Zika misinformation tracking in social media

Presenter: Amira Ghenai - PhD Candidate U.Waterloo

Project with:
- Yelena Mejova
- Luis Fernandez Luque

Date: 19/09/2016
Outline

- Zika outbreak
- Project Goal
- Data Description
- Location extraction
- Topic extraction
- Labeling
- Timeline
The Zika virus infection across the Americas is considered a serious outbreak nowadays. WHO has declared an international health alert. PAHO & WHO send key messages to the population to minimize the risks (mosquito control, avoiding mosquito bites and pregnancy risks).
Vaccines cause microcephaly in babies

@Ex_MisTech #Zika is a sick joke on the uninformed. Mandated vaccines 2 pregnant moms causes microcephaly of course. ✌

Get your Zika-hoax poison here!

#FACTSONVAX

96% of babies tested with microcephaly did NOT have Zika virus. So yeah—that actually PROVES Zika doesn't cause shrunken heads.
Microcephaly is caused by genetically modified mosquitoes.
Fish can help stop Zika

A tiny fish to fight spread of Zika and other vector borne diseases - new research shows promise @APEIRCO

A New Weapon to Fight Mosquito-Borne Diseases: Tiny Fish
Scientists hope the simple solution can help slow the spread of dengue, malaria, Zika, and other illnesses.
takepart.com
Project Goal

• The objective of this project is to study the feasibility of using social media monitoring as a tool to help the communication effort in the health crisis

• Monitor potential treats to the communication effort such as people spreading rumors and misinformation about Zika infection to the world
Data Description

• Twitter data collection is happening on AIDR.
• Keywords related to Zika: microcephaly, Zika, Aedes, Zika fever…
• Period from 2016-01-13 to 2016-08-22
• All languages
• Total collected tweets ~ 13 million tweet
Data Description

The diagram shows the total number of tweets across different languages. According to the data:

- English: 6.27 million tweets
- Spanish: 3.69 million tweets
- Portuguese: 2.30 million tweets
- Other: 1.48 million tweets

The graph indicates that English is the most commonly used language, followed by Spanish and Portuguese, with Other languages having the least number of tweets.
• Location is important
• To extract the exact country name from tweets is not a trivial task.
• The explained method gives a very high coverage for the tweets to locate by country name
Location Extraction

- Very high coverage for the tweets to locate by country name

<table>
<thead>
<tr>
<th>Language</th>
<th>Coverage Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>68%</td>
</tr>
<tr>
<td>Spanish</td>
<td>63%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>64%</td>
</tr>
</tbody>
</table>

- English is spread in more than one location
- Spanish: Most tweets come from the southern Americas
- Portuguese: Most tweets are located in Brazil
Location Extraction – English Map
Automatic topic extraction
  – Latent Dirichlet allocation – LDA

Preprocessing:
  – Remove stop words/ highly frequent words
  – Remove Twitter special characters
  – Lower case
  – Tokenization
  – Stemming
• Run LDA for:
  – English
  – Spanish
  – Portuguese
• Example of English top 5 topics:
  – women_pregnant_travel_cdc_warn
  – case_first_confirm_report_transmit
  – birth_caus_babi_microcephali_link
  – spam_just_like_blood_look (weird!!)
  – mosquito_help_fight_control_can
Labelling

- Each topic comes with a set of words that best describe it, and we also extracted most related tweets associated with this topic

- Topic classification of LDA tweets: (manual)
  - Spam / hashtag
  - Misuse
  - Joke
  - Reporting of a specific case involving zika
  - General zika information
  - Advise about zika
  - Misinformation about zika
Timeline – Language Volume
Timeline – Country distributions
Future plans

• Improve LDA results
• Find better way to extract rumors and misinformation from the dataset