COMPUTATIONAL CONTENT ANALYSIS AND STUDY OF ZIKA VIRUS OUTBREAKS ON TWITTER

Arabh Kumar (2014IPG-020)
Vipin Kumar (2014IPG-103)
Buddh Priy Maury (2014IPG-116)

ATAL BIHARI VAJPAYEE INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT
Gwalior-474 010, MP, India

September 20, 2017

Outline

- 1 Introduction
- 2 Objective
- 3 Motivation
- 4 Literature Review
- 5 Methodology
  - Data Extraction
  - Data Preprocessing
  - Analysis of data
- 6 Result and Discussion
- 7 Conclusion
- 8 Future Scope
- 9 Reference

Introduction

- Twitter is one of the most popular social media site where people express their views on different topics which also include health related topics.
- People's opinion towards any major health crisis which they share on twitter could help public health agencies to predict awareness level within society about that particular health crisis or epidemic.
- The challenge is to gather all such relevant data,detect and summarize the overall sentiment on a topic(Zika Virus in our study).

objective

- To examine level of concern on Zika virus by analyzing sentiment polarity of tweets and what number of people are twitting about prevention, transmission, treatment, symptom, mosquito, and pregnancy.

Motivation

- Twitter is a large social media channel where users tweet about various topic which also includes health issues.
- Traditional disease surveillance was done manually by selecting some target population and collecting their view about any particular disease.
- Social media channels, like Twitter, provides continuous information on public opinion about any epidemic and other health issues which can help public health agencies in performing real time surveillance.

Literature Review

- The first work on sentiment analysis aimed at classifying text by overall sentiment, not just focused on any one topic[1].
- Collection of a large amount of data has been helpful to find out what people are thinking or presuming. Recently with the boom in social media sites, data available for opinion mining is very large [2].
- Natural Language Toolkit (NLTK) is a library. This library is a combination of many script modules, a big set of structured files, different tutorials, numerous statistical functions, machine learning classifiers, etc [3].

Literature Review

- French Polynesia went through the biggest Zika virus outbreak between 2013 and 2014. Increase in Guillain-Barre syndrome was identified during the period of the Zika virus outbreak. There was an expected relation between Zika virus and Guillain-Barre syndrome [4].
- In traditional survey based methods, there is a big time gap but in new techniques of big social data mining help us to get rid of that time gap and also take care of privacy concerns in order to study public behavior on specific issues [5].

Methodology

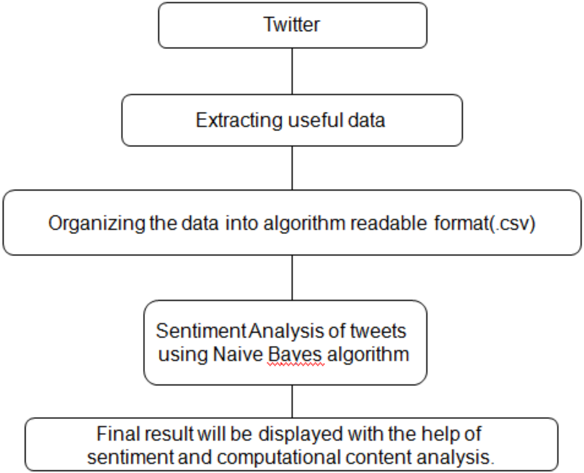

Twitter
Extracting useful data
Organizing the data into algorithm readable format(.csv)
Sentiment Analysis of tweets using Naive Bayes algorithm
Final result will be displayed with the help of sentiment and computational content analysis.

Figure: Flowchart of proposed activities

Data Extraction

- Data is extracted from Twitter using keywords 'zika' , 'zika virus'.
- Twitter API is used for data extraction from Twitter.
- Original data set is in the .json format.
- Useful data is kept and unwanted data is removed.

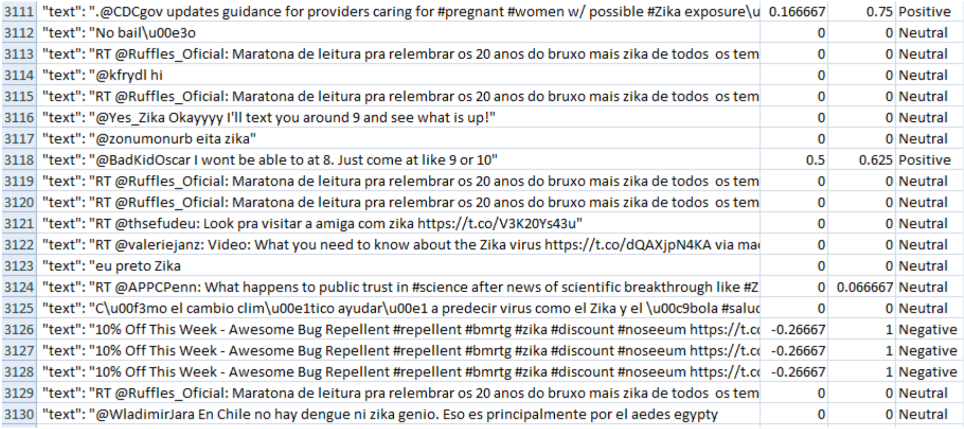
Data Preprocessing

- The original dataset obtained using Twitter API contains information about date , time , language , location , links etc.
- We cleaned the data and make a dataset of which contains only tweets.
- We have converted that dataset into .csv format to be later read by sentiment analysis algorithm.

Analysis of data

- We have used Natural Language Toolkit (NLTK) and different Python libraries .
- We used Naive Bayes Classifier.
- We classified tweets in different classes of polarity
  - Positive
  - Negative
  - Neutral

continued..



3111	"text": ".@CDCgov updates guidance for providers caring for #pregnant #women w/ possible #Zika exposure\u	0.166667	0.75	Positive
3112	"text": "No bail\u00e3o	0	0	Neutral
3113	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3114	"text": "@kfrydl hi	0	0	Neutral
3115	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3116	"text": "@Yes_Zika Okayyyy I'll text you around 9 and see what is up!"	0	0	Neutral
3117	"text": "@zonumonurb eita zika"	0	0	Neutral
3118	"text": "@BadKidOscar I wont be able to at 8. Just come at like 9 or 10"	0.5	0.625	Positive
3119	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3120	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3121	"text": "RT @thsefudeu: Look pra visitar a amiga com zika https://t.co/V3K20Ys43u"	0	0	Neutral
3122	"text": "RT @valeriejanz: Video: What you need to know about the Zika virus https://t.co/dQAXjpN4KA via ma	0	0	Neutral
3123	"text": "eu preto Zika	0	0	Neutral
3124	"text": "RT @APPCPenn: What happens to public trust in #science after news of scientific breakthrough like #Z	0	0.066667	Neutral
3125	"text": "C\u00f3mo el cambio clim\u00e1tico ayudar\u00e1 a predecir virus como el Zika y el \u00c9bola #salu	0	0	Neutral
3126	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3127	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3128	"text": "10% Off This Week - Awesome Bug Repellent #repellent #bmrtg #zika #discount #noseeum https://t.c	-0.26667	1	Negative
3129	"text": "RT @Ruffles_Oficial: Maratona de leitura pra relembrar os 20 anos do bruxo mais zika de todos os tem	0	0	Neutral
3130	"text": "@WladimirJara En Chile no hay dengue ni zika genio. Eso es principalmente por el aedes egypty	0	0	Neutral

Figure: Filtered tweets with output

Result and Discussion

- We collected 26,239 tweets between 30th July 2017 and 6th August 2017.
- A .json file is generated each time when we ran our Python script for tweets fetching.
- We removed all other unnecessary information and created a data set which consists only original text.
- We found out that 17.54% tweets were positive, 5.08% were negative and 77.37% were neutral

continued..

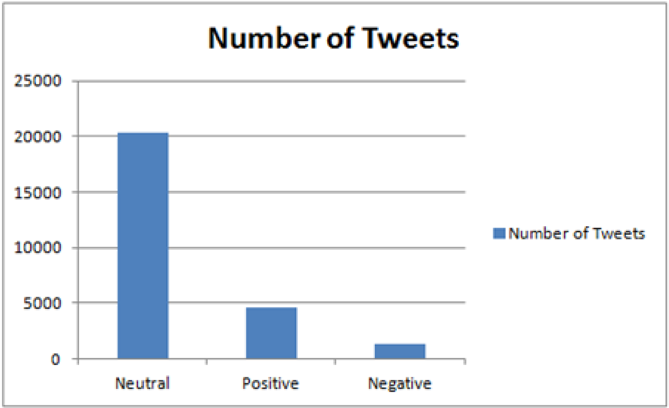


Figure: Graph between polarity and number of tweets

continued..

- We found that very high number of people were tweeting about the four diseases characteristics, mosquito, and pregnancy.
- It shows that very large number of people were tweeting relevant to this epidemic.

continued..

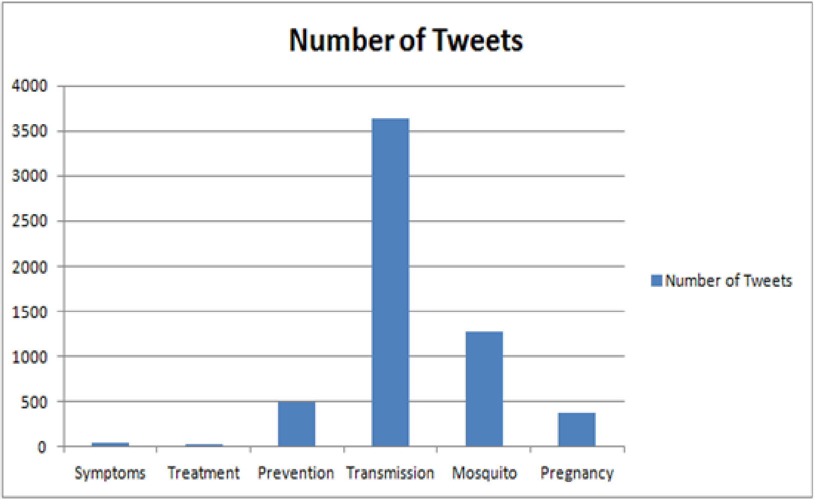


Figure: Number of tweets in each categorization after running all tweets

Conclusion

- 59% of tweets on Zika virus were negative between 2016-02-24 and 2016-04-27
- This shows very high level of concern between 2016-02-24 and 2016-04-27.
- In our study, we found that only 5.08% tweets are negative because we have done our study at such a time when Zika virus is no longer a matter of high concern among people.

Future Scope

- We can initiate to work in multiple languages to provide analysis to more locations where multiple languages are spoken.
- We can improve our system to filter out sentences which are not relevant to the topic but contains the common keywords of tweet collection..

References

1 Pang, B., Lee, L. and Vaithyanathan, S.: 2002, Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 79–86.

2 Pang, B., Lee, L. et al.: 2008, Opinion mining and sentiment analysis, Foundations and Trends R in Information Retrieval 2(1–2), 1–135.

3 Bird, S.: 2006b, Nltk: the natural language toolkit, Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, pp. 69–72.

4 Cao-Lormeau, V.-M., Blake, A., Mons, S., Lastère, S., Roche, C., Vanhomwegen, J., Dub, T., Baudouin, L., Teissier, A., Larre, P. et al.: 2016, Guillain-barrè syndrome outbreak associated with zika virus infection in french polynesia: a case-control study, The Lancet 387(10027), 1531–1539.

References

5 Purohit, H., Banerjee, T., Hampton, A., Shalin, V. L., Bhandutia, N. and Sheth, A. P.: 2015, Gender-based violence in 140 characters or fewer: A bigdata case study of twitter, arXiv preprint arXiv:1503.02086 .

Thank You