statistics of different time zones across united states. so our major objectives are

- To collect twitter data through out united states over a period of time.
- To Analyze the health factor in various time zones in United States.
- To Find the trend of diseases being reported by people across Social networks.

The data from social media has been successfully used in various cases that motivated us to do this study. Our main motivations for this work are inspired from past works which are

- Detection of Influenza Outbreaks using twitter messages.
- Analyzing tweets for H1N1 Pandemic using predictive models.
- These served as motivation for a general approach to analyze twitter for various diseases throughout the country

DATA & TECHNOLOGY USED

DATA USED

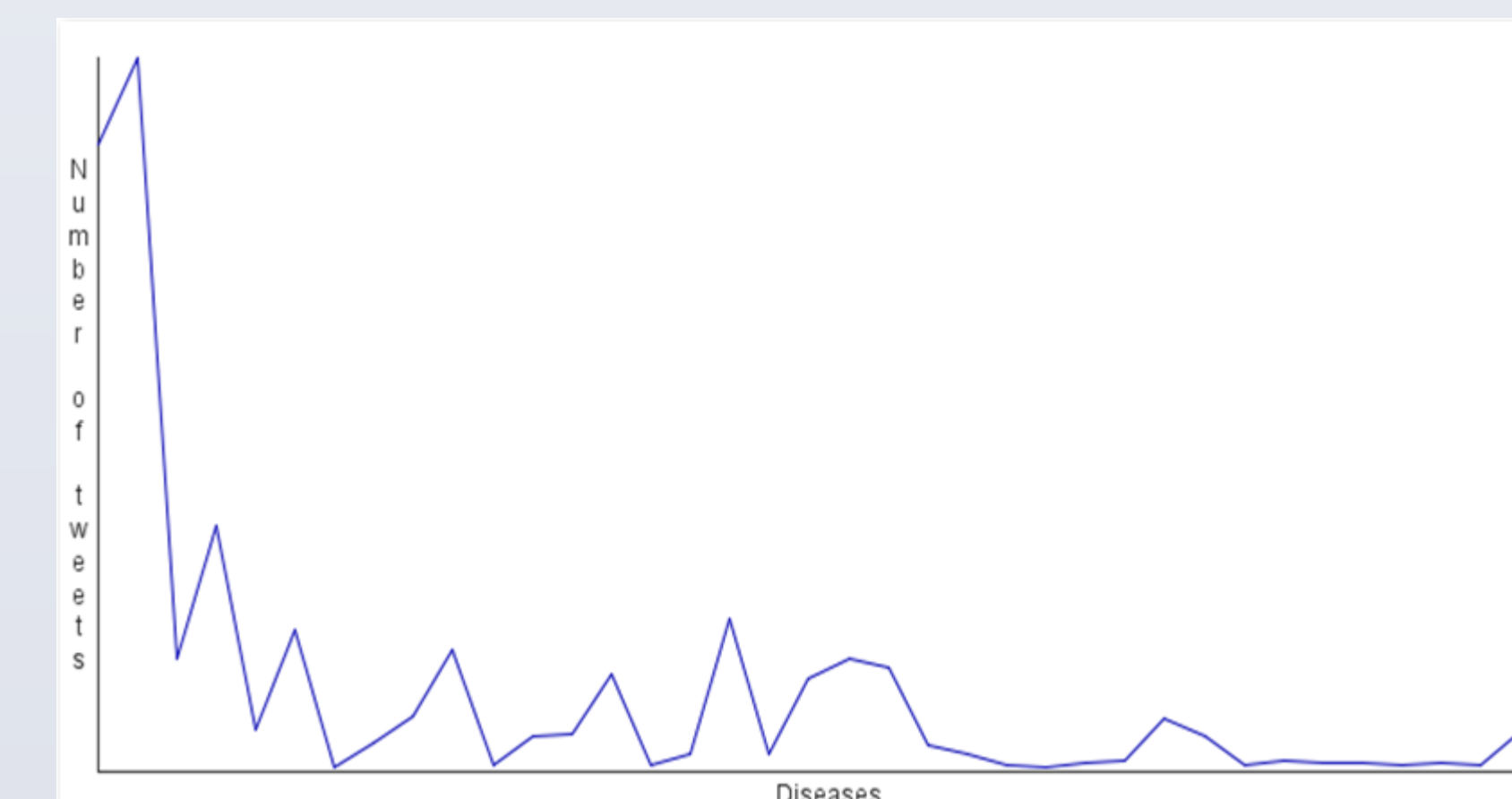
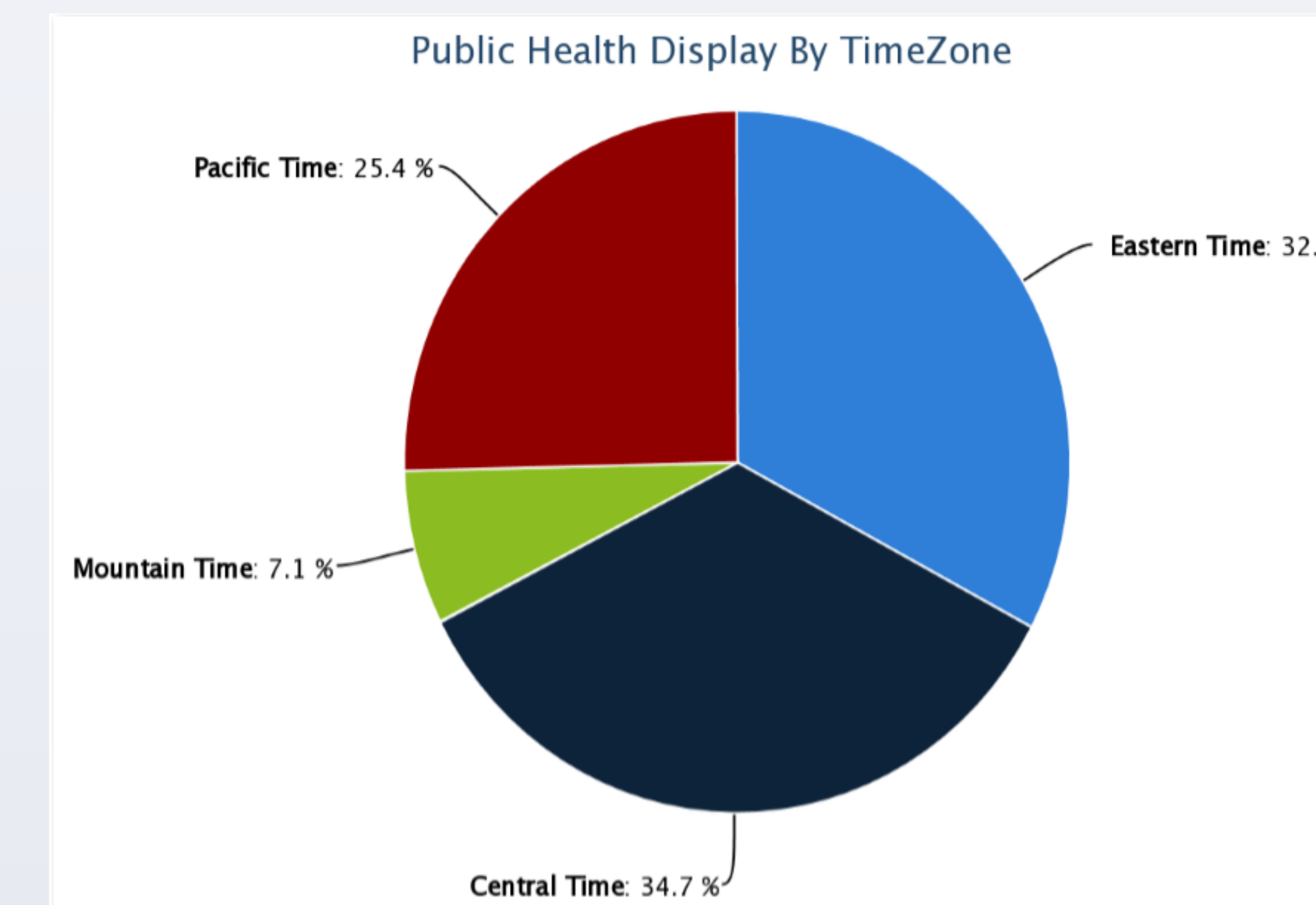
- Data is downloaded from twitter public feed using Twitter API from twitter.
- Data Cleaning has to be done to remove unwanted hash tags, user names and retweets
- Around 2.5GB of data has been downloaded from twitter which correspond to health related factors. We identify the tweets that contain health information and create the data set
- Removed the retweets(containing RT's) and tweets containing URL's in the tweet text
- URL's were mostly false positives as they contained news articles rather than users health
- The following fields were extracted from twitter which were useful. Tweets related to US time zone were filtered
 - Tweeted text
 - Time zone of the tweet
- A Simple filtering would not be sufficient as tweet with a given keyword can be used in many contexts
 - E.g.: "I have Spring fever"
- A sample of 2000 tweets were taken from the parsed tweets. These tweets were classified for to be used as training data for a supervised classifier.
- These tweets were taken to label as
 - Health related: User is suffering from some disease
 - Not health related: User is not suffering from any disease.
- Supervised Learning helps to analyze the data and recognize the patterns used for classification and regression analysis.

SUPPORT VECTOR MACHINE

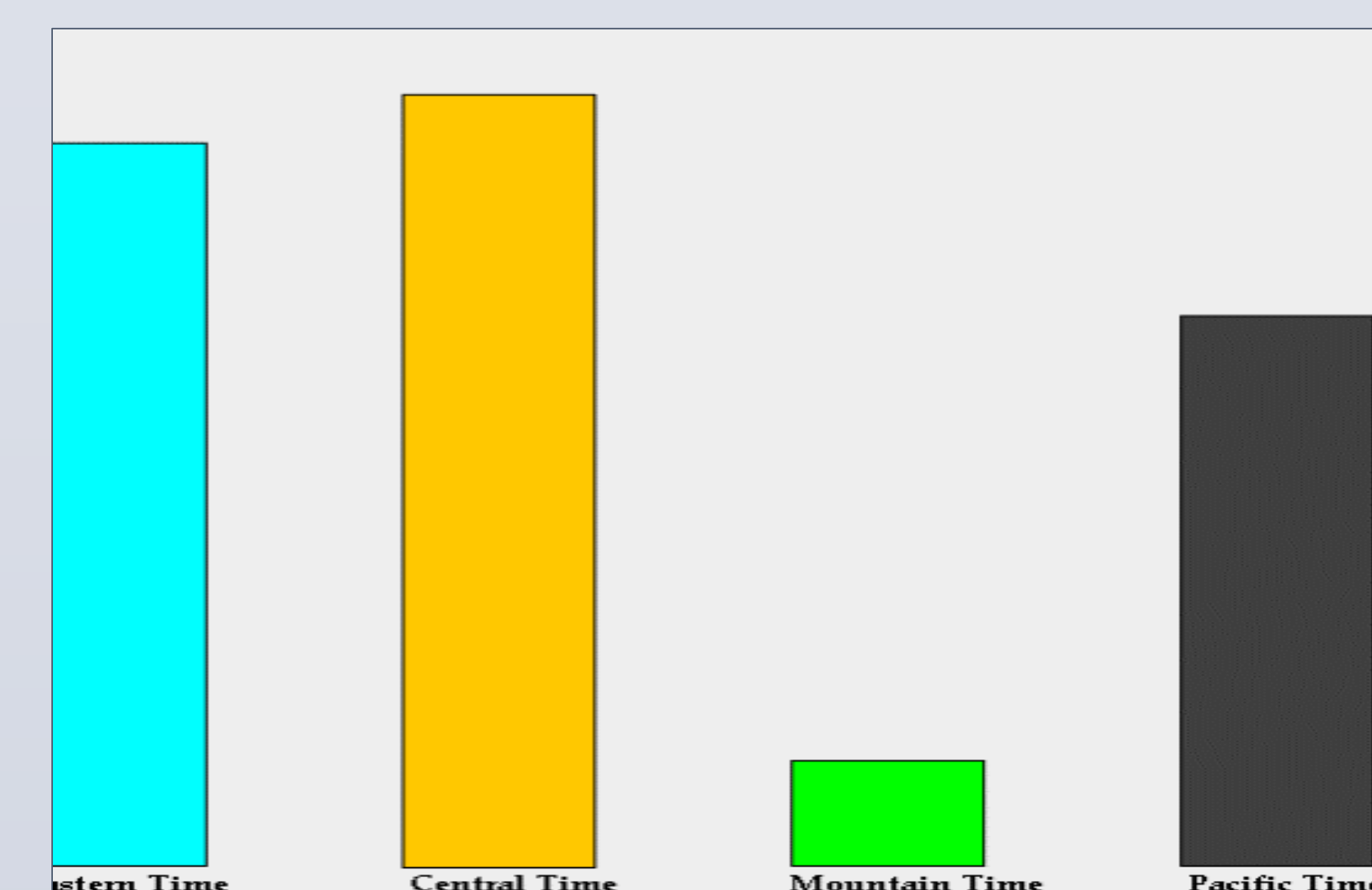
- Supervised Learning helps to analyze the data and recognize the patterns used for classification
- In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification.
- The SVM will be used to create a model file from the set of training examples we give.
- A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on
- The model file is then used to classify the whole input file into health related and not health related using SVM training.

RESULTS

After training and classifying the tweets we analyzed the data. The health related tweets are analyzed based on different time zones and the health statistics are then calculated



This trend graph displays the percentage amount of tweets received for a particular disease. We have constructed this graph by taking 38 diseases.



This bar graph represents the statistical data resulted from our study. The collected tweets are analyzed based on time zone for health statistics. This data is represented in form of a bar graph. This graph gives us a rough idea about the trend of diseases in a particular time zone.

CONCLUSION

We have demonstrated that public health information can be extracted from Twitter. We created a training corpus of 2000 messages labeled for relevance to health and produced a labeling of 154000 Twitter messages. Our machine learning model learns to group tweets into health related and unrelated topics. Then we analyzed the related tweets to get the statistical result about public health in different time zones.

We plan to extend present models tailored to specific problems to implement a better training of the SVM and to Compare the result with government survey reports. We also plan to do state wise analysis and a more specific application to extract more knowledge from public tweets.

REFERENCES

- [1] <http://svmlight.joachims.org/>
- [2] <http://www.support-vector.net/>
- [3] Dustin Boswell, "Introduction to Support Vector Machines", 2002
- [4] Michael J. Paul; Mark Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health", Association for the Advancement of Artificial Intelligence, 2011
- [5] Paul, M., and Dredze, M. 2011, "A model for mining public health topics from twitter", Technical report, Johns Hopkins University

CONTACTS

Soumya Pani
Chaitanya Panuganti

soumyadebabrata@ufl.edu
chaitanyapvsk@ufl.edu