Tweetycs

Understanding Real-Time Discussion of Health Issues on Twitter Through Visual Analytics (Ongoing Project)

Amir Haghighati

Insight Lab

April 10, 2019

ahaghig3@uwo.ca

Agenda

1. Introduction
   - Why Social Media?
   - Challenges
   - Opportunities
2. Background
   - Existing Research
   - Analysis
3. Proposed Method
   - Part One: Non Real-Time Analysis
   - Part Two: Real-Time Visual Analytics

Social Media Usage

- People gather health information from diverse mediums, including social media.

Social Media Usage

- People gather health information from diverse mediums, including social media.
- Social media allows us to explore conversations in a rapid fashion.

Social Media Usage

- People gather health information from diverse mediums, including social media.
- Social media allows us to explore conversations in a rapid fashion.
- Twitter is one of the largest social media platforms with more than 300 million monthly active users.

Social Media Usage

- People gather health information from diverse mediums, including social media.
- Social media allows us to explore conversations in a rapid fashion.
- Twitter is one of the largest social media platforms with more than 300 million monthly active users.
- The unrestricted access to opinions and large user base has made Twitter a source for the collection and dissemination of information for various domains including health.

Social Media Usage (Cont'd)

- Health organizations are using social media to:
  - promote healthy lifestyle choices,
  - identify disease outbreaks,
  - explore human behaviour, and
  - assess the public's perception of health issues

Social Media Usage (Cont'd)

- Health organizations are using social media to:
    - promote healthy lifestyle choices,
    - identify disease outbreaks,
    - explore human behaviour, and
    - assess the public's perception of health issues
- Individuals, news organizations, businesses, interest groups, and other groups also discuss health on Twitter.

Challenges of Social Media

- On any given day, over 500 million tweets are posted.

Challenges of Social Media

- On any given day, over 500 million tweets are posted.
- The sheer number of tweets, variety in quality of information, and identity of user accounts makes it harder to analyze public discourse on Twitter.

Challenges of Social Media

- On any given day, over 500 million tweets are posted.
- The sheer number of tweets, variety in quality of information, and identity of user accounts makes it harder to analyze public discourse on Twitter.
- There exist challenges for the public to improve their knowledge on a wide variety of health issues on Twitter.

Challenges of Social Media

- On any given day, over 500 million tweets are posted.
- The sheer number of tweets, variety in quality of information, and identity of user accounts makes it harder to analyze public discourse on Twitter.
- There exist challenges for the public to improve their knowledge on a wide variety of health issues on Twitter.
  - Following a particular health organization may be beneficial for learning about a specific health hazard.

Challenges of Social Media

- On any given day, over 500 million tweets are posted.
- The sheer number of tweets, variety in quality of information, and identity of user accounts makes it harder to analyze public discourse on Twitter.
- There exist challenges for the public to improve their knowledge on a wide variety of health issues on Twitter.
  - Following a particular health organization may be beneficial for learning about a specific health hazard.
  - Obtaining a high-level understanding of social discourse on a wide variety of health issues, remains a challenge.

What is The Purpose of Utilizing Social Media?

For Understanding Public Discourse

What is The Purpose of Utilizing Social Media?

For Understanding Public Discourse

- A high-level understanding can:
  1. help addressing misinformation
  2. equip individuals with a better mental structure

  to assess how health issues are discussed.

What is The Purpose of Utilizing Social Media?

For Understanding Public Discourse

- A high-level understanding can:
  1. help addressing misinformation
  2. equip individuals with a better mental structure

  to assess how health issues are discussed.
- Health professionals and social scientists can use this lens to:
  1. better understand public perception of health issues and
  2. determine how to better utilize Twitter for health promotion

Existing Research

- Previously, a combination of manual content annotation and computational models have been used to analyze the sentiment of discourse of health issues (e.g., marijuana usage, perception of H1N1 vaccine).

Existing Research

- Previously, a combination of manual content annotation and computational models have been used to analyze the sentiment of discourse of health issues (e.g., marijuana usage, perception of H1N1 vaccine).
    - Sentiment analysis is concerned with the use NLP and computational linguistics to identify and extract subjective information from human language.

Existing Research

- Previously, a combination of manual content annotation and computational models have been used to analyze the sentiment of discourse of health issues (e.g., marijuana usage, perception of H1N1 vaccine).
    - Sentiment analysis is concerned with the use NLP and computational linguistics to identify and extract subjective information from human language.
- Some existing works also used machine learning techniques to classify tweets based on user description, genre, theme, and relevance to the topic of discussion (e.g., e-cigarettes, breast-cancer, dental pain).

Existing Research

- Previously, a combination of manual content annotation and computational models have been used to analyze the sentiment of discourse of health issues (e.g., marijuana usage, perception of H1N1 vaccine).
  - Sentiment analysis is concerned with the use NLP and computational linguistics to identify and extract subjective information from human language.
- Some existing works also used machine learning techniques to classify tweets based on user description, genre, theme, and relevance to the topic of discussion (e.g., e-cigarettes, breast-cancer, dental pain).
- Existing research has focused predominantly on understanding one or two health topics.

Our Aim

What can we do?

- Existing research has focused predominantly on understanding one or two health topics.

Our Aim

What can we do?

- Existing research has focused predominantly on understanding one or two health topics.
- Building a tool to provide insight into a variety of health issues is possible through a visual analytic perspective.

Our Aim

What can we do?

- Existing research has focused predominantly on understanding one or two health topics.
- Building a tool to provide insight into a variety of health issues is possible through a visual analytic perspective.
  - Visual Analytics (VA) enhances the understanding of data by combining computational models and techniques (e.g., machine learning techniques) with interactive visualizations.

Our Aim

What can we do?

- Existing research has focused predominantly on understanding one or two health topics.
- Building a tool to provide insight into a variety of health issues is possible through a visual analytic perspective.
  - Visual Analytics (VA) enhances the understanding of data by combining computational models and techniques (e.g., machine learning techniques) with interactive visualizations.
  - Visual Analytics $\neq$ some charts!

Our Aim

What can we do?

- Existing research has focused predominantly on understanding one or two health topics.
- Building a tool to provide insight into a variety of health issues is possible through a visual analytic perspective.
    - Visual Analytics (VA) enhances the understanding of data by combining computational models and techniques (e.g., machine learning techniques) with interactive visualizations.
    - Visual Analytics $\neq$ some charts!
- But how can we combine machine learning with visualization and interaction?

Our Aim (Cont'd)

Questions We Want to Answer

- Who talks about what?
- Is there a theme for these discussions? How many themes are out there?
- What is purpose of these discussions?
- Is there any relationship between a specific health issue and the sentiments from different sides of the discussions (e.g., media corporations and government officials)?
- ...!?

Part One

Data Collection

- Using Tweepy (a Twitter API) and 117 search terms, a collection of 535,973 unique English language tweets over a 1 month period was curated.

Part One

Data Collection

- Using Tweepy (a Twitter API) and 117 search terms, a collection of 535,973 unique English language tweets over a 1 month period was curated.
    - Search terms: causes identified by the Institute for Health Metrics and Evaluation (IHME)

Part One

Data Collection

- Using Tweepy (a Twitter API) and 117 search terms, a collection of 535,973 unique English language tweets over a 1 month period was curated.
  - Search terms: causes identified by the Institute for Health Metrics and Evaluation (IHME)
  - Metadata about the tweeter (account description, number of followers and following accounts, verification status) was also retrieved.

Part One

Data Collection

- Using Tweepy (a Twitter API) and 117 search terms, a collection of 535,973 unique English language tweets over a 1 month period was curated.
    - Search terms: causes identified by the Institute for Health Metrics and Evaluation (IHME)
    - Metadata about the tweeter (account description, number of followers and following accounts, verification status) was also retrieved.
- Retrieved data was stored in a MongoDB database.

Part One (Cont'd)

Analysis - Sentiment and Categories

- Initially, AlchemyAPI (now acquired by IBM) was used for sentiment analysis: each tweet got a sentiment score in the range (-1,1).

Part One (Cont'd)

Analysis - Sentiment and Categories

- Initially, AlchemyAPI (now acquired by IBM) was used for sentiment analysis: each tweet got a sentiment score in the range (-1,1).
- Based on previous research and an analysis of 500 sample tweets:
    - five content themes:
        1. Educational
        2. Fundraising
        3. Personal
        4. Promotional
        5. Unrelated

Part One (Cont'd)

Analysis - Sentiment and Categories

- Initially, AlchemyAPI (now acquired by IBM) was used for sentiment analysis: each tweet got a sentiment score in the range (-1,1).
- Based on previous research and an analysis of 500 sample tweets:
    - five content themes:
        1. Educational
        2. Fundraising
        3. Personal
        4. Promotional
        5. Unrelated
    - and six user categories:
        1. Businesses
        2. Celebrities
        3. Interest Groups
        4. Media
        5. Official Agencies
        6. General Public

Part One (Cont'd)

Analysis - Model Construction and Model Selection

- Classification of tweets and the users into the specified categories was done by variations in classification techniques:
  Toggling the inclusion of a specific feature in training models using bag of words and linear support vector classifiers.

Part One (Cont'd)

Analysis - Model Construction and Model Selection

- Classification of tweets and the users into the specified categories was done by variations in classification techniques:
  Toggling the inclusion of a specific feature in training models using bag of words and linear support vector classifiers.
- Taking only *'description'* of the tweeter into consideration and running 100 experiments on 3000 labeld tweets (20:80 - train:test)
  $\rightarrow AverageAccuracyRate = 86.86\%$ in classifying users.

Part One (Cont'd)

Analysis - Model Construction and Model Selection

- Classification of tweets and the users into the specified categories was done by variations in classification techniques:
  Toggling the inclusion of a specific feature in training models using bag of words and linear support vector classifiers.
- Taking only *'description'* of the tweeter into consideration and running 100 experiments on 3000 labeld tweets (20:80 - train:test)
  $\rightarrow AverageAccuracyRate = 86.86\%$ in classifying users.
- Taking *'text'*, *'count of keywords'*, and *'user verification status'* of the tweets into consideration and running 100 experiments on the same set
  $\rightarrow AverageAccuracyRate = 81.44\%$ in classifying tweets.

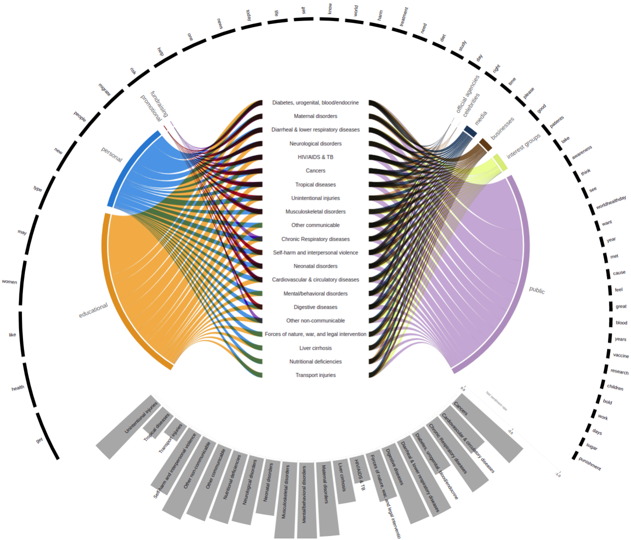
Part One (Cont'd)

Analysis - Model Construction and Model Selection

- Classification of tweets and the users into the specified categories was done by variations in classification techniques:
  Toggling the inclusion of a specific feature in training models using bag of words and linear support vector classifiers.
- Taking only *'description'* of the tweeter into consideration and running 100 experiments on 3000 labeld tweets (20:80 - train:test)
  $\rightarrow AverageAccuracyRate = 86.86\%$ in classifying users.
- Taking *'text'*, *'count of keywords'*, and *'user verification status'* of the tweets into consideration and running 100 experiments on the same set
  $\rightarrow AverageAccuracyRate = 81.44\%$ in classifying tweets.
- Removing unrelated tweets → 416,900 tweets remained.

Part One (Cont'd)

Visualization



Part Two

Where is The Analytics?

- Different supervised/unsupervised ML techniques → Different utility for each user

Part Two

Where is The Analytics?

- Different supervised/unsupervised ML techniques → Different utility for each user
- Online streams of tweets → Need for stream processing and asyncronous programming

Part Two

Where is The Analytics?

- Different supervised/unsupervised ML techniques → Different utility for each user
- Online streams of tweets → Need for stream processing and asyncronous programming
- Different possibilities for users → Customized utilities through interaction with ML techniques and visualizations

Part Two

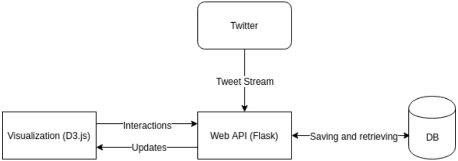
Where is The Analytics?

- Different supervised/unsupervised ML techniques → Different utility for each user
- Online streams of tweets → Need for stream processing and asyncronous programming
- Different possibilities for users → Customized utilities through interaction with ML techniques and visualizations
- Increasing epistemic utility for the user in order to understand the public discourse in the context of his/her interest.

Part Two (Cont'd)

Ongoing Framework

- A Python API is subscribing to streams and saving the tweets in the database.
- The API incorporates several machine learning techniques and these techniques are being executed with respect to the old data and the new incoming chunks of data.
- The user who is interacting with the Visualization will receive the updates on tweets and would be given the option of choosing the result of different ML techniques.



Thank you!

Questions?

Accuracy Rate for User Category Model Construction

Model	Avg. Acc. Rate (%)
A1: description	86.86
B1: description + screen name	79.83
C1: description + name + influence score	79.84
D1: description + name + influence score + verified	79.75

Accuracy Rate for Tweet Theme Model Construction

Model	Avg. Acc. Rate (%)
A2: tweet	80.99
B2: tweet + reserved keywords	81.09
C2: tweet + verified	81.14
D2: tweet + reserved keywords + verified	81.44