Pandemics in the Age of Twitter:
A Content Analysis of the 2009 H1N1 Outbreak

by

Cynthia Mei Chew

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Department of Health Policy, Management and Evaluation
University of Toronto

© Copyright by Cynthia Mei Chew 2010
Pandemics in the Age of Twitter: A Content Analysis of the 2009 H1N1 Outbreak

Cynthia Mei Chew
Master of Science
Department of Health Policy, Management and Evaluation
University of Toronto
2010

Abstract

This thesis reports on the use of Twitter during the 2009 H1N1 pandemic to explore its use as an “infoveillance” approach complementary to traditional surveys and content analysis. This study aimed to: 1) report on the use of “H1N1” versus “swine flu”, 2) conduct a qualitative analysis of tweet content, and 3) assess the feasibility of Twitter as a real-time content, sentiment, and public attention trend tracking tool.

A manual content analysis of tweets revealed that H1N1 resources were the most commonly shared. Few tweets provided inaccurate information. News websites were the most popular resources while official agencies were rarely referenced directly. Our automated analysis correlated well with manual results and showed that Twitter activity was influenced by external events.

This study describes the character and quality of Twitter communications during the H1N1 pandemic, and illustrates the potential of using social media to conduct real-time “infodemiology” studies for public health.
ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Dr. Gunther Eysenbach for his invaluable guidance, insight, and expertise.

I greatly appreciate Dr. Jennifer Keelan for helpful discussions while serving on my thesis committee.

I am grateful to the Canadian Institutes for Health Research for their financial support.

I thank Claudia Lai for her assistance with coding.

I thank my parents, Sam and Shirley, my sisters, Catherine and Carolyn, and my wonderful friends for their love and support.

And lastly, I thank my husband, Charles, for his encouragement, prayers, and unconditional love.
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ ii
ACKNOWLEDGEMENTS .................................................................................................................. iii
TABLE OF CONTENTS .................................................................................................................. iv
LIST OF TABLES ............................................................................................................................ viii
LIST OF FIGURES ........................................................................................................................... ix
LIST OF APPENDICIES ................................................................................................................ x
CHAPTER 1: INTRODUCTION ......................................................................................................... 1
CHAPTER 2: LITERATURE REVIEW ............................................................................................... 4
  2.1 Web 2.0 and Social Media ........................................................................................................ 4
  2.1.1 Web 2.0 and Apomediation ................................................................................................. 4
  2.1.2 Social Media ........................................................................................................................ 5
  2.1.2.1 Social Media Applications ............................................................................................... 6
  2.1.3 Internet & Social Media Usage ............................................................................................ 7
  2.1.4 Social Media Analytics ....................................................................................................... 10
  2.1.5 Summary ............................................................................................................................ 11
  2.2 Public Health Emergencies .................................................................................................... 12
  2.2.1 Crisis Risk Communication & the Social Amplification of Risk ....................................... 12
  2.2.2 The Social Amplification of Risk & Social Media ............................................................... 14
  2.2.3 Public Health 2.0 ................................................................................................................ 15
  2.2.3.1 Centers for Disease Control and Prevention (CDC) ..................................................... 16
  2.2.3.2 Federal Emergency Management Agency (FEMA) ....................................................... 18
  2.2.3.3 Public Health Agency of Canada (PHAC) ..................................................................... 19
  2.2.3.4 The Red Cross .............................................................................................................. 19
  2.2.4 The Public’s use of Social Media & Web 2.0 in Emergencies ......................................... 20
  2.2.4.1 2003 SARS Epidemic ..................................................................................................... 20
  2.2.4.2 December 26, 2004 Indian Ocean Earthquake & Tsunami ......................................... 21
5.6 Automated Analysis........................................................................................................ 86
5.7 Methodological Limitations & Advantages.................................................................. 86
5.8 Practical Implications for Public Health................................................................. 90
5.9 Research Objectives Revisited.................................................................................... 91
5.10 Future Directions...................................................................................................... 92
5.11 Conclusion................................................................................................................ 93

REFERENCES.................................................................................................................... 95
LIST OF TABLES

Table 4-1: Descriptions and Examples of Content Categories...................................................... 63
Table 4-2: Descriptions and Examples of Qualifier Categories.................................................... 64
Table 4-3: Descriptions and Examples of Link Categories........................................................... 65
Table 4-4: Content, Qualifiers, and Links of Manually Coded H1N1 Tweets............................... 66
Table 4-5: Descriptions and Examples of Automated Coding Search Patterns............................ 68
Table 4-6: Content and Qualifiers of Automatically Coded H1N1 Tweets.................................... 69
Table 4-7: Automated Coding Trends Over Time........................................................................ 70
Table 4-8: Correlations between Manual and Automated Coding............................................. 70
Table 4-9: RT Trends over Time................................................................................................... 79
LIST OF FIGURES

Figure 2-1: The Social Amplification of Risk Framework (adapted from Renn, 1991).............. 13

Figure 2-2: Social Amplification of Risk Framework with emphasis on social media users....... 15

Figure 3-1: Research design overview.......................................................................................... 55

Figure 4-1: Absolute number (lines) and relative percentage (bars) of tweets containing the keywords H1N1 or Swine Flu, between May and December 2009.................................................. 62

Figure 4-2: Non-linear pattern of tweeted misinformation identified via manual coding........ 67

Figure 4-3: Example scatterplots of manual versus automated coding proportions............... 71

Figure 4-4: The proportion of tweets expressing humour.......................................................... 72

Figure 4-5: The proportion of tweets expressing frustration...................................................... 72

Figure 4-6: The proportion of tweets sharing personal experiences............................................ 73

Figure 4-7: The proportion of tweets expressing concern............................................................ 73

Figure 4-8: The proportion of tweets expressing relief................................................................. 74

Figure 4-9: The proportion of tweets expressing misinformation............................................... 74

Figure 4-10: The proportion of tweets expressing downplayed risk.......................................... 75

Figure 4-11: The proportion of tweets sharing personal opinions & interest............................ 76

Figure 4-12: The proportion of tweets sharing questions............................................................ 76

Figure 4-13: The proportion of tweets sharing resources............................................................ 77

Figure 4-14: Comparison of content proportions between RTs and nonRTs............................ 77

Figure 4-15: Comparison of qualifier proportions between RTs and nonRTs........................... 78

Figure 4-16: Comparison of link proportions between RTs and nonRTs................................. 78
CHAPTER 1: INTRODUCTION

“In the era of the 24-hour news cycle, the traditional once-a-day press conference featuring talking heads with a bunch of fancy titles has to be revamped and supplemented with Twitter posts, YouTube videos and the link. The public needs to be engaged in conversations and debate about issues of public health, they don’t need to be lectured to.”

-Andre Picard, 2010

Public health agencies do not act in a void, but rather are part of a larger feedback loop that includes both the media and the public. The Social Amplification of Risk Framework postulates that psychological, social, cultural, and institutional factors interact with emergency events and thereby intensify or attenuate risk perceptions (Kasperson et al., 1988). A feedback loop connects these parties and allows authorities to tailor the future health messages and actions based on responses by the media and the public (Prue et al., 2003). Traditionally, print media, TV and radio are the major transmitters of information from public health agencies to the public, and play a large role in risk intensification and attenuation. However, with the increasing adoption of the internet and social media, the way in which the public receives information and their preferred modes of communication are changing.

With the rise of Web 2.0 (i.e., the “participatory” Web) and social media, and the resulting proliferation of user-generated content, the public potentially plays a much larger role in all stages of knowledge translation, including information generation, filtering, and amplification. As a result, for public health professionals, it is increasingly important to establish a strong feedback loop and monitor the public response and perceptions during emergency situations in order to examine the effectiveness of knowledge translation strategies and tailor future communications and educational campaigns.

Surveys (polling) are the traditional methods for public health officials to understand and measure public anxiety, knowledge, rumours, misinformation and sentiments, and behavioural responses. Such measures are crucial for public health planning (e.g., to predict vaccination demand). For example, public polling was used during the 2001 anthrax attacks in order to
determine what precautions the public was taking in response to these threats (Blendon et al., 2003). In this study, the authors emphasized the need for rapid-turnaround surveys as these best capture the changes in attitudes and behaviour influenced by specific events, and thereby produce the most relevant information for agencies to act upon. Unfortunately, short-duration surveys come with many barriers as it can take precious time for agencies to gather resources, funding, and survey instruments in order to conduct polling (Blendon et al., 2003). Content analysis of public/media publishing and communication activity is also a common method of gaining qualitative data on public/media attitudes, attention, and behaviour. However, manual content analysis is typically time-consuming as documents must be reviewed line-for-line and coded using a systematic classification scheme.

But with increasing use of the Internet and social media such as Twitter (www.twitter.com), new “infoveillance” methods, for example mining, aggregating, and analysing such data in real-time, become available (Eysenbach, 2009). Twitter is particularly suitable for longitudinal text mining and analysis. The brief (up to 140 characters) text status updates (posts or tweets) that users openly communicate with friends and followers contain a wealth of data as they share their thoughts and feelings, disclose what they are currently doing, share links to what they are reading (news articles, blogs, etc.), and voice their comments, opinions or interpretations of articles or events. Mining such data allows for a snapshot of the public’s opinions and behavioural responses to be taken almost instantaneously and longitudinal tracking allows for identifying changes in opinions or responses. Various approaches could be used, such as counting the number of keywords used or ranking the most cited URLs can provide quantitative and qualitative insights into what people are reading and discussing. It is possible that searching for positive or negative emoticons or other keywords may even provide insights into the “collective mood” of users over time. In addition to quantitative trend analysis, the method also allows a qualitative exploration on the likely reasons for why sudden changes have occurred (for example, due to a widely read news report) and may provide an indication of what is holding the public’s attention (Ripberger, 2010).

Apart from being a potentially useful research and surveillance tool, social media are increasingly important communication tools for public health professionals. For example, the
Centers for Disease Control (CDC) has begun using a Twitter feed as part of their overall dissemination strategy, as has several Canadian agencies, including the Public Health Agency of Canada (PHAC). The realms of social networking and Web 2.0 have only been established in the past few years and as such their use has not been thoroughly investigated in large-scale emergency or pandemic situations. Much of the previous communication work on biological threats such as the anthrax scare of 2001 and SARS in 2003, focused on describing, analyzing, and evaluating traditional media and government agency coverage (e.g., Prue et al., 2003; Mebane et al., 2003; Drache et al., 2003).

This outbreak marks the first instance where a global pandemic has occurred in the age of Web 2.0, and adoption is increasing. Using social media, users can share and/or comment on traditional media stories and create content. This event presents an opportunity to investigate the potential role that these technologies can play in a state of emergency for public health agencies.

**Research Objectives:**

This report of H1N1-related communication patterns is part of a broader project where health-related information-seeking/sharing behaviour is tracked on the Internet and in social media to understand the role of social media in public health emergencies and inform public health education and communication initiatives. The overall goals of this study are to:

1) Track the public’s terminology use during the pandemic as a measure of knowledge translation effectiveness, i.e., the use of recommend H1N1 versus colloquial swine flu terminology;
2) determine what Twitter users are communicating on Twitter, how they are expressing themselves, what information sources they are using, and if this content is changing over time;
3) explore the use of Twitter as a real-time content, sentiment, and public attention trend analysis and tracking tool.
CHAPTER 2: LITERATURE REVIEW

2.1 Web 2.0 and Social Media

2.1.1 Web 2.0 & Apomediation

“Web 2.0” is an umbrella term used to describe the current generation and state of the internet today. In contrast with earlier versions of the web (Web 1.0 or the “static” web), today’s internet focuses on accessible user-generated content, publication, and collaboration instead of top-down information dissemination by business, organizations, and governments (Boulos & Wheeler, 2007). This “social”, “participatory” or “democratic” web encourages dynamic and flexible interactivity, interoperability, and sharing. This set of new technologies and approaches allowed for many new online activities such as blog, wikis, mashups, folksonomies and social networking (Boulos & Wheeler, 2007).

A growing number of developers and researchers have experimented with Web 2.0 technologies and created many health-related tools, services, and applications for both health providers and consumers (e.g., Frost & Massagli, 2008). Eysenbach (2008) proposed that the term “Medicine 2.0” be used to describe this next generation of online medicine and emphasized 5 major aspects of the concept: social networking, (user) participation, collaboration, openness, and apomediation.

The term *apomediation* refers to the search strategy where users bypass formal intermediaries (traditional “gatekeepers” or middlemen) and instead of acting completely independently, are guided by peers and web tools to find credible and relevant information (Eysenbach, 2008). However, unlike gatekeepers who give access to information, these apomediaries are not a prerequisite to obtaining the resource or service. With the popularization of social networks where content is passed along by users, any user can act as an apomediary or informal knowledge broker, although the credibility and quality of information will vary. The medical importance of apomediation in social networks is emphasized by findings that online peer feedback can influence health-related decision making (Lau & Coiera, 2008). As a result, it is
crucial not only to provide health resources but also the appropriate network that can direct users to relevant and trustworthy information (Lau & Coiera, 2008).

2.1.2 Social Media

Building on Web 2.0 advancements, social media can be defined as:

"a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content"
-Kaplan & Haenlein, 2010

In essence, social media allow for the rapid exchange of ideas and information using internet-based applications that rely chiefly on social interaction, user participation, and user-generated content. Social media can be seen as allowing the “democratization of information” as users are encouraged to generate, comment on and propagate news or information that they find useful or interesting. This user-driven process leads to the amplification of what the masses “like” and guides even more readers towards the original or similar content. Users can also find and connect to those who share their interests. Viral marketing or dissemination often plays on these features of social media to quickly spread information among networks and raise awareness. The use of viral marketing in health promotion appears to be a promising new way of communicating health messages (e.g., Freeman & Chapman, 2008; Gosselin & Poitras, 2008).

Due to public adoption and inexpensive production, businesses, governments, charities, and health organizations alike are beginning to utilize social media applications as part of their communications and branding strategy (examples will be presented later in the chapter). There are many different forms of social media, and these can be broadly grouped based on their primary functions: communication (blogs, social networking, micro-blogging), collaboration (wikis, social bookmarking, folksonomies), multimedia (photo, music, and video sharing), entertainment (virtual worlds), and reviews (aggregated reviews, community-built FAQs). SMS (short message service) text messaging and cellular phones are not technically a form of social media, but like computers are platforms on which numerous social media applications can be accessed.
2.1.2.1 Social Media Applications

While there are countless numbers of social media applications available, some have become more popular and widely used than others. The following section briefly introduces some of the most popular platforms used today.

Facebook
Originally intended as a social networking site for university students, Facebook has since expanded to allow anyone over the age of 13 with a valid e-mail address to join. Now in its 6th year of operation, Facebook has over 400 million registered users (Cutler, 2010) and a growing variety of features that allows for users to interact. Each individual has a profile page on which they can display a photo and other personal information. Profiles can be made publically searchable to make more social connections or “friends”. The profile page serves as a hub for “friends” to post messages, links, or media that might interest the user. The user can upload their own media and links to their profile page as well as provide status updates to notify their friends of their actions, whereabouts, and thoughts. Users can also set up special interest groups to share common interests, causes, or affiliations.

Flickr
Founded in 2004, Flickr is a widely used image and video storage and sharing site. As of October 2009, over 4 billion photos have been uploaded to the site (Flickr, 2009). Flickr allows for users to upload their photos, tag them with metadata, edit, organize, and share them with others. Various levels of privacy are available. Users can create special interest groups for sharing related media and can comment on and discuss photos if authorized to do so. One of the latest features, geotagging, allows users to indicate where on a map their photos were taken and browse maps to see what others have seen in the area.

YouTube
Since its launch in 2005, YouTube has become one of the most popular video sharing websites in the world and today it exceeds over 2 billion views a day (YouTube, 2010). On YouTube, a wide variety of user-generated amateur content and professional or official content is available
for viewing via Adobe Flash Video or HTML5. If cookies are enabled, YouTube can make recommendations to visitors based on their past viewing history. To comment on videos or upload their own, users must register. Video owners can annotate their videos to provide further information, personal opinions, or insights (comments in text boxes pop-up during the video). Subscribing to video feeds allows users to stay up-to-date with their favourite channels and follow the administrator’s activity.

**Twitter**

Twitter is a social networking and micro-blogging platform that allows users to send and read short messages (tweets) from devices such as cell phones (using SMS) or computers. These 140 character free-text messages can be posted as a stand-alone comment or directed to particular users (@username) and may contain content such as personal status updates, news briefs, or web links of interest. Users can subscribe or “follow” favourite personalities, friends or news feeds and vice versa to receive tweets automatically. Coming to prominence in recent years, it is now the third most used social network in the world (Kazeniac, 2009) and has 100 million users worldwide (Oreskovic, 2010). A number of Twitter text conventions have been created to aid user communication. Twitter authors use the @username convention in their posts to reference of direct a tweet to another user. Building on this notation, users can use the retweet notation, (RT @original-author), to pass on tweets of interest and give credit to the original author. The hashtag (#) notation allows for keyword tagging and tweet organization via topic. For instance, tweets containing the phrase “#H1N1” would be found when the term “H1N1” was searched or monitored for. The hashtag also allows users to quickly understand the topic or context of the tweet. The follow @username convention can be used to recommend other users by telling others to follow their posts.

### 2.1.3 Internet & Social Media Usage

Internet usage in 2010 is almost ubiquitous among both young American adults and teens (93%) and is growing among adults in general (74% of adults 18 and over) (Lenhart et al., 2010). Similarly, in Canada, approximately three-quarters of the population are connected to internet (Internet World Stats, 2009). The population is also becoming less tethered to wired
connections, with more than half of Americans using wireless internet on their laptops and on their cell phones (Lenhart et al., 2010). According to an extensive study conducted by Pew Internet (Lenhart et al., 2010), mobile phone usage has also increased, but is slowing. Today, 75% of American teens own a cell phone as do nearly all young adults aged 18-29 (93%). Cell phones are also being adopted at younger ages at increasing rates. In 2004, only 18% of 12 year olds had a personal cell phone. In 2009, this number has more than tripled (58%). In Canada, 7 out of 10 households own a cellular phone (Zamaria & Fletcher, 2007). Overall, the North American population is becoming more internet savvy and mobile, although barriers to penetration do exist.

The results of the most recent study of social media usage by online American teens and adults by Pew Internet (Lenhart et al., 2010) show that the adoption rates of many social media platforms are increasing, sometimes at the expense of older applications. For instance, blogging has decreased significantly among online teens (14%, down from 28% in 2006), while it is relatively steady among adults over 30 (7% in 2007, 11% in 2009). Social networking sites (SNS) such as Facebook and MySpace are continuing to grow in popularity. In 2009, 73% of teenagers had a profile page on a SNS, up from 55% in 2006. SNS adoption among adults has also increased, up to 47% from 37% in 2008. Of these SNS, Facebook was the most commonly used among adults and teens, 74% and 71%, respectively. MySpace was the second most popular (48% of adults, 66% of teens). Conversely, Twitter does not have a mass following among teens (8%), although 1 in 5 adults use Twitter or a similar service (Lenhart & Fox, 2009). However, the US Health Information National Trends Study found that social media use was not evenly distributed across all age groups (Chou et al., 2009).

A growing number of these online “netizens” are using the internet to look up health information online. In 2000, 25% of American adults used the internet to find health information. This number has more than doubled since then (61%) (Fox & Jones, 2009). This process of searching for health information is also incredibly social. More than half of all health inquiries are typed in on behalf of someone else and two-thirds of e-patients discuss their findings with friends or family (Fox & Jones, 2009). E-patients actively seek out other patients’ experiences or thoughts on their health situation or treatment in these “apomediated environments” (Eysenbach, 2008),
but the majority do not post new content (Fox & Jones, 2009). Social media and web 2.0 plays an important role in this process, with 53% of e-patients consulting Wikipedia and 39% using Facebook in their queries. Twitter is less frequently used for this purpose; only 12% of e-patients currently use the service to obtain health information or updates (Fox & Jones, 2009).

The quality and credibility of health information on the internet is particularly important (e.g., Eysenbach & Kohler, 2002) as 6 out of 10 e-patients report that their searches have provided information that impacted their own health or the health of someone that they care about in one way or another (Fox & Jones, 2009). However, most health consumers are not particularly aware of how to evaluate internet information sources (Eysenbach & Kohler, 2002). Yet, thus far, it appears that internet health information has not been tremendously harmful for the majority of cases; the amount of people harmed due to such misinformation is relatively low (3%) and has not increased since 2006, according to a Pew Internet study (Fox & Jones, 2009).

**Twitter Usage**

According to a study conducted by Cheng et al. (2009), Twitter is predominantly used by Americans and they account for 50.8% of all users. Pew Internet reported that approximately 19% of all online American adults are using Twitter or a similar application (Lenhart & Fox, 2009). And as of April 10, 2010, it is estimated that in the United States, 55% of users are female, 45% are between the ages of 18-34, 69% are Caucasian, 49% have less than a college degree, and 58% make over $60K a year (Quantcast, 2010). Although there are over 100 million users worldwide (Oreskovic, 2010), it has been reported that only 1 in 4 users tweet every month and the majority (75%) have tweeted fewer than 10 times total (Moore, 2009). Similarly, an analysis of 300,000 tweets found that a disproportionate number of tweets (90%) were produced by the top 10% of power users (Heil & Piskorski, 2009). In addition, not all users have the same visibility and reach on Twitter: approximately 50% of users have only 1-5 followers while 20% have no followers at all. While Twitter claims to have a large following, based on the data it appears that the service is primarily being used by a small subset of users.
2.1.4 Social Media Analytics

With the increasing growth of publicly available user-generated content, carefully crafted, official messages are no longer the only pieces of information that influence consumer decisions. As a result, companies and businesses have been quick to develop new market research methods to analyze publicly generated content and measure the effectiveness of social media campaigns. Social media analytics in this case often utilize methods such as social network analysis, machine learning, data mining, and natural language processing (Lawrence et al., 2010).

However, there is no commonly agreed upon set of metrics to measure campaign effectiveness or returns on investment and few social media marketers believe that such actions can be measured accurately (Solis, 2010; eMarketer, 2009; MarketingSherpa, 2010). The majority of companies are currently measuring the impact of their social media initiatives via the number of visitors and the size of their network (number of followers, etc.) (MarketingSherpa, 2010). Approximately half of companies surveyed used search engine ranking position, sentiment analysis, and quantity of internet commentary as additional metrics (Marketing Sherpa, 2010). In a study of public blogging, Melville et al. (2009) described methods that companies may use to address the challenges of evaluating the impact of social media campaigns. For businesses wishing to understand how the public reacts to their products or actions, important methodological issues include: 1) finding relevant blogs, 2) determining the sentiment expressed in posts, and 3) determining which bloggers are authoritative or influential. A small number of highly relevant blogs are manually identified and then used as models for text classification to identify other relevant blogs via keyword searches. Alternatively, researchers can use URL links between blogs to identify other related blogs, effectively using snowball sampling. This cross-referencing method avoids the difficulty that arises when search terms are ambiguous or have multiple meanings (i.e., a search for “Apple” may yield a technology company or fruit). Manually labelled documents and words can be used as models for automated sentiment detection programs to learn and expand on existing vocabularies. One way to measure the level of authority or influence of bloggers is to study and quantify the number of links to a specific blog. Link analysis can be performed using the Google-owned PageRank algorithm which calculates the authority of a blog based on the number and authority of other blogs that link to it or by using
more simple measures such as ranking blogs based on the number of times it has been cited over the past 30 days (e.g., www.blogpulse.com).

Much of Twitter’s user activity data is publically available (e.g., number of tweets/followers) and through the application programming interface (API), tweets can be automatically downloaded for analysis. As a result, several web-based Twitter analytics tools are available for public use. For example, Twist (www.trendistic.com) allows users to search for keywords and see the resultant frequencies graphed longitudinally. Twist also visualizes emerging Twitter topic trends. Another web application, Twitturly (www.twitturly.com) allows for the tracking of URLs on Twitter and display the most popularly shared links over set time frames. URL shortening services such as Bit.ly (www.bit.ly) also provide visitor statistics for their URLs. While many market research companies such as Sysomos and IBM provide Twitter and social media analytics to their clients, their methods are not publically available. At the time of this writing few academic papers that detailed the development of Twitter metrics were available. Several researchers have performed content and activity analysis of tweets sent during crises such as Hurricane Gustav and Ike (Hughes & Palen, 2009) and the Red River Flooding of 2009 (Starbird et al., 2010). These examples are detailed later in this chapter.

2.1.5 Summary

Web 2.0 and social media technologies have given rise to a new range of participatory applications that allow users to share information and communicate in novel ways. Rates of internet and mobile phone usage are high and continue to grow. At the same time users are using computer-mediated technologies to take a proactive role in learning more about their health. These new technological advances and communication behaviours have potential implications for how health agencies and governments transmit messages to the public.
2.2 Public Health Emergencies

2.2.1 Crisis Risk Communication & the Social Amplification of Risk

During times of crisis, it is necessary for governments and authorities to communicate with the public by broadcasting accurate, timely, direct, and relevant messages using a wide range of mediums such as television, radio, and print ads. This exchange of information regarding immediate health risks can be defined as crisis risk communication (Glik, 2007). Risk communication messages are typically designed based on empirical evidence from the fields of sociology, psychology, and communications and theoretical models (Covello et al., 2001; Wray et al., 2004; Glik, 2007).

One useful model, the Social Amplification of Risk Framework (SARF), is a population-level model that emphasizes the interconnectedness between message senders and receivers and the role each plays in message transmission (Kasperson et al., 1988). The basic assumption of SARF is that risk events capture few people’s attention, are mostly irrelevant, or have a very localized impact, unless communicated to society on a broad level (Renn, 1991). Drawing upon the metaphor of signal amplification from technical communications literature, SARF postulates that risk events interact with psychological, social, cultural, and institutional processes that either amplify (increase) or attenuate (decrease) risk perceptions (Kasperson et al., 1988). SARF can be specifically applied to describe the social processes of information-exchange (Pidgeon & Henwood, 2010). In SARF, there are 3 main groups in the amplification/attenuation process: sources, transmitters, and receivers (Figure 2-1).
As seen in Figure 2-1, sources may be individuals or groups who either personally experience the risk event or are primary information brokers (i.e., eyewitnesses, government agencies, scientists). Typically, these sources pass information to transmitters for dissemination, usually through press conferences, news briefs, or reports.

Transmitters of signals (messages) or “amplification stations” may be individuals, social units, or institutions. Traditionally, the media play the major role of information transmitters. These amplification stations will transform the original data before passing it along further. These transformations may include the reinterpretation or elaboration of the message, increasing or decreasing the volume of information about an event, or deciding what specific facts or events are amplified or attenuated (Kasperson et al., 1988). Risk amplification occurs when a fairly minor risk is given a disproportionate amount of attention in the media, thereby increasing the perceived level of risk. One example of risk amplification is the intense media attention on a few isolated shark attacks along the Florida coast in 2001. Although the number of shark attacks was less than in previous years, tourists were driven away by the increased perceived threat (McComas, 2006). Conversely, risk attenuation occurs when a serious threat receives less attention from the media and consequently, the public.
The general public are the primary message receivers. However, they are interconnected with both the sources and transmitters in a feedback loop. Public and media response and reactions to emergency situations can influence future communication and educational strategies of both sources and transmitters. Public polling and surveying (Blendon et al., 2003) and media monitoring (Prue et al., 2003) are often conducted for this reason. The public also play an important feedback role within their communities. For example, during the World Trade Center bombing in 1993, studies found that workers in larger offices took longer to evacuate than those in smaller offices as they required more time to come to a common understanding of the situation before leaving the premises (Aguirre et al., 1997), and during the Denver South Platte River flood in 1965, 60% of interviewees reported that they checked with family, friends, and neighbours to confirm evacuation notices that they had received before taking any action (Drabek et al., 1975).

2.2.2 The Social Amplification of Risk & Social Media

Traditionally, print media, TV and radio are the major mediums of information transfer from public health agencies to the public, and play a large role in risk intensification and attenuation. However, the role of the Internet and social media in public health emergencies and disaster scenarios is increasing.

Studies during the anthrax scare of 2001 have found that approximately 25% of the population turned to the internet for information in these circumstances (Pollard, 2003) with Internet users changing their behaviour as a result of information that they had found (Kittler et al., 2004). Similarly, during the 2003 SARS outbreak in Toronto, almost 50% of hospital patients used the internet to search for relevant health information (Rizo et al., 2005) and a variety of internet-based applications were created to facilitate information sharing (Eysenbach, 2003). After the September 11th attacks, one out of four internet users went online to supplement information from traditional media sources (Rainie, 2002) and used multiple sources in order to confirm the credibility and veracity of information that they had received (Rainie, 2003). Indicative of this media change, during the most recent public health emergency, respondents cited the internet as their most frequently used source of information for the 2009/10 H1N1 pandemic (Jones &
Salathe, 2009). The results of these studies indicate the importance of the internet as an alternative source of information and the active role that the public plays in information seeking.

In addition, with the rise of Web 2.0 and social media, and the resulting proliferation of user-generated content, the public potentially plays a much larger role in all stages of knowledge translation, including information generation, filtering, and amplification. Users share their activities on their social networking pages or microblogs, write about their experiences or thoughts on their blogs, and use social bookmarking to share links of interest. Web 2.0 and social media technologies facilitate this rapid, socially-mediated process of knowledge translation. As a result, for public health professionals, it is increasingly important to establish a strong feedback loop and monitor the public response and perceptions during emergency situations in order to examine the effectiveness of knowledge translation strategies and tailor future communications and educational campaigns. Figure 2-2 depicts SARF with an emphasis on the public use of social media and Web 2.0.

**Figure 2-2: Social Amplification of Risk Framework with emphasis on social media users**

![Social Amplification of Risk Framework](image)

### 2.2.3 Public Health 2.0

Recognizing the power of web 2.0 and social media, numerous public health and disaster relief organizations have adopted social media in various forms to create an online presence and
communicate with internet users. The following section briefly reviews the online activities and tools that are being utilized by major authorities and associations.

### 2.2.3.1 Centers for Disease Control and Prevention (CDC)

The CDC has an extensive list of social media tools to support both their emergency and ongoing health campaigns which can be accessed via a computer or mobile device (CDC, 2010a). The health authority uses several different blogs to address pertinent issues, for example health marketing, public health, and health science. Although the CDC provides authoritative information, they are aware that the public does receive health information from a variety of external sources, including other blogs. To manage this concern they host interactive webinars specific for bloggers (bloginars) to answer users’ questions and provide accurate information to inform their posts.

The CDC also participates in micro-blogging and has 3 separate Twitter accounts to update its followers. CDCflu is used to provide updates about pandemic H1N1, seasonal flu, and general influenza. The CDC_ehealth account is primarily aimed at health professionals who are interested in staying current with CDC’s social media campaigns and applications. Lastly, the CDCemergency profile allows followers to receive updates regarding emergency preparedness and response information from CDC and its federal partners. The CDCemergency account has the most followers (1,242,661) of the three and is ranked 168th in terms of followers (Twitaholic, 2010). However, the Twitter profiles of news organizations such as Good Morning America (1,704,920 followers, ranked 67th), the BBC (1,749,976 followers, ranked 60th), the New York Times (2,444,055 followers, ranked 24th), and CNN (3,099,003 followers, ranked 12th) are more popular among Twitter users (Twitaholic, 2010). In comparison, the five most popular Twitter accounts (Britney Spears, Ashton Kutcher, Ellen DeGeneres, Lady Gaga, and Barack Obama) have over 4,000,000 followers each (Twitaholic, 2010).

Online media such as podcasts, images, and video are also used to disseminate information. The CDC Streaming Health channel on YouTube currently has 98 public health advertisements and news clips available and has been viewed over 3,495,672 times. However, the number of
subscribers is relatively low (5,032) compared to popular music artists (Lady Gaga has 143,753 subscribers) and do-it-yourself YouTube stars (make-up artist Michelle Phan has 729,677 subscribers). This lack of subscriptions may indicate that YouTube users will only seek out CDC videos when necessary (i.e., in emergencies) and they are not part of their day-to-day viewing routine. In addition to YouTube, CDC has its own video-sharing site (CDC TV) that also allows users to download videos should they choose.

The CDC has profiles on three social networking sites: Facebook, MySpace, and Daily Strength (a specialized network where users provide emotional support and share their health experiences). All provide videos, news, and resources for emergency and ongoing health campaigns.

Virtual worlds are a growing online trend among youth and young adults, and CDC has moved into two online environments as an avenue to promote public health campaigns to this niche population. Like many health care agencies, research groups, and government bodies (see review by Beard et al., 2009), CDC has set up a designated area (“CDC Island”) to provide health information to virtual visitors via video, posters, and virtual health fairs. A virtual conference center, lab, and CDC representative avatar are also available for users to interact with and ask questions. In Whyville, a popular destination for tweens, CDC is working to raise awareness about seasonal flu and the importance of vaccination. To do so they have worked with the developers to create an online flu (ill avatars get red spots on their face and sneeze while chatting) which requires virtual vaccination to avoid. In the 6 weeks of vaccination clinics, CDC reported that “thousands of Whyvillians were vaccinated” (CDC, 2009a).

Specific CDC Social Media Campaigns
In 2009, the FDA recalled various peanut-containing products for Salmonella Typhimurium and worked with CDC to create awareness and distribute information about the recall via social media (CDC, 2009b). This strategy included a peanut recall-specific blog for updates, hosted bloginars, podcasts, and RSS feeds. The FDArecalls Twitter account posted updates and the CDC social networking profiles provided links to relevant resources. CDC Second Life Island offered bracelets and biohazard suits for avatars to raise awareness and virtual microscopes for
avatars to examine the bacteria. Two specific widgets were created for users to add to their webpage or blog; one that enabled searching for peanut product recalls and another that displayed a map of the US with reported cases of salmonella.

Commencing in April 2009, CDC partnered with the US Department of Health and Human Services (HHS) to promote information dissemination using social media tools regarding the 2009 pandemic H1N1 outbreak (CDC, 2009c). The CDC YouTube channel played a larger role in this campaign and uploaded several videos regarding vaccination safety and demonstrating good hygiene procedures. Six H1N1 widgets were developed for specific audiences. Three were broad in scope (to test users flu IQ, give H1N1 prevention information, and the latest government updates), two were created to communicate updates to schools and students (higher education and grades K-12), and one distributed quick facts to clinicians. In September, CDC piloted a 3-month text messaging service dedicated to providing H1N1 mobile updates (CDC, 2009d). To subscribe, individuals could text “HEALTH” to a designated number and receive approximately 3 updates per week. Both the CDCflu and Flu.gov (HHS administered) Twitter feeds provided updated pandemic information. However, it is unknown how effective these tools were and how many people adopted them.

2.2.3.2 Federal Emergency Management Agency (FEMA)

To support emergency preparedness and response, FEMA has integrated a social media toolbox, albeit less extensive than the CDC (FEMA, 2009). The official Twitter profile, FEMAInFocus, provides updates nation-wide and has 12,390 followers (ranked 26,598th overall) (Twitaholic, 2010). Ten region-specific accounts are also in place to distribute more specific information to affected areas. The FEMA YouTube stream currently has 336 videos with 2,278 subscribers and 184,168 views. These videos primarily show how FEMA works to prepare for, respond to, and help communities recover from disasters. Similar to CDC, FEMA also has their own media website to share videos, photos, and audio files of public health information, news, and recent emergency efforts. A FEMA Facebook account provides similar information. Most recently, the home department of FEMA, the Department of Homeland Security (DHS), began experimenting with social media monitoring to follow recovery, response, and rebuilding activity in the
aftermath of the 2010 Haiti earthquake (DHS, 2010). The Haiti Social Media Disaster Monitoring Initiative aggregates and analyzes text on social networking sites, blogs, public websites, and message boards and disseminates findings that contribute to situational awareness and recovery efforts (DHS, 2010).

2.2.3.3 Public Health Agency of Canada (PHAC)

The PHAC has implemented a few select forms of social media in their public knowledge translation strategy (PHAC, 2009). PHAC has two Twitter feeds to provide updates in both of the official languages of Canada (English and French); each has 2,135 and 377 followers, respectively. However, neither feed is largely followed; the English PHAC feed ranks 269th in followers from Canada and 56,411th worldwide. PHAC uses its Facebook profile to provide news, health quizzes, FAQs, interviews, and guidelines for other users to browse. PHAC also has a YouTube channel which it uses to distribute news and a few select TV public health announcements. Currently they have uploaded 14 unique videos (each video is uploaded in both English and French), which have been viewed 6,631 times. Since launching in April 2008 they have gained 48 subscribers. In addition, both RSS feeds and widgets are available for live updates.

2.2.3.4 The Red Cross

As a non-profit charity and relief organization, the Red Cross has similarly utilized social media to maintain an online presence and update their audience with relevant information. However, they have also employed these technologies to raise funds and support (Red Cross, 2010). The Red Cross currently manages a Facebook “Cause page” which allows users to donate money and recruit others in their social network to join the cause. Top recruiters and fundraisers gain online social recognition for their efforts. Thus far, almost 70,000 Facebook users have joined and raised over $40,000. The Red Cross also has profiles on Social Vibe, Good2gether, and Ammado, all of which allow for both social networking and fundraising. Their most recent and highly publicized social media fundraising campaign was mounted in response to the 2010 Haiti earthquake. A Twitter message informed users in the various networks that they could donate
$10 to the Red Cross relief effort just by texting a short message; within a few short days over $3 million dollars were raised (Siegel, 2010).

2.2.4 The Public’s Use of Social Media & Web 2.0 in Emergencies

In times of crises and emergency, the members of the public and communities are often times the first to react, respond, and mobilize in order to help others in need (Palen et al., 2010). With the advent of web 2.0 and social media technologies, both bystanders and victims can and have been using these tools to communicate, document (i.e., citizen journalism), and rally aid in innovative ways. The following section presents several crisis cases that highlight these applications and their evolution over time.

2.2.4.1 2003 SARS Epidemic

The 2003 SARS epidemic in China occurred just prior to many of today’s existing social media platforms (i.e., Twitter, Facebook). Instead, cell phones played a major role in public communication during this health emergency. Due to China’s strict censorship policy little information regarding an increasing number of “atypical pneumonia” cases was released, however, Gordon (2007) suggested that the people of Guangdong province were aware of SARS and the potential problem before the mainstream media as the number of text messages tripled during the days leading up to the Chinese government’s official report to the WHO on February 11th, 2003. Following February 14th, a media and internet blackout on SARS was enforced across China and news providers did not report on or acknowledge the existence of the disease (Dahong, 2003). Without any means of acquiring or verifying information, the public began to circulate texts regarding SARS outbreaks, folk remedies (most of which were inaccurate, e.g., drinking teas and vinegars), and rumours (Dahong, 2003). The Hong Kong government was quick to arrest one offending teenager and sent out mass texts with official messages to quell the rumour (Gordon, 2007). Humourous and satirical text messages were also spread to voice opinions and provide an outlet for public discourse in China. Cell phone applications were also built to help the public battle SARS. Sunday Communications, a cell phone service provider,
allowed subscribers to receive alerts by text if they were within one kilometre of an infected building in Hong Kong (Eysenbach, 2003).

In other cases, dissatisfied computer-savvy Chinese citizens created independent websites listing areas of suspected or confirmed SARS cases (Jardin, 2003). One of the most popular grassroots sites, “SoSick.org”, was originally released only to family and friends of the administrators. However, after an overwhelming response of 200,000 hits within 2 days the site was opened to the public and quickly picked up by the media (SoSick.org, 2003). SoSick.org and others like it eventually pressured the government to release their infected areas list and provide daily updates (Jardin, 2003).

2.2.4.2 December 26, 2004 Indian Ocean Earthquake & Tsunami

Mobile phone technology, blogging, and photo-sharing were at the forefront of the Indian Ocean disaster. Cell phones with cameras, then a novel technology, were widely used to capture images of the devastation and citizens shared them with the world before the mainstream media could respond (Skillings, 2005). Liu et al. (2008) reported that the tsunami was the first instance of disaster-related activity on Flickr. Over 7 photo groups were created to share news, strengthen the community, document history, educate distant observers, and rally for aid (Liu et al., 2008). Mobile phones were also heavily used for texting for help and locating survivors as phone landlines were down and voice calls were often dropped due to high bandwidth use (Gordon, 2007).

Public blogs also played an unprecedented role. The “Southeast Asia Earthquake & Tsunami blog” was launched by 3 individuals to provide aid, news, and information about family members to affected peoples. The blog also allowed visitors to post their needs or what help they could offer. A list of confirmed deaths, image galleries, and links to aid agencies were also constantly updated. The blog was so successful that it reached 1.1 million hits within 10 days of its launch; a worldwide blogging landmark (Sharma, 2005).
2.2.4.3 August 29, 2005 Hurricane Katrina

When the Category 5 Hurricane Katrina hit the US Gulf Coast, Louisiana and Mississippi took the brunt of the damage. Hundreds of thousands were displaced and at least 1,400 people lost their lives (Louisiana Department of Health and Hospitals, 2006). The storm severely damaged communications infrastructures and caused widespread power failures. The resulting devastation left many relief and federal organizations overwhelmed. However, a large number of grassroots efforts such as “Katrinahousing.net” by the University of Michigan and “Craigslist Katrina Relief” emerged to provide aid, housing, necessities, and employment to those affected (Hughes et al., 2008). Torrey et al. (2007) observed the giving activity on 2 types of online communities: large forums such as “Craigslist Katrina Relief” and “Nola.com” and small or personal blogs. Their content analysis found that while small blog communities had more centralized authority (e.g., a single group administrator with a clear role) and were more immediately successful in managing information, planning, and developing trust between members, this activity was hard to sustain. On the other hand, the activity of large forums took longer to develop social norms, trust, and standard procedures, but remained active even 6 months after the disaster, long after smaller blog activity had ceased. The authors suggest that the theory of weak ties (Granovetter, 1973) may play a role in the success of larger networks as members are more likely to come from a wide range of backgrounds, with different skill sets, information, and expertise. However, it is notable that the smaller communities were more adept at creating a sense of group identity and offered more emotional support than larger forums.

Similar to the Indian Ocean tsunami, citizen photojournalism during Hurricane Katrina was also widespread. Liu et al. (2008) reported that on Flickr 7 major photo groups arose, the largest with 600 members and almost 7000 photos. This was the first disaster where more formalized posting procedures were attempted by group administrators, but these were met with limited success. For example, on one group photo tagging instructions such as “KatrinaMissing”, “KatrinaFound”, “KatrinaOkay” were issued to try and create a database of survivors, victims, and missing persons, but the movement was not adopted by members (Liu et al., 2008). While the effort had good intentions, the authors speculate that Flickr lacked the critical mass to serve such a purpose.
To understand how victims used technology during both the tragedy and aftermath, Shklovski et al. (2010) conducted longitudinal interviews and surveys with surviving New Orleans musicians. Survivors reported that cell phones were vital during that period, with text messaging being the most important way to communicate as phones could not receive calls due to damaged telecommunications receivers. Many relied on their laptops, library computers, and internet cafes to e-mail family and friends in attempts to locate them or provide status updates. Several tech-savvy participants created safe lists for fellow musicians to check-in and provide information about their status and location. Interviewees who used the service found that it created a sense of community and hope. Many reported to be frustrated and critical of mainstream media reporting as the news coverage was not specific enough for their needs and was often times sensationalized. To deal with the information dearth, locals turned for information elsewhere and actively worked to generate and disseminate accurate information. “Nola.com”, a local newspaper, created discussion boards that were neighbourhood specific and provided maps and satellite images. “Sciponius.com”, a visual wiki, allowed users to annotate maps with information about the physical environment and the conditions of buildings, highways, and other public infrastructures (Palen et al., 2007a). Many participants in the Shklovski et al. (2010) study praised these resources and found them invaluable. Additionally, as participants were trying to return to their normal lives, the internet became increasingly important to find financial support and employment and share that information with others (Shklovski et al., 2010).

2.2.4.4 April 16, 2007 Virginia Tech Shootings

On April 16, 2007, a Virginia Tech student murdered two students then proceeded through the campus, shooting dozens of fellow students and professors and ended the crisis by killing himself (New York Times, 2007). Before noon that day, 33 people were dead and the community was both grieving their loss and frustrated with the University’s lack of communication and slowness to warn students during the crisis; information that could have potentially saved lives (ABC News, 2007; Virginia Tech Review Panel, 2007). Following the shootings, Virginia Tech also released inaccurate information (e.g., the gunman was in custody)
and took another 39 hours to release the names of the deceased. As one of the first crises to occur in the age of Facebook, Palen et al. (2007b) and Vieweg et al. (2008) studied the public’s use of the Facebook and web 2.0 technologies to deal with the information dearth, generate and disseminate information, and conduct collective problem solving. Within a half hour of the last shooting, students began to post on Facebook asking if their friends were okay. Within 90 minutes, the first Wikipedia page on the tragedy was published and the Facebook group “A Tribute to those who passed at the Virginia Tech Shooting” was created. Shortly thereafter the “I’m Ok at VT” Facebook group started, encouraging students to check in and let others know they were safe. All three would become central sources of information and problem solving for the next 24 hours as students worked together to determine the names of the victims. Facebook activity, instant messaging statuses, personal social contacts, news outlets, and the Virginia Tech student newspaper all served as sources. Students shared what they found while other members would ask for verification and attempt to cross-reference with other sources. As a result, the communities were self-correcting and established reporting norms (e.g., students had to explain their relation to the deceased or information source). The content and activity analysis conducted by Palen et al. (2007b) and Vieweg et al. (2008) found that the online community was able to accurately compile the names of all 32 victims before Virginia Tech officially released their list; an illustrative example of the power of social media collaboration by the public in emergencies. However, bloggers and mainstream media alike purportedly produced misinformation, such as disseminating the wrong identity of the killer (Kantor, 2007).

Hughes et al. (2008) reported on the use of social media as an outlet to grieve and memorialize the tragedy. Numerous Facebook support groups were created (e.g., “Praying for Virginia Tech”) to show support and express sympathy. Some users changed their profile picture to a Virginia Tech memorial ribbon as a symbolic sign of support. On Second Life, a graveyard with a tombstone for each victim, along with his or her photograph and biography was created. Virtual memorial sites also popped up allowing users to place objects similar to the ones at physical memorial sites on the campus. A Flickr photo group was started not just for the purpose of disseminating news but also as an “experience aggregator”; to put a personal context on the tragedy and allow people to express their condolences, thoughts, and prayers (Liu et al., 2008).
Following this crisis and concerns over privacy, news organizations began to contact Flickr group administrators for permission to use the posted photos in their reporting (Liu et al., 2008)

2.2.4.5 October 2007 California Wildfires

In the fall of 2007, over 20 wildfires raged in California from Santa Barbara to San Diego, burning 500,000 hectares and forced large-scale evacuations. According to a survey of those affected, locals were not satisfied with quality and quantity of information from traditional media providers or authorities (Sutton et al., 2008). Sutton et al. (2008) found that similar to criticisms of mainstream news coverage of Hurricane Katrina, locals believed that national news was not accurate or specific enough and was biased towards metropolitan areas. Citizen reviews of the local news were better, but they complained that these providers were unable to keep up-to-date with rapid changes and were not accessible via TV or radio after evacuating the local area. Worst yet, the county emergency website was not able to handle the traffic and frequently crashed. Instead, several community websites emerged or changed their focus to aid residents. “Rimoftheworld.net”, a longstanding community website for residents in San Bernardino, allowed residents to submit news stories, discuss evacuation routes and fire prevention strategies on discussion boards, and maps of the surrounding area. The administrators of the website worked with local firefighters and emergency services to circulate both official and eyewitness information. KPBS, an NPR affiliate in San Diego, provided radio and TV updates and was one of the few commercial sources to offer Twitter feeds and annotated Google maps marked with information such as with fire areas and emergency meeting points. Residents cited these sources as very helpful and more timely, frequent, and useful than traditional sources.

Similar to the study conducted by Torrey et al. (2007) of large versus small online community giving activities during Hurricane Katrina, Shklovski et al. (2008) studied the activity of an emergent community-based volunteer site (“RuralSite”) and a longstanding local information site (“MountainSite”) during the emergency. The authors found that the activity of the longstanding site was more sustained than the emergent site, whose readership and activity dropped off after the initial crisis ended.
The public also found new and innovative ways to use Flickr during the wildfires. Seven main groups were dedicated to posting photos of the event, and many of them utilized geotagging to produce spatial-based photos (Liu et al., 2008). Several news providers such as Yahoo! News, KPBS, and News 8 also created Flickr groups and invited citizen photojournalists to upload their own photos to be used on their official websites. Liu et al. (2008) saw this action as the legitimizing of Flickr and its members as being an important and authoritative part of the information dissemination cycle. Locals also started using the photo-sharing site to upload photos of their property and possessions and tagging them with the keyword “fire” for insurance purposes before evacuating (Liu et al., 2008).

This emergency was one of the first to occur after the popularization of Twitter. In a survey of affected residents, 10% reported using Twitter for information, most of them using the service for the first time (Sutton et al., 2008). In particular, two San Diego residents dedicated themselves to gathering information from friends, news sources, and their personal eyewitness reports and posted all of their findings on Twitter (Poulsen, 2007). These men provided very unique and specific details by venturing around the city, taking photos of their friends’ houses and listing inventory of local supermarkets and stores and told others where they could buy supplies (Poulsen, 2007).

The importance of Twitter hashtags (#) came into focus during this event. The inventor of the hashtag notation, Chris Messina, began to urge people to start using #sandiegofire for related posts to aid those looking for information (Gannes, 2010). Although users began adopt the convention, there was no one agreed-upon term and a variety of keywords emerged (Needleman, 2007).

2.2.4.6 May 12, 2008 China Sichuan Earthquake

When the 2008 Sichuan earthquake occurred, it was Twitter that purportedly first brought the story to the world. AFP (2008) reported that a blogger by the name of Robert Scoble posted the event on Twitter before either the mainstream media or the US Geological Survey could issue news releases. Scoble, a prominent Twitter user, announced on his website that his friends in
China had reported the quake to him via Google Talk; he then published the news of the quake on Twitter (Scoble, 2008).

Due to a combination of extreme telecommunications infrastructure damage and overwhelming call volume, both landline and cell phone services in the area failed. Instead, many turned to the internet for help and information, in particular “Tianya Club”. Tianya Club is one of the most popular internet forums (Bulletin Board System) in China with over 20 million registered accounts. Within 1 minute of the quake (14:28 China Standard Time), earthquake-related posts started to appear on Tianya (Qu et al., 2009). Within 10 minutes, 56 threads reported feeling tremors in 22 different cities, demonstrating that in times of crisis people will turn to the Internet almost immediately for information. Qu et al. (2009) investigated the role that the high-traffic website played in the aftermath of the earthquake and conducted a content analysis to describe the 4 major categories of its use: information-related (seeking, sharing, gathering, and integrating), opinion-expressing (providing feedback on the relief efforts), aiding action (giving, planning and coordinating relief), and emotional (allowing the public to express support, sympathy, etc.). Qu et al. (2009) found that informational posts were the most viewed and replied to and comprised of the majority of threads. Authentic posts and rumours alike were questioned and cross-referenced with multiple sources by members. Moderators did their part to demarcate unconfirmed or disproved information to slow the spread of misinformation. Opinion-related threads were the second most popular (32%) and generated lengthy debates. While not as commonly viewed, action-related threads were used by relief organizations to coordinate between online members, on-site volunteers, and government officials. Often times, relief foundations would post their request for supplies (e.g., gloves, masks, medicines) and members would offer what they had and together they would coordinate to arrange transport. Daily reports of donations and on-site progress further stimulated a sense of community and group accomplishment. Within 10 days over $6.3 million was given online by 660,000 donors.

There are two widely publicized success stories of Tianya members providing information that proved useful to the authorities (Winerman, 2009; Qu et al., 2009). In the first case, the military was attempting send relief to a remote and difficult area but were unable to find suitable terrain for a landing strip and had to delay their efforts. Upon hearing this, a young woman who had
grown up in the area but was currently away for school posted on Tianya the location of a suitable helicopter landing spot. The post was forwarded thousands of times to all of the major online communities until it eventually reached the military. Upon contacting the student the military was able to land where she had described and deploy troops and equipment to those in need. In the second case, the forum provided valuable feedback to government officials. A message that raised much concern from members provided details about the possible embezzlement of relief supplies by officials. This post attracted the attention of the government, who quickly investigated the situation and punished the offending individuals.

Qu et al. (2009) also discussed the use of “Google XunQin” (Finding Relatives), an application created to help family and friend connect. Google XunQin had 3 main features, it: 1) indexed the names of victims in hospitals and made them searchable, 2) indexed “where are you” or “are you okay” messages posted to the 5 major online communities in China for search capabilities, and 3) allowed the batch posting of messages submitted to be posted to all 5 websites. While the total number (or percentage) of success stories was not reported, Qu et al. (2009) stated that the service was effective and allowed thousands to find their relatives.

2.2.4.7 2008 Hurricane Gustav & Ike

Hurricane Gustav and Ike occurred within one week of each other in southern USA (August 25 and September 1, respectively). While in actuality neither hurricane was on the same scale of destruction as Hurricane Katrina, residents and governments alike were concerned and the events were highly publicized. Hughes & Palen (2009) took this opportunity to study hurricane-related Twitter tweets and found that activity spiked when the hurricanes represented the most danger (i.e., when the hurricanes made landfall). During this same time period the number of tweeted URLs and authors were tracked for hurricanes, general discussion, and US political conventions. Hughes & Palen (2009) reported that the minority of users sent a large number of tweets, and this percentage was constant across all events. The authors suggested that this finding supports the idea that a few select users act as information hubs to disseminate information while the majority are bystanders. In addition, the amount of tweeted URLs was higher among emergency events than others. For example, tweets regarding Hurricane Gustav contained URLs 52.4% of the
time, while 36-43% of convention tweets and 24.6% of general tweets had URLs. Although the authors believe this result indicates that crises have higher information demands, it is also possible that events which are significant, of broad interest, and highly publicised are the focus of more public attention and the resultant information supply by the media is large and highly disseminated among users.

2.2.4.8 2009 Red River Flooding

Several researchers took the opportunity to study Red River-related Twitter activity during the 2009 flood season. A detailed analysis of over 7,000 Red River tweets for content and activity by Starbird et al. (2010) found that individuals made up the largest proportion of users (37%), but in terms of tweet volume, dedicated flood information accounts produced the most tweets (44%). Twitter activity was also affected by the public’s risk perception; tweet activity spikes occurred when the threat was growing, and peaked when the risk was highest.

The authors reported that original, first hand or autobiographical information (generative information) was the least common type of content (less than 10%) and was produced almost entirely by locals or those personally connected to the event. Synthesized information was tweeted most often by media outlets and was the second most popularly tweeted content. Derivative information accounted for the most tweets (over 75%) through a user-driven circle of information shaping and sharing using retweets (RT @), follow@ notation, re-sourcing (conducted by webcrawling bots that find and post news from other sources), and URLs. Retweets accounted for 5.7% of the sample, with the majority posted by the media or flood specific services. The follow@ convention was used exclusively by locals or personally invested individuals. Based on this result, the authors suggest that the follow@ notation is essentially a form of recommendation that only those affected by the event would have the credentials to use. Re-sourcing accounted for 69% of tweets and was heavily used by the media, automated feeds, and news aggregators, but not locals, likely because the media do not have first-hand knowledge or experience and must rely on other sources to produce information. Lastly, URLs were found in 56% of sampled tweets and were used much less frequently by locals who had first-hand information to rely on.
A secondary analysis of the Red River dataset by Starbird and Palen (2010) focused on retweet activity. Their analysis found that retweets in their sample were more likely than non-retweets to be related to the emergency. Nearly all of the information that was retweeted was generated by the mainstream or local media, service organizations, and dedicated flood-specific news groups or Twitter accounts and only three of the most highly retweeted users were individuals. The authors found that two major categories of information were retweeted: 1) general information with broad appeal and 2) information that had local utility. The first category was usually shared by those not directly affected by the flooding (e.g., photos of the flooding) and were not retweeted among locals. On the opposite side of the spectrum, information that was useful to locals (e.g., sandbagging coordination efforts, road closures, river levels) was highly circulated by locals. Based on these findings, Starbird and Palen (2010) concluded that users selectively choose to spread certain types of information over others using retweets, depending on what they deem is important or newsworthy.

One innovative use of Twitter that the authors reported on was the creation of a flood-services account by a local individual (Starbird et al. 2010). This tech-savvy citizen programmed a script that would take information published by the US Geological Survey and publish it automatically on Twitter whenever it was updated (usually on the hour). Based on this example, Starbird et al. (2010) emphasized that while locally generated information is important, official information is still authoritative and necessary and can be made more useful and relevant with local input and technological modifications.

2.2.4.9 January 13, 2010 Haiti Earthquake

Similar to other examples of citizen photojournalism in disasters, moments after a catastrophic magnitude 7.0 earthquake struck Haiti, affected citizens were using their mobile phones to take photos of their plight and send them via Twitter (Parr, 2010). For some Haitians who lost their phone landlines, Facebook was the only way to communicate their status to loved ones and learn about the fate of others (Keen, 2010). Meanwhile, users around the world spread news of the quake from traditional news sources using Twitter and other forms of social media (Nielsen Company, 2010). Over 2.3 million tweets were sent between January 12 to 14 (Sysomos, 2010).
and over 1,500 Facebook status posts per minute contained the word “Haiti” (Keen, 2010). Grassroots relief websites by both Haitians (Le Roux, 2010) and tech-savvy volunteers (Brodigan, 2010) quickly sprung up to offer aid.

Mobile giving had been piloted in other disasters, but the 2010 earthquake was the most successful to-date. In response to the tragedy, Twitter updated their official blog to announce a Red Cross texting campaign where users could text “HAITI” to donate $10 to the relief effort (Twitter, 2010). Within 48 hours, over $3 million dollars had been raised, thanks in large part to viral dissemination via Twitter (Siegel, 2010). The Red Cross enlisted the help of popular celebrities and Twitters, including first lady Michelle Obama, to re-tweet important messages on behalf of the organization (Morgan, 2010). Haitian musician Wyclef Jean’s Haiti relief organization, Yele Haiti, also used a similar text message and Twitter approach and raised $1 million dollars within 24 hours (Siegel, 2010). Similarly, Oxfam America raised $7,500 in 12 hours using Facebook (Keen, 2010).

While mobile and social media giving was being heralded by some as an innovative and rapid way to reach an untapped audience of donors, others were concerned that smaller text donations would cannibalize potential larger donations as givers would feel as though they had already helped (Wallace, 2010). Speed is also a factor in mobile giving. Although awareness and donor response in mobile giving campaigns can be rapid, these funds can take up to 90 days to reach charities as cell phone service providers must work with 3rd party companies to tabulate and transfer money (Wortham, 2010). Additionally, the potential for misinformation and scams to propagate on social media is great. For example, a Facebook group was established to pledge $1 for every person that joined. While over 500,000 people joined, it was not clear where the money would come from and to what organization it would be given to as no further details were given (Morgan, 2010). In addition, rumours such as:

"AMERICAN AIRLINES is taking doctors and nurses to Haiti for free. Please call 212 697-9767. Spread the word. Thanks!"

and
"UPS IS SHIPPING TO HAITI FOR FREE Today!! IT HAS TO BE UNDER 50 lbs!! If you have ANYTHING to give, round up your stuff ...

were circulated so heavily on Twitter that companies such as UPS, American Airlines, and JetBlue were forced to issue official statements (Akre, 2010).

### 2.2.5 Summary

The Social Amplification of Risk Framework indicates that risk messages from authorities can be amplified or attenuated based on the social processes involved in message transmission. Additionally, feedback between individuals and groups is important to the tailoring of future messages and official responses. The media and traditional mediums such as newspaper, radio and television have been the most common way for risks to be attenuated or amplified and for feedback to be gathered. But with the growing pervasiveness and adoption of the internet and social media, the public are playing a larger role in creating content, transmitting information, and amplifying and/or commenting on traditional news stories, as seen by the case studies of crisis communication by the public during emergencies. As such, a greater emphasis must be placed on social media users in the Social Amplification of Risk Framework and they should be considered as important sources, transmitters, and receivers of messages. However, this introduces new challenges such as quality control.

In a Web 2.0 environment, it is more difficult for institutions to send out and manage one consistent and dominant message. Instead, they must compete with other sources as they are no longer the sole gatekeepers to information. In response, public health authorities have begun adopting these technologies to better communicate and interact with the public. But even prior to adoption by officials, the public has used these communication technologies over the years in grassroots efforts to seek and mobilize aid and rapidly disseminate information in emergencies. Platforms such as Twitter are becoming more popular for these purposes and the public are generating a wealth of crisis-related posts as a result. Twitter may be suited for crisis communications as users can send messages using mobile devices and broadcast them to a wide audience wherever cell phone or wireless internet service is available.
2.3 Traditional Methodologies & Infodemiology

2.3.1 Traditional Methods of Public Polling & Content Analysis

Traditionally, telephone surveys and face-to-face interviews are the most common methods for public health officials to understand and measure public anxiety, knowledge, perceptions, attitudes, and behaviour in regards to health issues. Without survey research, authorities would not know how the public is responding during a crisis. However, this data must be collected rapidly and within a short time frame to elicit the most relevant information (Blendon et al., 2003). Public perceptions, opinions, and behaviour can change quickly in response to unforeseen events and are particularly sensitive to media coverage as the situation develops (Rubin et al., 2008). Long data collection periods which produce only aggregated data can also miss changes in public behaviour and perceptions influenced by specific events (Blendon et al., 2003). Lastly, the closer that data can be collected to the occurrence of the crisis, the less risk there is of recall bias (Wessely et al., 2003). Consequently, rapid-turnaround surveys are the method of choice for most researchers. However, short-duration surveys can come with many barriers as it can take precious time for agencies to gather resources, funding, and survey instruments in order to conduct polling (Blendon et al., 2003).

Carefully designed surveys or questionnaires for any medium (e.g., mail, telephone, in-person, internet) require time to write, design, and test (Fowler, 2008). For instance, it is necessary to ensure that questions have clear instructions, precise language, and carry consistent meanings (Fowler, 2008). Validation of the survey instrument or use of a pre-existing, validated instrument should also be considered (Fowler, 2008). Answers must be properly mapped if a numeric scale (e.g., Likert scale) is used and be exhaustive if multiple choices are given (Fowler, 2008). Questionnaires can also elicit reactive effects such as social desirability; this can lead to systematic under-reporting of attitudes or behaviours that do not fit with social norms and the over-reporting of those that do (Tourangeau & Yan, 2007). Ideally, cognitive interviewing is conducted prior to pilot testing to address or mitigate these design issues, yet performing the interviews and running the subsequent analyses can be time and resource consuming (Willis, 2005). Surveys or interviews conducted face-to-face have the additional problem of sending
trained staff to the geographic area of interest, something that may not be feasible or safe to do in emergencies.

While internet-based surveys can help to achieve large samples over a geographic spread within a short time frame and at a low expense, the representativeness of the sample is not always guaranteed for specific populations (Fowler 2008; Rubin et al., 2008), particularly in “open” web surveys were self-selection bias is high (Eysenbach & Wyatt, 2002). Representative internet panels are becoming a popular method of dealing with this problem, but telephone surveys are seen as the most feasible way to rapidly access a representative population from near or far locations. However, a discussion by Rubin et al. (2008) highlighted some of the methodological challenges of conducting telephone surveys. Generally, women are more likely to answer the phone, those who share a phone line with many others are less likely to be selected, those who have multiple phone lines are more likely to be selected, and those without any land phone line have no chance of selection at all (Ruben et al., 2008). The later point is becoming a larger and larger issue in survey research as the mobility of the population increases and turns to cell phones. Additionally, if phone studies are conducted primarily during working hours, those who do not have jobs (e.g., housewives, elderly) are more likely to be selected (Fowler, 2008). Lastly, in emergencies, telephone surveys may not be feasible if people evacuate their homes (and land lines) and if telecommunication structures are damaged.

Content analysis of online publishing activity is another common method to gain rich qualitative and quantitative insights into public opinions, attitudes, and discourses surrounding public health issues (e.g., Keelan et al. 2007; Keelan et al., 2010). However, these content analyses are often time consuming, as posts must be read line-for-line by researchers and systematically coded. The long process of collecting and analysing this data can be detrimental to the timeliness and relevance of the findings, particularly when data is needed to quickly address issues.

2.3.2 Infodemiology & Infoveillance

In recent years, various methods have been developed to mine, aggregate, and analyze in real-time the vast amount of text data being produced by Internet users with the goal of potentially
informing public health and policy (Eysenbach, 2009). This branch of science, termed infodemiology ("information epidemiology"), can be defined as the “distribution and determinants of information in an electronic medium, specifically the internet, or in a population” (Eysenbach, 2002, 2006, 2009). Data sources may include information generated by the public on search engines, news websites, social media websites, discussion groups (forums), blogs, and microblogs (Eysenbach, 2009). The term infoveillance can be used specifically to describe studies that focus on the analysis of internet information demand (e.g., search queries) and supply (e.g., blog or status publishing) for the purposes of surveillance (Eysenbach, 2009). In practice, understanding online information supply is important to public health professionals, not only to deal with harmful or prolific misinformation on the internet. Knowledge of information demands allows authorities to potentially detect disease outbreaks, address “epidemics of fear” (Eysenbach, 2003), or address true concerns and questions. Infoveillance intelligence has the potential to inform ongoing or emergency public health education and campaigns.

2.3.2.1 Infodemiology Applications in Research

Researchers and public health organizations alike have experimented with different infodemiology applications. The first infodemiology study was a report by Eysenbach (2006) on the correlation of influenza-related Google searches and influenza cases in Canada. As Google search logs were not made publically available, a Google ad campaign for flu keywords was created to assess the public’s search activity. Following a search for flu-related keywords the sponsored flu ad would appear and could be clicked on by the viewer. Google statistics measured both the amount of views and clicks. Eysenbach (2006) found that internet clicks on the flu ad correlated better with flu cases in Canada for the 2004/2005 flu season than influenza-like illness sentinel physician data (ILI-SPR). Interestingly, the ad clicks were able to predict flu events for the following week while the ILI-SPR data best correlated with the current week. This novel pioneering study only cost $365.64 for the entire flu season; much less costly than traditional surveillance or polling methods.
Following the study conducted by Eysenbach (2006), several other researchers have replicated these findings using other search engines or techniques. Polgreen et al. (2008) took Yahoo! search query logs from 2004 to 2008 and correlated them with US influenza data at the US census-region level. Similarly, Hulth et al. (2009) found that search queries submitted to a Swedish medical website correlated with Swedish influenza data. Recently, Ginsberg et al. (2009) of Google replicated the seminal Eysenbach study using five years of their search log data. Google flu search data is now available free for use on the Google Flu Trends website and has since been used to show correlations to other diseases such as 2001 West Nile Virus (Carneiro & Mylonakis, 2009) and 2009 H1N1 (Wilson et al., 2009). Search query correlations have also been demonstrated for the 2008 listeriosis outbreak (Wilson & Brownstein, 2009) and the 2009 FDA salmonella peanut-butter recall (Brownstein et al., 2009). However search activity for other epidemics such as SARS (Eysenbach, 2003) and avian flu (Carneiro & Mylonakis, 2009) were not sensitive enough to detect true outbreaks.

Using the same search query methodology, Cooper et al. (2005) of the CDC used Yahoo! search activity of the 23 most common cancers in the United States to correlate with estimated incidence, mortality, and volume of media coverage of each cancer. The authors found that cancer-specific search queries correlated with all three variables but most strongly with media coverage. The authors postulated that this finding emphasized the important and influential role that the media played in the public’s attention on cancer.

In addition to syndromic surveillance and disease monitoring, infoveillance provides novel avenues for research on public attention and opinion. Ripberger (2009) proposed that the same information demand-based methods that Eysenbach (2006) applied to influenza could potentially be applied to measuring public attention. Ripberger (2009) found that preliminary work using Google Insight to compare news articles (supply) and search trends (demand) had both face and convergent validity. Face validity in this study was defined as whether the relationship between supply and demand activities was expected or made logical sense. For instance, one would expect the volume of “Afghanistan War” searches to be low until media coverage intensified, and Ripberger (2009) found that this was the case. Ripberger (2009) measured convergent validity by using the correlation between the volume of search terms for “health care”, “global
warming”, and “terrorism” with the volume of media articles produced using those keywords and found significant correlations in all three cases. A time-series analysis found that public attention to some variables, such as health, preceded media coverage, while other keyword searches for other topics (like terrorism) lagged behind or occurred in parallel. The findings demonstrate the potential for infoveillance applications to public attention research, although Ripberger (2009) emphasizes that potential issues such as the representativeness of Google searches and keyword selection methodology would need to be addressed.

2.3.2.2 Infodemiology Applications in Public Health

The ever present threat of infectious diseases, epidemics, and bioterrorism in a globalized world has lead health authorities to be vigilant in their monitoring and tracking of local crises in order to prevent and control them before they become global health issues (Hartley et al., 2010). International Health Regulations (IHR) dictate the international legal framework for the early detection, reporting and response to infectious diseases (Baker & Forsyth, 2007; Wilson et al., 2008). Under IHR, WHO nations must report diseases of global concern to the WHO within 24 hours of first knowledge and are authorized to use non-governmental sources of information for surveillance purposes. Several infoveillance applications play important roles in these global surveillance processes.

ProMED-mail

Unlike infoveillance applications, ProMED-mail is a website and e-mail list that is run by a community of users that manually search and review health-related reports. This organization acts as a warning system and communication network for public health experts, health professionals, and the public to share emerging outbreak information and collaborate in emergencies. As a free program supported by the International Society of Infectious Diseases it remains unaffiliated with any official authorities or governments. E-mail updates from subscribers and manual searches of informal and official reports by staff serve as the main data sources. Top moderators or editors send reports to subject area experts to verify any scientific information as well as provide any necessary context or background. Depending on the urgency of the event, finalized reports are sent out within 24 hours or less via e-mail lists and the
ProMED-mail website. Despite being entirely manual, ProMED-mail has frequently been the first to report information on emerging outbreaks, such as the 2003 SARS epidemic, and by doing so, aided health professionals on the front lines of pandemics (Woodall, 1997; Madoff, 2004). However, one study suggested that while the value of ProMED-mail is real, it is also limited (Zeldenrust et al., 2008). Zeldenrust et al. (2008) conducted a retrospective study of 13 months of Netherlands Early Warning Committee (NEWC) reports to determine the added value of ProMED-mail reports. Of the 25 potential threats, five were potentially threatening to the Netherlands. Two (40%) of these alerts were found to be reported only by ProMED-mail. Although the sample size was small, the authors concluded that ProMED-mail did have some limited value to authorities, but mostly praised the speed of reporting; the lag time between the first ProMED-mail report and other reporting agencies ranged from 0-42 days.

**Global Public Health Intelligence Network (GPHIN)**

In 1994, when the pneumonic plague broke out in Surat, India, it was CNN, not the WHO or the government of India that first reported on the epidemic (Mykhalovskiy & Weir, 2006). Instead, the WHO’s information system failed under old technology (then fax machines and phones), poor coordination, and a lack of foreign field agents to provide up-to-date situational information (Galaz, 2009). The failure of the WHO to quickly verify and disseminate critical information became the impetus for collaborating with Health Canada to develop an automatic online news monitoring system, the Global Public Health Intelligence Network (GPHIN).

GPHIN gathers information on public health events by monitoring global media news 24 hours a day, 7 days a week (Mykhalovskiy & Weir, 2006). The system uses Factiva and al Bawaba (Arab language news) news aggregators as its primary data source. Factiva alone covers over 9,000 news sources in 22 languages. GPHIN scans for potential infectious disease events, as well as biological, chemical, environmental, radioactive, and natural disasters. A customized taxonomy identifies articles, filters out duplicates, and creates relevancy scores. Events with low relevancy scores are discarded and those with high scores are evaluated by human analysts who then disseminate reports to subscribers. The verification of reports is carried out by the WHO as they have the international authority to make inquiries of both governments and health authorities (Grein et al., 2000). GPHIN processes 2,000 to 3,000 news items per day and
supplies the WHO with 40% of its early warning information (Mykhalovskiy & Weir, 2006). Notably, GPHIN is credited with detecting the first alert of SARS in Guangdong, China in November 2002 (Eysenbach, 2003). Subsequent alerts provided to the WHO in February 2003 prompted a WHO investigation and confirmation of the epidemic from the Chinese government. Thus far no peer-reviewed evaluation studies of GPHIN have been published.

**HealthMap**

Launched in 2006, HealthMap is a real-time web-based mashup (web application hybrid) that scans the internet for public health news and then aggregates, filters, and integrates these reports using both taxonomies and human judgement (Freifeld et al., 2006; Brownstein et al., 2008). Information is collected mainly through Google News, ProMED-mail, and authorities such as the WHO, although blogs, microblogs, and social networking sites are being incorporated (Brownstein et al., 2010). Users can also submit first-hand reports via the website, e-mail, texting, iPhone applications, or Twitter. Aggregated health events are displayed on an interactive map and can be viewed by location, date, or disease. Thus far the automated classifier has yielded 84% accuracy (Freifeld et al., 2006). HealthMap acts as a web 2.0 combination GPHIN and ProMED-mail; it possesses the automated scanning ability of GPHIN mixed with the public participation of ProMED-mail. Most recently, HealthMap partnered with the New England Journal of Medicine to provide up-to-date situational information during the 2009 H1N1 outbreak (Brownstein et al., 2010). From April 1 to December 31, 2009, over 87,000 H1N1 reports were collected and categorized with good sensitivity according to suspected or confirmed cases or deaths, as well as time and location (Brownstein et al., 2010).

**Infovigil**

Infovigil (Eysenbach, 2009) is an infoveillance system developed at the Centre for Global eHealth Innovation in Toronto. Infovigil continuously gathers and mines textual information on the internet (e.g., webpages, blogs, Twitter) and is capable of generating descriptive statistics on a number of metrics (e.g., keyword frequency, number of relevant posts per day, keyword ratios). The data can be visualized and graphed in a number of ways and outputs can be configured using specific parameters such as date range. Visualizations can be annotated to aid with data interpretation.
2.3.2.3 Advantages & Disadvantages of Infodemiology Approaches

Official health reporting, while authoritative, does have its shortcomings. These reports may deny facts that may have potential negative impact on trade and tourism or become delayed due to a lack of human resources or because of the many levels of government that require clearance before a public news release (Woodall, 1997). Additionally, official reports typically err on the side of conservatism in terms of the severity and extent of outbreaks and reveal crucial information only after many series of tests and confirmations (Woodall, 1997). Although some of these prudent processes are used to convey the most accurate information, they can often times lead to public unease, confusion and costly delays. The 2003 SARS outbreak in China is one example of the pitfalls of non-transparent government reporting. Instead, infoveillance allows for thorough searches of available material to be conducted around the clock in a mostly automated fashion, providing rapid information for analysis, dissemination, and action. Data captured this relatively inexpensive way is also outside of normal communications and channels and can supplement information-poor situations (Brownstein et al., 2009).

In developing countries, surveillance is extremely important but often times they have minimal or little traditional disease reporting due to a lack of government infrastructure or funding (Woodall, 1997). Webcrawling systems such as GPHIN may not detect diseases in these countries if the supply of local news reporting is low and published in uncommon or unfamiliar languages (Grein et al., 2000). In these situations, the lack of information technology and supply limits potential for infoveillance. However, the adoption of hand-held devices and mobile phones with internet access or SMS by citizens in these nations is rapidly growing. Additionally, foreigners such as missionaries, relief organizations, contractors, and mining outfits often times come equipped with their own telecommunications (Woodall, 1997). In these resource-poor settings the public may be able to act as citizen reporters and provide valuable on-the-ground situational information and reports on public behaviour to fill existing gaps (Woodall, 1997; Chretien et al., 2008).

However, there are many unknowns about infoveillance and its use for public health (Brownstein et al., 2009; Wilson & Brownstein, 2009; Eysenbach 2009). Firstly, the information used is
unstructured, free-form text. This requires careful and complex query and algorithm development to yield meaningful results with good sensitivity. Similarly, infoveillance keywords and signals can lack specificity and give rise to false positives; in these cases more human verification, time, and resources is required. Additionally, because reports are coming from all different types of sources across the globe, verification and follow-up may be extremely difficult. For example, the WHO found that over 2 years, 3% of events reported by infoveillance methods such as GPHIN could not be followed up on or were inconclusive (Grein et al., 2000).

Information supply (news reports, blog posts, status updates) and demands (search queries) are also sensitive to external forces, such as the newsworthiness of events. Increased activity generated by a “celebrity effect” (Polgreen et al., 2008) may lead to false positives or “epidemics of fear” (Eysenbach, 2003). For instance, Cooper et al. (2005) demonstrated the heavy influence that the media had on cancer-related searches. Although this issue is problematic, it emphasizes the need for careful human analysis of infoveillance reports. Regardless, false positives can still be used by public health officials to signal public information needs that must be met by authorities (Eysenbach, 2003).

However, infoveillance should not be seen as a replacement for traditional methods or human intervention, but rather as a supplement. Traditional surveillance will still be necessary to estimate morbidity, mortality, incidence of disease, demographic factors, and fatality rates (Brownstein et al., 2010). Infoveillance can instead be seen as a means to fill information gaps, provide timely updates, and allow rapid global public health collaboration.

2.3.3 Summary

Up-to-date situational information and data on public responses to emergencies are crucial for authorities as they provide valuable details that can be acted upon. But traditional methods of public polling can be difficult to conduct quickly or limited in times of crisis. New infodemiology approaches that analyze online information supply and demand in real-time may help to supplement information-poor situations. Potential applications of infodemiology have been demonstrated by researchers and today public health authorities and experts use a number
of infoveillance systems to maintain situational awareness. With infodemiology methods, publically generated health-related content can be analyzed to gain potential understanding about public attention, attitudes, behaviours, and information demands. In our study, we aim to use real-time infoveillance methods to study the content of posts written on Twitter during a public health emergency, the 2009 H1N1 pandemic influenza.

### 2.4 The 2009 H1N1 Pandemic Influenza

The following section provides a brief timeline and summary of relevant H1N1 literature in order to provide the context for our case study on Twitter use in public health emergencies.

#### 2.4.1 A Brief Timeline of H1N1

The 2009 swine-origin influenza A was first noted in mid-March in La Gloria, Mexico after 60% of the town’s population fell ill with an unknown respiratory disease. Later that month the first case of H1N1 was confirmed in the United States (MMWR, 2009a). After a series of genetic sequence testing, the Public Health Agency of Canada confirmed on April 23 that the series of illnesses in Mexico was due to the H1N1 virus (MMWR, 2009b). On April 24, the World Health Organization (WHO) issued its first H1N1 Disease Outbreak Notice and formally declared a “public health emergency of international concern” shortly thereafter (WHO, 2009a). Following a significant increase in the number of cases worldwide over the ensuing months, the WHO raised the Pandemic Alert Level to Phase 6 on June 11th to indicate that widespread human infection was imminent (WHO, 2009b). The fall of 2009 saw a second wave of H1N1 and the start of mass vaccinations in several major countries, with the US and Canada starting their vaccination strategies in October 2009 (CNN, 2009; CBC, 2009b). Overall, the pandemic was milder than originally predicted and claimed at least 12,200 lives worldwide (Reuters, 2010).

#### 2.4.2 Public Perceptions & Attitudes towards H1N1

Numerous research groups aimed to study the attitudes, behaviour, and perceptions of the public in response to the epidemic in order to inform public policy and add to the growing body of
pandemic literature. At the time of this writing, at least 20 peer-reviewed studies had been published (see Appendix 2-1 and 2-2 for search methodology and study selection). Studies covered a range of geographic areas, with the majority focused on respondents in Europe and Australia. The length of data collection ranged from 1 day (Woien & Tonsberg, 2009) to four months (Van et al., 2010). The majority reported their results in aggregate, although a few reported selected results longitudinally (e.g., Wong & Sam, 2010b; Van et al., 2010; Sypsa et al., 2009). The majority of studies (8) utilized random digit telephone sampling to meet proportional quotas, and three used representative online research panels (Quinn et al., 2009; Schwarzinger et al., 2010; Maurer et al., 2009). Two studies sent surveys to the whole school population (Van et al., 2010; Effler et al., 2010). Three utilized convenience sampling at malls (Balkhy et al., 2010; Seale et al., 2009; Seale et al., 2010), three distributed internet-based surveys using convenience sampling or open web surveys (Goodwin et al., 2009; Goodwin et al., 2010; Jones & Salathe, 2009), and in one study the sampling method was stated to be random, but it was unclear if this was actually the case (Kamate et al., 2010). Appendix 2-3 also contains a summary of selected studies.

2.4.2.1 Risk Perception of H1N1

Internationally, the perceived risk of H1N1 was low and the majority of the public did not believe they were likely to contract the disease (Jones & Salathe, 2009; Schwarzinger et al., 2010; Wong & Sam, 2010a; Quinn et al., 2009; Seale et al., 2009; Sypsa et al., 2009; Rubin et al., 2009; Lau et al., 2009a, Lau et al., 2009b; Seale et al., 2010; Van et al., 2010). The rate of respondents stating they had high or very high personal risk of becoming infected ranged from 8% (Schwarzinger et al., 2010) to 42% (Van et al., 2010), with the majority of studies reporting 20% or less. These ratings of personal susceptibility also decreased over time (Van et al., 2010). Perceptions of the risk to family or the public were also low (Seale et al., 2010; Quinn et al., 2009; Lau et al., 2009a). Congruent with ratings of low personal risk, feelings of personal control over infection were high and more than half of respondents believed they had the ability to avoid the disease (Rubin et al., 2009; Jones & Salathe, 2009; Goodwin et al., 2009, Seale et al., 2010).
Respondents from different locations failed to recognize the severity of H1N1 infection in respect to their health (Eastwood et al., 2010; Lau et al., 2009a; Rubin et al., 2009; Sypsa et al., 2009; Seale et al., 2010; Schwarzinger et al., 2010). 43% (Seale et al., 2009) to 22% (Eastwood et al., 2010) of respondents believed that H1N1 would have very or highly severe health effects over the course of the pandemic.

Compared to other threats such as diabetes, HIV, injuries, terrorism, heart disease, cancer, avian flu, or seasonal flu, one study found that, early on in the pandemic, H1N1 was perceived as the most threatening second only to physical injury (Jones & Salathe, 2009). However, when compared to SARS in terms of fatality, consequences to the community, duration, and number of cases, H1N1 was perceived as less serious, although one in five respondents believed that H1N1 would have a higher fatality rate than avian flu (Lau et al., 2009a). The majority correctly stated that the fatality rate of H1N1 was lower than seasonal flu (Lau et al., 2009a), although the absolute death rate of seasonal flu was still underestimated (Goodwin et al., 2009).

2.4.2.2 Knowledge & Misconceptions of H1N1

Several studies reported on the misconceptions and level of knowledge participants had of H1N1, starting with its nature as a virus. The majority of participants were aware of the viral nature of H1N1 (Balkhy et al., 2010), however, 28% also believed that it was an immunodeficiency disease (Balkhy et al., 2010). Only 18% knew that swine flu was caused by the H1N1 virus (Kamate et al., 2010), and 43% erroneously believed that H1N1 was a new avian flu instead of a swine influenza subtype (Lau et al., 2009a). Kamate et al. (2010) also reported that awareness of H1N1 and swine flu terms was not equal among the public (40% versus 80%).

There was also considerable confusion regarding routes of transmission. While nearly all participants recalled that H1N1 was transmitted via droplets (Lau et al., 2009a; Lau et al., 2009b; Balkhy et al., 2010), the other routes of transmission (touching an infected person or contaminated object) were less known with only 60-80% of respondents answering correctly (Lau et al., 2009a; Lau et al., 2009b; Balkhy et al., 2010). Overall, 40% were not aware of at least one of the three routes (Lau et al., 2009a). Several erroneous transmission routes were also
commonly believed: 40% via water sources (Lau et al., 2009a), 38% via sexual route (Balkhy et al., 2010), 39% airborne (Lau et al., 2009a), and 25% via insects (Lau et al., 2009a). Lau et al., (2009a) reported that nearly 70% of respondents had at least 1 of these misconceptions. The majority were not aware of the incubation or communicability period (Balkhy et al., 2010).

Although the pork industry was affected by H1N1 and “swine flu” terminology, only 7% of Europeans were found to stop eating or reduce their intake of pork (Goodwin et al., 2009) and 7% of Hong Kong respondents believed that eating pork could cause H1N1 (Lau et al., 2009a). Slightly more than half knew that H1N1 originated in pigs in one survey (Kamate et al., 2010).

Misconceptions about vaccines were also documented. Several studies reported that a significant proportion (25-40%) of their respondents believed that vaccines for seasonal influenza were also effective against H1N1 (Lau et al., 2009a, Goodwin et al., 2009). Eastwood et al. (2010) also reported that 4% of participants who were vaccinated for seasonal influenza did so because of this misconception. In addition, 40% of Australians sampled believed that vaccines caused people to become infected with the virus (Seale et al., 2010).

### 2.4.2.3 Behavioural Responses to H1N1

The majority of studies investigated the public’s response or outward behaviour towards H1N1. These responses can be categorized into 3 major categories: intent to change behaviour, attitudes towards behaviour changes, and actual behaviour change.

**Intenstions to Change Behaviour**

Intended precautionary behaviours seemed to differ between Europe and Asia. Lau et al. (2009a) found that in Hong Kong, 1 in 4 respondents were likely to wear a facemask during the pandemic and 42% would do so if symptomatic. Malaysians were similarly concerned with purchasing products such as facemasks in order to protect themselves (Goodwin et al., 2009). In contrast, only 15% of European respondents intended to purchase goods (Goodwin et al., 2009). In both Hong Kong and Malaysia, 40-50% of participants were likely to practice avoidance behaviours (e.g., avoid crowded places or travel) (Lau et al., 2009a; Goodwin et al., 2009). Respondents
from Europe did not share this need for avoidance; only 1 in 5 participants stated that they would avoid public transit or delay travel plans (Goodwin et al., 2009).

In a survey of staff and students at the University of New South Wales, the majority reported that they would not come to the university if they were experiencing any influenza-like symptoms (79%) (Van et al., 2010). However, most concerning was the lack of precautions that students would take if they had an exam or assignment due; 67% reported that if they were symptomatic under these situations they would continue attend class.

**Attitudes toward Precautionary Measures**
Quarantine was seen as a highly effective precautionary measure by 80% of respondents in Australia (Seale et al., 2009), but only 37% of Indians (Kamate et al., 2010). The majority of participants in Saudi Arabia (Balkhy et al., 2010) and Hong Kong (Lau et al., 2009a) supported hypothetical quarantine sanctions by the government. Hong Kong respondents also highly supported the real government quarantine of Metropark Hotel on May 2, 2009 (Lau et al., 2009a). In contrast, a study of Perth school closures in Australia found that only half of parents believed the closure was appropriate (Effler et al., 2010). 33% believed that H1N1 was too mild to warrant the action (Effler et al., 2010).

Agreement with or belief in avoidance behaviours varied between eastern and western cultures. The majority (80-95%) of respondents from Eastern countries (Saudi Arabia and Hong Kong) believed that avoidance of infected peoples, areas, and crowds was an effective way of preventing H1N1 (Balkhy et al., 2010; Lau et al., 2009a). In a UK study, less than half of participants agreed that these methods were necessary (Rubin et al., 2009).

Although simple hand washing was a CDC-recommended precautionary behaviour (CDC 2009e), it was not seen as highly effective by all respondents. Several studies found that only half of participants believed that hand washing was highly effective (Rubin et al., 2009; Seale et al., 2009), while others reported even lower numbers (Goodwin et al., 2009; Kamate et al., 2010). When results were summed to include those who perceived that hand washing had medium or higher effectiveness, this number climbed to over 80% in the UK and Australia.
(Rubin et al., 2009; Seale et al., 2009) and 98% in Hong Kong (Lau et al., 2009a). However, using this method the number of respondents in India grew only to 50%, perhaps suggesting a low uptake of public health messages in the region (Kamate et al., 2010).

The effectiveness of facemasks against H1N1 differed across regions. In Hong Kong, where the use of face masks for flu prevention is common, nearly all (93%) believed that the product was effective against H1N1 and 41% erroneously believed that the government recommended face masks as a precautionary measure during the pandemic (Lau et al., 2009a). In India, 38% saw facemasks as offering a high level of protection (Kamate et al., 2010). However, participants from western countries such as Australia and the UK did not share the same view of the product, with less than a quarter of participants believing that masks were highly effective (Rubin et al., 2009; Seale et al., 2009).

Only two studies, from Australia and India, asked respondents about the effectiveness of vaccines, antivirals, antibiotics, and herbal medicines as prevention methods side-by-side with previously discussed methods (Seale et al., 2009, Kamate et al., 2010). Seale et al. (2009) found that pandemic vaccines were seen by Australians as highly effective by approximately 65% of the population while only 36% of Indians shared that view, equal to the amount of people who judged face masks to be effective (Kamate et al., 2010). Antibiotics, which are not effective against viruses, were rightly perceived to have low effectiveness against H1N1 among Australians (Seale et al., 2009). Interestingly, Kamate et al. (2010) found that the overall perceived effectiveness of antibiotics in India was greater than that of antivirals. The use of herbal medicines was not seen as effective in either study (Seale et al., 2009; Kamate et al., 2010).

**Actual Behaviour Change**

The actual use of hand washing as a precautionary measure ranged from 28% (Rubin et al., 2009) to nearly 80% (Jones & Salathe, 2009). The majority of studies reported that half or more were increasing the frequency of hand washing (Balkhy et al., 2010; Lau et al., 2009a; Kamate et al., 2010; Seale et al., 2010; Jones & Salathe, 2009). Disinfecting surfaces such as door handles and counter tops was also a CDC-recommended behaviour (CDC, 2009e). However, only 17-
36% of respondents reported doing so (Rubin et al., 2009; Kamate et al., 2010; Seale et al., 2010). Only 38% of participants were covering their nose and mouths when sneezing (Balkhy et al., 2010).

The frequency of avoidance behaviours (e.g., avoiding public transit) taken by the public ranged from 60% (Jones & Salathe, 2009) to 0.4% (Rubin et al., 2009). Avoidance behaviour was particularly low (less than 15%) in the UK, Australia, and Hong Kong (Rubin et al., 2009; Seale et al., 2010; and Lau et al., 2009a). Taking time off of work or school was the least common avoidance behaviour, with less than 5% of respondents using this kind of protective measure (Jones & Salathe, 2009; Rubin et al., 2009; Seale et al., 2010). Although Kamate et al. (2010) reported a significantly higher number of Indians taking time off of work (25%) and not taking their children to school (16%), these behaviours were the lowest compared to other avoidance behaviours.

Several studies found that the public’s overall behavioural response to the pandemic to be fairly moderate with roughly only 40% of respondents making a specific health change (Rubin et al., 2009; Van et al., 2010). But perhaps of more concern is the public’s lack of self-isolation during symptomatic or potential outbreaks. In India, Kamate et al. (2010) found that nearly 30% of participants continued their normal daily activities if they had symptoms. When the number of H1N1 cases began to rise in Australian school systems, officials in Perth closed several schools from June 8-14, 2009 to stop the spread of disease. However, Effler et al. (2010) reported that during this time almost 75% of children left their home to engage in sports, play at the park or beach, go shopping, or attend social events. On average, children left the house 3.7 times during the closure.

Despite few behaviour changes in response to H1N1, it is very likely that the public’s response was influenced by news and events, and as such aggregated data or short study durations may not give the most accurate depiction of public behaviour if the context is ignored. For instance, Wong and Sam (2010b) found temporal trends in H1N1 protective behaviour over 6 weeks of data collection. Their findings indicated that during the acute phase of the pandemic in
Malaysia, protective and avoidance behaviours increased and then tapered off after the immediate threat was over (Wong & Sam 2010b).

2.4.2.4 Emotional Responses to H1N1

The level of concern or fear regarding H1N1 varied across studies (Quinn et al., 2009; Schwarzinger et al., 2010; Goodwin et al., 2009; Goodwin et al., 2010; Woien & Tonsberg, 2009; Eastwood et al., 2010; Wong & Sam, 2010a). The range of concern was as low as 15% (Schwarzinger et al., 2010) to as high as 26% of respondents being very worried (Goodwin et al., 2009). Other emotional states such as panic and depression were found to be less than 5% and overall, H1N1 caused less emotional disturbance than SARS (Lau et al., 2009a). Only one study used a validated measure of anxiety, the State Trait Anxiety Inventory (Rubin et al., 2009). In this study, 24% of participants scored high enough to suggest anxiety and 2% of these respondents were highly anxious (Rubin et al., 2009). Two studies looked at anxiety longitudinally (Van et al., 2010; Jones & Salathe, 2009). Van et al. (2010) found that less than 5% of respondents were anxious and this number was sensitive to media reports and new case incidences; the level of anxiety peaked with high levels of case reporting and decreased over the next few months. Jones and Salathe (2009) reported that anxiety levels were fairly stable over the short duration of the study (1 week), but levels of public calm were sensitive to media reports.

2.4.2.5 H1N1 Vaccination Attitudes & Intentions

Attitudes toward Vaccination
Concerns regarding vaccine safety and side effects were amongst the most cited reasons for not choosing to vaccinate (Sypsa et al., 2009; Eastwood et al., 2010; Schwarzinger et al., 2010; Seale et al., 2010) with up to 71% of respondents using these reasons to explain their rationale to not vaccinate (Schwarzinger et al., 2010). Many also did not believe that the current H1N1 situation necessitated vaccination (Seale et al., 2010; Eastwood et al., 2010). Those who did intend to vaccinate did so for both self protection and protection of the community (Schwarzinger et al., 2010; Seale et al., 2010). However, less than half believed the H1N1 vaccination would be
effective (Seale et al., 2010; Lau et al., 2009b) and had not undergone enough clinical testing (Seale et al., 2010).

**Intention to Vaccinate**

International studies assessing the willingness to receive the 2009 H1N1 vaccine reported rates that ranged from 17% in France (Schwarzinger et al., 2010) to 55% in Australia (Seale et al., 2010) and the USA (Quinn et al., 2009). The proportion of those intending to vaccinate also decreased over time, likely due to the public’s declining perception of risk (Sypsa et al., 2009). Several factors such as cost, the presence of safety data, clinic location, and vaccine administrator were found to influence willingness to vaccinate (Lau et al., 2009b; Eastwood et al., 2010; Quinn et al., 2009). Increasing the out-of-pocket cost for the H1N1 vaccine was a detriment to uptake (Lau et al., 2009b). A lack of safety information presented at the time of vaccination also negatively affected willingness to vaccinate (Quinn et al., 2009; Lau et al., 2009b). If fact sheets on safety or efficacy information were not provided, only 5-15% of participants would proceed with their vaccination. Location was also a factor. 11% of previously willing people were no longer willing to vaccinate if vaccination clinics took place at community halls rather than at their GPs office (Eastwood et al., 2010). Public trust of health providers also played an influential role. 72% would get the H1N1 vaccine if it was recommended to them by their physician (Seale et al., 2010) and only 10% would proceed with vaccination if it was distributed by a non-health professional (Quinn et al., 2009).

Unsurprisingly, personal risk perception and prior vaccination history had a strong effect on intentions to vaccinate. Respondents with higher perceptions of personal risk were more likely to vaccinate (Sypsa et al., 2009; Eastwood et al., 2010; Seale et al., 2010; Schwarzinger et al., 2010). Those who had received vaccinations in the past were also more likely to vaccinate against H1N1 (Maurer et al., 2009; Quinn et al., 2009; Sypsa et al., 2009; Eastwood et al., 2010; Seale et al., 2010; Van et al., 2010; Schwarzinger et al., 2010).


2.4.2.6 Perceptions of Authorities

The public’s trust in, and approval of, the government and health authorities varied from region to region. Hong Kong citizens, who learned many pandemic lessons during the SARS outbreak in 2003, greatly supported the government and the vast majority believed any outbreak would be controlled (Lau et al., 2009a). Norwegians had similar levels of confidence in their health authorities (Woien & Tonsberg, 2009). In a UK telephone survey, Rubin et al. (2009) found that respondents on average highly agreed that the authorities were to be trusted, giving a mean score of 4 out of 5. Government ratings in America were positive, although less so. An online survey using a representative panel of Americans found that 30-40% had complete or very high confidence in the government’s ability to deal with pandemic in an open and honest fashion, while the majority (40-50%) trusted the government to a lesser extent (Quinn et al., 2009). Roughly half of Australians in a convenience sample believed health authorities would be truthful in handling H1N1, and slightly more trusted the government’s effectiveness (Seale et al., 2009). Trust in the authorities’ abilities was lower in India with only 33% believing that the government would be effective (Kamate et al., 2010).

In contrast, one trend that held across multiple studies in different geographic locations was the perceived level of hype around H1N1 as well as the truthfulness of incidence reporting. In Norway, India, and Australia, 40-50% of participants believed that health authorities were exaggerating the level of risk and danger regarding the pandemic (Woien & Tonsberg, 2009; Kamate et al., 2010; Seale et al., 2009). Almost 40% of respondents in both Hong Kong and Saudi Arabia believed that the government was hiding the true number of real cases (Balkhy et al., 2010; Lau et al., 2009b).

2.4.2.7 Information Needs & Sources

Seven studies reported on respondents’ information needs and sources. Not surprisingly, television, newspapers, and the internet were the primary ways of receiving information (Wong & Sam, 2010a; Kamate et al., 2010; Balkhy et al., 2010; Jones & Salathe, 2009). However, authoritative sources such as health care providers were not frequently referred to (Wong & Sam,
2010a; Balkhy et al., 2010). In a telephone survey of Chinese, Indians, and Malays in Malaysia, Wong and Sam (2010a) found that information needs and sources varied between these ethnic groups. Chinese and Indian respondents were more likely to get information from newspapers and preferred to be informed by health care professionals compared to Malays. Malays overall had less information needs and preferred to receive their news from the TV. A further analysis dividing low and high education participants found that those with lower education also preferred television as their information source and received less information overall.

In the United States, Jones & Salathe (2009) reported that the internet was the most common resource amongst respondents, with social network sites being the least frequently used for information. However, the statistical results supporting this finding are difficult to interpret and should be taken with caution. For those who were searching for H1N1 resources, the most sought after information was regarding H1N1 treatment and prevention (Wong & Sam, 2010a). Overall, citizens from the UK and Norway were satisfied with the information they had received and deemed it clear (Rubin et al., 2009; Woien & Tonsberg, 2009). However, this was not the case worldwide; only 55% of Australians (Seale et al., 2009) and one third of Indians (Kamate et al., 2010) believed they had enough information regarding the pandemic.

2.4.3 Summary

During the 2009-10 H1N1 pandemic, the literature reported that the public’s perception of risk and disease severity was low. Emotional responses and anxiety were also mild. The perceived level of threat was low and this was the likely reason that many chose not to make any specific behaviour changes to protect themselves against H1N1. Vaccine acceptance was varied, but the majority of those sampled had no intention of vaccinating against the epidemic. Trust in government and health authorities varied by region, but were mostly positive. Traditional mediums such as television and newspapers were the main source of pandemic information for most, although the internet played a growing role in the dissemination of information. The 2009 H1N1 pandemic is one of the first global epidemics to occur in the age of web 2.0 and social media. As such, it presents a unique opportunity to observe how the public are using tools such as Twitter in health emergencies. The following chapters report on the use of traditional content
analysis and novel infoveillance methods to analyze this user-generated content in order to a) inform public health authorities of public information demands and concerns during emergencies and b) demonstrate new methods using social media that may complement traditional public polling.
CHAPTER 3: RESEARCH DESIGN

3.1 Research Design Overview

The goals of this infodemiology project are to:

1) Track the public’s use terminology use during the pandemic as a measure of knowledge translation effectiveness, i.e., the use of recommended H1N1 versus colloquial swine flu terminology;
2) determine what Twitter users are communicating on Twitter, how they are expressing themselves, what information sources they are using, and if this content is changing over time;
3) explore the use of Twitter as a real-time content, sentiment, and public attention trend analysis and tracking tool.

The research design was split into three streams based on these research objectives (Figure 3-1). For the purposes of this study we limited our analysis to tweets from May to December 2009 in order to allot sufficient time and resources for data collection and analysis. We chose to analyze H1N1 content, qualifiers, and links in general and not focus on sub-concepts such as vaccination or misinformation as this project is the first part of a larger infodemiology study and as such we aimed to study H1N1 tweets broadly prior to investigating specific phenomena. We were particularly interested in tracking trends in the data over time as this information is important to noticing changes and the effects of specific events over time. As such, we chose to perform statistical analysis on both longitudinal and aggregated data to ensure that potential patterns in the data were not lost by aggregating data over our time frame. Tweet authorship, although potentially important, was not included in this study as it was not feasible for the scale of this project, although it will be investigated at a later point.
3.2 Data Collection & Database

Using the infoveillance system, Infovigil (Eysenbach, 2009), developed at the Centre for Global eHealth Innovation in Toronto, we are able to continuously gather and mine textual information from the internet. Twitter allowed us to access their server through the Application Programming Interface (API) every few seconds to gather new tweets. All public user-generated tweets were accessible when using this method. Tweets containing our keywords of interest were written into an internal relational database, together with metadata such as the profile name of the author and time/date stamp using Eastern Standard Time.

Between May 1 and December 31, 2009, we archived over 3 million tweets containing the keywords or hashtags (#) “H1N1”, “swine flu”, and “swineflu”. These three keywords were chosen at the beginning of data collection. In studying emergency events it is necessary to make rapid decisions regarding data collection, particularly when data is short-lived. In doing so,
decisions may be made before being sure of the scope and language being used in communications (Starbird et al., 2010). Consequently, other derivations may exist, however, the keywords that we chose likely represent the most relevant and common ways to refer to H1N1. The hashtag (#) convention is used in tweets to help other readers to understand the context of the tweet and to support keyword term searching. These hashtags can be used within the text, but are often times applied at the end of the tweet. Hashtag use was not required for our data collection. Non-English tweets were archived, but not included in our analyses as translation was not feasible.

Archival is crucial for analysis as Twitter itself does not archive tweets for more than a few days. In addition to recording tweets, starting in September 2009 we also archived the actual resources (web pages) cited in those tweets (i.e., not only the URL but also the content of the URL) using Web Cite (www.webcitation.org). Infovigil can also rank URLs by frequency per day. This database of tweets served as the primary dataset for this study. All statistical analyses used SPSS 16.0 or Microsoft Excel 2007. Tweets coding was documented using Excel 2007.

### 3.3 Knowledge Translation: H1N1 versus Swine Flu Terminology

To investigate the effectiveness of knowledge translation approaches to shift from “swine flu” to WHO-recommend “H1N1” terminology (WHO, 2009a), a linear regression for the proportion of tweets with H1N1 over time was performed using English-only tweets from May 1 to December 31, 2009. Tweets utilizing both “swine flu” and “H1N1” were counted towards the overall total, but not the proportion of H1N1 or swine flu tweets.

### 3.4 Manual Content & Sentiment Analysis

The qualitative manual coding of tweets commenced on Monday, May 11, 2009 as this represented the first set of complete data available at the time (we have since archived tweets back to May 1). In order to look for changing content over time at systematic periods, Mondays, 4 weeks apart were selected over the remainder of 2009 (May 11, June 8, July 6, August 3, August 31, September 28, October 26, November 23, and December 21). As we were mainly
interested in trends, we held the day of the week constant (Monday) to avoid finding any artificial peaks in the data caused by sampling from different days of the week as posting activity has been found to vary with different days of the week (Pear Analytics, 2009). Only 9 data points were selected for analysis due to limitations in time and resources. 25 tweets were randomly selected from every hour of the aforementioned days using the “RAND()” function in Excel. This method of sampling was chosen in order to avoid any time bias associated with posting. While several studies have looked at the use of Twitter in emergencies (e.g., Starbird & Palen, 2009), these studies analyzed the entire population of disaster-related tweets and no prior methodologies have been created for the sampling of tweets for representativeness. Consequently, we were unable to perform a formal sample size calculation. Instead, we chose our sample size based on feasibility and determined that a minimum sample size of 25 tweets per hour (or 600 tweets per day) would be sufficient to capture a “snapshot” for that day. Any re-posted or retweeted tweets that used the notation “RT @ username” or “RT@username” were excluded in order to prevent popular posts or spam from saturating the sample. In addition, retweets serve a specific recommendation purpose and as such may be systematically different from regular tweets. The retweet convention is generally used to attribute authorship to original tweet authors while re-broadcasting the tweet (Starbird and Palen, 2010). Non-English tweets were also excluded as translation was not feasible.

3.4.1 Codebook

A tri-axial coding scheme was created using an iterative process to reflect: 1) the tweet’s content, 2) how it was expressed (its qualifier), and 3) the type of link posted, if any. These macro categories were chosen prior to looking at the data and were based on the research goals. Preliminary coding of 1200 tweets was used to create the initial categories and codebook. The first pass of code creation used a ground-up approach where categories were emergent from the data. Upon review and discussion, infrequently used categories were collapsed into larger concepts, definitions were refined, and a subset of tweets (125) was coded by two raters to establish coding reliability (kappa). The last iteration of the codebook with criteria, definitions, and examples was finalized when a sufficient kappa level (0.7 and higher) was reached for each axis of the coding scheme.
3.4.2 Content Analysis

Each tweet was coded with one content category, following codebook definitions and hierarchies. If multiple qualifiers were present within a tweet, all applicable qualifiers were applied. Neutral or ambiguous statements were left uncoded. URLs were assessed by visiting each webpage and checking the “about us” section and disclaimers. If no information could be found, a Wikipedia and Google search were conducted for more details. In cases were web pages could not be easily attributed to one category, sites were categorized according to their lowest ranking type. For example, if a web page was either a “news blog” or “other website”, it was categorized as “other website”.

Tweets were categorized as misinformation if the information or opinion presented was unsubstantiated by our reference standards: the Centers for Disease Control and Public Health Agency of Canada for scientific claims and a panel of credible online news sources (e.g., CNN, BBC) for news-related claims. If resources gave confusing information, but not necessarily inaccurate information these were not classified as misinformation. If posts contained parodies or jokes, these were not classified as misinformation as the purpose of jokes is not to communicate fact or information, but to entertain.

3.4.3 Statistical Analysis

The chi-square test for trend was used to determine if the proportion of content, qualifiers, or links tweeted changed linearly over our analysis timeframe. This test assumes that one variable takes ordered categories (e.g., numerical or date order) and the second variable is dichotomous (e.g., yes or no) or also has ordered categories (Campbell, 2005). In our case, the tweet posting date took ordered categories as we proceed through the year (i.e., April comes before May and so forth) and the presence of specific content/qualifier/link categories was dichotomous (e.g., H1N1 resources are present/not present). Essentially, the chi-square test for trend evaluated whether or not the number of tweets present in a particular category was increasing or decreasing over our timeframe. Trends were considered to be significant if the $p$-value was less than 0.05. Prior to
testing for linearity, scatterplots were performed on each category in order to detect any non-linear patterns.

### 3.5 Automated Content & Sentiment Analysis

#### 3.5.1 Query Development

The Infovigil system was configured to allow for real-time analysis and visualization of the continuously archived Twitter database. To be consistent with the manual coding, the global search pattern was modified to filter out retweets that used the notation “RT@” or “RT @”. While tweet searches included data from May 1, 2009 to the present day, we only utilized data from the 9 selected days as comparison points with the manual coding. All qualifiers along with selected content categories (resources, personal experiences and personal opinions/interest) were transformed into concept queries. These concepts were selected as they were potentially the most feasible to translate into queries. Concepts not selected, such as advertisements and spam lacked specific keywords and vocabularies, while a search query for jokes/parodies would likely be very similar and therefore redundant to the query for the humour qualifier.

Initial search patterns (keywords or phrases) for each concept were derived from the codebook and an ongoing list of common phrases. Common misspellings, emoticons, internet slang, and variants of keywords were also included. Keywords were modified to include or exclude specified prefixes or preceding words (i.e., “not”, “don’t”, “un”). Groups of keywords within a concept were grouped into subconcepts if they reflected a common idea (e.g., concern for self) to refine the results and find any trends that may not be present when keywords were combined. Queries were not case-sensitive and utilized SQL syntax. At the time, the search engine lacked any natural language processing or full sentiment analysis capabilities. As such, the results from each keyword were manually audited to estimate its precision. Audits were primarily conducted by viewing the hits for each for three randomly selected days. Search patterns were modified or deleted if their results were too unspecific (approximately more than 30% of tweets did not reflect the concept). Audits were also conducted based on overall tweet volume. Large spikes
within concepts were reviewed to determine if spam campaigns or specific keywords were responsible for creating misleading patterns.

3.5.2 Validation Analysis

Concept query totals from the 9 selected days were recorded. Pearson’s correlations were used to measure the relationship between the proportions of selected categories resulting from the manual coding and the automated analyses. Automated proportions were obtained by taking the amount of tweets that were returned by a concept search query (e.g., tweets labelled as “personal experience”) and dividing by the total amount of tweets for that day. Correlations were considered significant if the $p$-value was less than 0.05. Additionally, chi-square tests for trend were used to determine if changes in the automated concepts over time were trending in the same direction and magnitude as in the manual coding.

3.6 Public Attention & Sentiment on Twitter

Infovigil automatically graphs the longitudinal results from the automated queries. These graphs were visually examined for large spikes in tweet volume (a potential indicator of public attention) and tweets on those days were qualitatively reviewed to see what media stories or external events influenced these peaks and were related to certain qualifiers or sentiments (i.e., concern, relief). For example, when the WHO made their pandemic level 6 announcement on June 11, 2009 (WHO, 2009b) we would expect to see a large spike in overall tweet volume due to this development. We would also expect to see a spike in “concerned” tweets in response to the announcement. For clarity, the largest peak in each longitudinal graph was scaled to 100 on the $y$-axis and all other peaks were plotted relative to that peak. In other words, values on the $y$-axis are not percentages. The denominator for each day was the total number of tweets with our given keywords (“H1N1”, “swine flu”, and “swineflu”).
3.7 Audit of Retweets

Both our manual and automated analysis excluded retweets (RTs). As these RTs may be systematically different from original or non-retweets (nonRTs), we performed a sub-analysis on this data. 3 RTs from every hour of the 9 selected days (12% of the manual sample) were manually coded using the same methodology described previously. This proportion was arbitrarily chosen. Chi-square tests were used to observe any difference in proportions between manually coded RT and nonRTs. Proportions were considered significantly different if the $p$-value was less than 0.05. Fisher’s exact tests were used in instances when cell counts were less than 5. Chi-square tests for trend were used to look for trends within the RTs and were compared to trends found in the nonRT results. Trends with a $p$-value of less than 0.05 were considered significant.

To compare RTs and nonRTs from the automated analysis, the queries for each of the 10 concepts were modified to include RTs in the search results. Longitudinal graphs of RT and nonRT results for each automated query concept were compared visually. Noticeable differences in graph shape or spike volume were noted. Tweets in these peaks were qualitatively examined to determine what content influenced the spike.
CHAPTER 4: RESULTS

4.1 Knowledge Translation: H1N1 versus Swine Flu Terminology

Between May 1 and December 31, 2009, the proportion of tweets using “H1N1” relative to the total number of tweets using “swine flu” or “H1N1” increased from 8.8% to 40.5% in an almost linear fashion ($R^2 = .788; p < .001$), indicating a gradual adoption of the WHO-recommended H1N1 terminology as opposed to “Swine Flu” (Figure 4-1). “H1N1” usage first became equal to “swine flu” usage on September 1.

**Figure 4-1: Absolute number (lines) and relative percentage (bars) of Tweets containing the keywords H1N1 or Swine Flu, between May and December 2009**

Blue = “swine flu” or “swineflu” Red = H1N1; Green = (“swine flu” or “swineflu”) AND H1N1.

4.2 Manual Content & Sentiment Analysis

4.2.1 Coding Scheme

After several codebook iterations, the inter-rater reliability (kappa) was estimated to be 0.80 for content, 0.74 for qualifiers, and 0.84 for links between two raters over 125 tweets.
In the first pass of coding, 31 content categories emerged from the data. These were grouped into six main content categories: resources (e.g., news), direct or indirect personal experiences (e.g., “I have swine flu”), personal reactions to or opinions (e.g., “I’m scared of H1N1”), jokes/parodies, marketing for H1N1-related products, and unrelated posts (spam). Examples and further descriptions can be found in Table 1.

In order to attribute only one content category to each tweet, a hierarchy was created. If information or resources were presented, but also commented on by the poster, the tweet was considered to be “personal opinions”. If the personal experiences that were shared were also humorous, the tweet was labelled as “personal experience” and not a joke. Humorous opinions or personal opinions were coded as “jokes”.

**Table 4-1: Descriptions and Examples of Content Categories**

<table>
<thead>
<tr>
<th>Content</th>
<th>Description</th>
<th>Example Tweets</th>
</tr>
</thead>
</table>
| Resource       | Tweet contains H1N1 news, updates, or information. May be the title or summary of the linked article. Contents may or may not be factual. | “China Reports First Case of Swine Flu (New York Times): A 30-year-old man who flew from St. Louis to Chengdu is... http://tinyurl.com/rdbhcg”
|                |                                                                             | “Ways To Prevent Flu http://tinyurl.com/r4l4cx #swineflu #h1n1”                 |
| Personal Experience | Twitter user mentions a direct (personal) or indirect (e.g., friend, family, co-worker) experience with the H1N1 virus or the social/economic effects of H1N1. | “Swine flu panic almost stopped me from going to US, but now back from my trip and so happy I went :-()”
|                |                                                                             | “Oh we got a swine flu leaflet. clearly the highlight of my day”
|                |                                                                             | “My sister has swine flu!”                                                   |
| Personal Opinions | Twitter user posts their opinion of or interest in the H1N1 virus/situation/news or expresses a need for or discovery of information. General H1N1 chatter or commentary. | “More people have died from Normal Flu than Swine flu, its just a media hoax, to take people's mind off the recession”
|                |                                                                             | “Currently looking up some info on H1N1”
|                |                                                                             | “Swine flu is scary!”                                                       |
| Jokes/Parody   | Tweet contains a H1N1 joke told via video, text, or photo; or a humourous opinion of H1N1 that does not refer to a personal experience. | “If you're an expert on the swine flu, does that make you Fluent?”           |
| Marketing      | Tweet contains an advertisement for an H1N1-related product or service.     | “Buy liquid vitamin C as featured in my video http://is.gd/y87r #health #h1n1” |
| Spam           | Tweet is unrelated to H1N1                                                 | “musicmonday MM lamarodom Yom Kippur Polanski Jay-Z H1N1 Watch FREE online LATEST MOVIES at http://a.gd/b1586f” |
Tweets that were not resource or spam-based were coded with a qualifier, if present. The codebook definitions of the 7 qualifiers took into consideration specific keywords and phrases, common internet expressions (e.g., “lol”), and emoticons (textual expressions representing a face or mood) (Table 4-2).

**Table 4-2: Descriptions and Examples of Qualifier Categories**

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Description</th>
<th>Example Emoticons or Internet Slang</th>
<th>Example Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humour or Sarcasm</td>
<td>Tweet is comedic or sarcastic.</td>
<td>:) ;( :P</td>
<td>“Rupert Grint had Swine Flu. It’s VOLDEMORTS COMEBACK!” “babysitting kids with h1n1, awesome. cant wait til thursday!!”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LOL, ROFL, j/k</td>
<td></td>
</tr>
<tr>
<td>Relief</td>
<td>Tweet expresses joy, happiness, or sense of peace.</td>
<td>:) =) :D</td>
<td>“in Egypt...without the swine flu...YAY!” “thank God my mom and bro don’t have that swine flu it was just a cold”</td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>Tweet attempts to de-emphasize the potential risk of H1N1 or bring it into perspective. May also express a lack of concern, disinterest, or indifference.</td>
<td>-_-</td>
<td>“90 ppl get swine flu &amp; everyone wants to wear surgical masks. 1 million ppl have aids &amp; nobody wants to wear a condom” “everyone needs to calm down about swine flu. it’ll all be ok.”</td>
</tr>
<tr>
<td>Concern</td>
<td>Tweet expresses H1N1-related fear, anxiety, worry, or sadness for self or others. May also express scepticism.</td>
<td>:( :@ :S OMG, r u ok?</td>
<td>“Hope the girl sitting on the tram next to me doesn’t have swine flu. She is sneezing all over the place” “my baby sister has the swine flu. :(” “omg i know someone who has the swine flu .. its so sad”</td>
</tr>
<tr>
<td>Frustration</td>
<td>Tweet expresses anger, annoyance, scorn, or volatile contempt. May contain course language or profanity.</td>
<td>&gt;:</td>
<td>&gt;:( FML, grr</td>
</tr>
<tr>
<td>Misinformation</td>
<td>Tweet contradicts the reference standard or contains unsubstantiated information. May make speculations or express distrust of authority or the media. May include conspiracy or doomsday theories.</td>
<td></td>
<td>“Swine Flu Vaccine will kill you! <a href="http://tinyurl.com/ph8sgx%E2%80%9D">http://tinyurl.com/ph8sgx”</a> “The swine flu is pharmaceutical conspiracy, a way to quick money”</td>
</tr>
<tr>
<td>Question</td>
<td>Tweet asks a question or contains a question mark.</td>
<td></td>
<td>“Why is it actually called #swineflu and not pigflu or hogflu?”</td>
</tr>
</tbody>
</table>
The original coding scheme for links had 20 URL categories. The majority of these were collapsed into “social networks”, “news blog, feed, or niche news”, and “other”. The final coding scheme categorized tweeted URLs into one of nine categories (Table 4-3).

### Table 4-3: Descriptions and Examples of Link Categories

<table>
<thead>
<tr>
<th>Link</th>
<th>Description</th>
<th>Example Webpage</th>
</tr>
</thead>
</table>
| Mainstream or Local News          | Webpage links to a local or international TV/radio/print/internet news service. | www.CNN.com  
|                                   |                                                                             | www.BBC.co.uk  
|                                   |                                                                             | www.Reuters.com                                      |
| News Blog, Feed, or Niche News    | Webpage primarily provides aggregated news content, news briefs, or has a specialized news focus. Webpage may allow for user-submitted articles to be published. | www.H1N1Alliance.com  
|                                   |                                                                             | www.MedicalNewsToday.com                              |
|                                   |                                                                             | www.SoccerNet.com                                    |
| Government or Public Health       | Webpage of a government or public health authority. Health professionals associations, private health organizations, and unofficial efforts are excluded. | www.CDC.gov  
|                                   |                                                                             | www.WHO.int                                           |
| Personal Blog                     | Personal webpage or blog that the user may post thoughts, opinions, or experiences. | www.blogger.com  
|                                   |                                                                             | www.wordpress.com                                    |
| Social Network                    | Webpage links to a social networking page, such as a user’s status update.    | www.Facebook.com                                      |
|                                   |                                                                             | www.MySpace.com                                       |
|                                   |                                                                             | www.Twitter.com                                        |
| Online Store                      | Webpage links to a store or advertisement.                                  | www.ebay.com                                          |
|                                   |                                                                             | www.Purell.com                                         |
| Other                             | Webpage is not described above. May be centered on providing services, editorials, media, or reference material. | www.YouTube.com                                        |
|                                   |                                                                             | www.PrisonPlanet.com                                   |
| No Reference                      | Tweet made claims or presented information without providing a URL.         | www.About.com                                          |
| Not Accessible                    | URL is not accessible                                                       |                                                      |

### 4.2.2 Content Analysis

We manually analyzed 5395 tweets for our content analysis (Table 4-4). The total number of tweets was short by five as we did not gather enough eligible tweets on September 28 for analysis. In other words, not enough tweets were available. H1N1 resources were the most common type of content shared (52.6%), followed by personal experiences (22.5%). 39% of tweets were coded with 1 or more qualifiers. Tweets expressing humour (12.7%), concern (11.7%), and questions (10.3%) were the most common. We classified 4.5% of tweets as possible misinformation, misleading, or highly speculative. 61.8% of all tweets had links; 23.2%
of all posts linked to a news website, while links to government and public health agencies were not commonly shared (1.5%). 90.2% of tweets provided links when a reference was necessary (i.e., tweet was providing information).

Table 4-4: Content, Qualifiers, and Links of Manually Coded H1N1 Tweets

<table>
<thead>
<tr>
<th></th>
<th>May 11 (n=600)</th>
<th>June 8 (n=600)</th>
<th>July 6 (n=600)</th>
<th>Aug 3 (n=600)</th>
<th>Aug 31 (n=600)</th>
<th>Sept 28* (n=595)</th>
<th>Oct 26 (n=600)</th>
<th>Nov 23 (n=600)</th>
<th>Dec 21 (n=600)</th>
<th>Total (5395)</th>
<th>p Value, trend</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resources</td>
<td>291 (49)</td>
<td>330 (55)</td>
<td>252 (42)</td>
<td>284 (47)</td>
<td>352 (59)</td>
<td>302 (51)</td>
<td>296 (49)</td>
<td>366 (61)</td>
<td>367 (61)</td>
<td>2840 (53)</td>
<td>.001*</td>
</tr>
<tr>
<td>Personal Experiences</td>
<td>107 (18)</td>
<td>119 (20)</td>
<td>140 (23)</td>
<td>150 (25)</td>
<td>94 (16)</td>
<td>158 (27)</td>
<td>176 (29)</td>
<td>138 (23)</td>
<td>132 (22)</td>
<td>1214 (23)</td>
<td>.01*</td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>81 (14)</td>
<td>80 (14)</td>
<td>147 (25)</td>
<td>93 (16)</td>
<td>76 (13)</td>
<td>82 (14)</td>
<td>78 (13)</td>
<td>51 (9)</td>
<td>52 (9)</td>
<td>740 (14)</td>
<td>.001*</td>
</tr>
<tr>
<td>Jokes</td>
<td>100 (17)</td>
<td>53 (9)</td>
<td>50 (8)</td>
<td>52 (9)</td>
<td>45 (8)</td>
<td>28 (5)</td>
<td>28 (5)</td>
<td>30 (5)</td>
<td>35 (6)</td>
<td>421 (8)</td>
<td>.001*</td>
</tr>
<tr>
<td>Marketing</td>
<td>7 (1)</td>
<td>10 (2)</td>
<td>6 (1)</td>
<td>15 (3)</td>
<td>10 (2)</td>
<td>2 (0)</td>
<td>2 (0)</td>
<td>9 (2)</td>
<td>11 (2)</td>
<td>72 (1)</td>
<td>.72</td>
</tr>
<tr>
<td>Spam</td>
<td>14 (2)</td>
<td>8 (1)</td>
<td>5 (1)</td>
<td>6 (1)</td>
<td>23 (4)</td>
<td>23 (4)</td>
<td>20 (3)</td>
<td>6 (1)</td>
<td>3 (1)</td>
<td>108 (2)</td>
<td>.91</td>
</tr>
<tr>
<td><strong>Qualifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humour</td>
<td>129 (22)</td>
<td>81 (14)</td>
<td>72 (12)</td>
<td>87 (15)</td>
<td>59 (10)</td>
<td>65 (11)</td>
<td>63 (11)</td>
<td>58 (10)</td>
<td>73 (12)</td>
<td>687 (13)</td>
<td>.001*</td>
</tr>
<tr>
<td>Relief</td>
<td>13 (2)</td>
<td>3 (1)</td>
<td>6 (1)</td>
<td>10 (2)</td>
<td>7 (1)</td>
<td>5 (1)</td>
<td>15 (3)</td>
<td>9 (2)</td>
<td>13 (2)</td>
<td>81 (2)</td>
<td>.18</td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>17 (3)</td>
<td>14 (2)</td>
<td>10 (2)</td>
<td>19 (3)</td>
<td>7 (1)</td>
<td>7 (1)</td>
<td>20 (3)</td>
<td>11 (2)</td>
<td>1 (0)</td>
<td>106 (2)</td>
<td>.014*</td>
</tr>
<tr>
<td>Concern</td>
<td>47 (8)</td>
<td>57 (10)</td>
<td>105 (18)</td>
<td>85 (14)</td>
<td>69 (12)</td>
<td>81 (14)</td>
<td>76 (13)</td>
<td>57 (10)</td>
<td>56 (9)</td>
<td>633 (12)</td>
<td>.677</td>
</tr>
<tr>
<td>Frustration</td>
<td>25 (4)</td>
<td>27 (5)</td>
<td>29 (5)</td>
<td>32 (5)</td>
<td>23 (4)</td>
<td>33 (6)</td>
<td>18 (3)</td>
<td>13 (2)</td>
<td>12 (2)</td>
<td>212 (4)</td>
<td>.002*</td>
</tr>
<tr>
<td>Misinform.</td>
<td>23 (4)</td>
<td>13 (2)</td>
<td>13 (2)</td>
<td>55 (9)</td>
<td>37 (6)</td>
<td>34 (6)</td>
<td>28 (5)</td>
<td>26 (4)</td>
<td>14 (2)</td>
<td>243 (5)</td>
<td>.756</td>
</tr>
<tr>
<td>Question</td>
<td>66 (11)</td>
<td>60 (10)</td>
<td>76 (13)</td>
<td>59 (10)</td>
<td>63 (11)</td>
<td>47 (8)</td>
<td>84 (14)</td>
<td>49 (8)</td>
<td>51 (9)</td>
<td>555 (10)</td>
<td>.125</td>
</tr>
<tr>
<td><strong>Links</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News Websites</td>
<td>141 (24)</td>
<td>159 (27)</td>
<td>119 (20)</td>
<td>142 (24)</td>
<td>172 (29)</td>
<td>140 (24)</td>
<td>138 (23)</td>
<td>126 (21)</td>
<td>116 (19)</td>
<td>1253 (23)</td>
<td>.04*</td>
</tr>
<tr>
<td>News Blogs, Feeds, Niches</td>
<td>72 (12)</td>
<td>76 (13)</td>
<td>40 (7)</td>
<td>62 (10)</td>
<td>76 (13)</td>
<td>47 (8)</td>
<td>69 (12)</td>
<td>79 (13)</td>
<td>111 (19)</td>
<td>632 (12)</td>
<td>.001*</td>
</tr>
<tr>
<td>Government or Public Health</td>
<td>9 (2)</td>
<td>8 (1)</td>
<td>11 (2)</td>
<td>9 (2)</td>
<td>5 (1)</td>
<td>10 (2)</td>
<td>11 (2)</td>
<td>8 (1)</td>
<td>10 (2)</td>
<td>81 (2)</td>
<td>.826</td>
</tr>
<tr>
<td>Personal Blogs</td>
<td>8 (1)</td>
<td>14 (2)</td>
<td>20 (3)</td>
<td>7 (1)</td>
<td>11 (2)</td>
<td>5 (1)</td>
<td>16 (3)</td>
<td>12 (2)</td>
<td>6 (1)</td>
<td>99 (2)</td>
<td>.348</td>
</tr>
<tr>
<td>Social Networks</td>
<td>6 (1)</td>
<td>13 (2)</td>
<td>9 (2)</td>
<td>12 (2)</td>
<td>6 (1)</td>
<td>19 (3)</td>
<td>26 (4)</td>
<td>19 (3)</td>
<td>22 (4)</td>
<td>132 (2)</td>
<td>.001*</td>
</tr>
<tr>
<td>Online Stores</td>
<td>12 (2)</td>
<td>1 (0)</td>
<td>7 (1)</td>
<td>17 (3)</td>
<td>8 (1)</td>
<td>1 (0)</td>
<td>1 (0)</td>
<td>6 (1)</td>
<td>13 (2)</td>
<td>66 (1)</td>
<td>.668</td>
</tr>
<tr>
<td>Other</td>
<td>39 (7)</td>
<td>33 (6)</td>
<td>41 (7)</td>
<td>48 (8)</td>
<td>75 (13)</td>
<td>73 (12)</td>
<td>45 (8)</td>
<td>92 (15)</td>
<td>64 (11)</td>
<td>510 (10)</td>
<td>.001*</td>
</tr>
<tr>
<td>No Reference</td>
<td>23 (4)</td>
<td>56 (9)</td>
<td>78 (13)</td>
<td>50 (8)</td>
<td>32 (5)</td>
<td>26 (4)</td>
<td>19 (3)</td>
<td>25 (4)</td>
<td>17 (3)</td>
<td>326 (6)</td>
<td>.001*</td>
</tr>
<tr>
<td>Not Accessible</td>
<td>37 (6)</td>
<td>30 (5)</td>
<td>13 (2)</td>
<td>14 (2)</td>
<td>29 (5)</td>
<td>20 (3)</td>
<td>14 (2)</td>
<td>33 (6)</td>
<td>45 (8)</td>
<td>235 (4)</td>
<td>.20</td>
</tr>
</tbody>
</table>

*2300 hrs on Sept. 28 had only 20 eligible tweets.

The percent total will not equal 100% as not all tweets had qualifiers and tweets with multiple qualifiers were coded multiple times. Approximately 40% of tweets contained qualifiers.

The total will not equal to 100% as not all tweets had links. Approximately 60% of tweets contained links.

*chi-square test for trend significant at p < 0.05.

The chi-square test for trend showed several linear trends in the data (Table 4-4). The proportion of tweets containing resources and personal experiences increased over time while the amount of
jokes and personal opinions decreased. Tweets that expressed humour, frustration, or downplayed the risk from H1N1 became less common. Linking behaviour also changed during the course of the pandemic. News websites were cited significantly less, while references to news blogs/feeds/niches, social networks, and other web pages increased. No significant trends were found for misinformation, but the data exhibited a non-linear pattern (Figure 4-2).

Figure 4-2: Non-linear pattern of tweeted misinformation identified via manual coding from May to December 2009

4.3 Automated Content & Sentiment Analysis

4.3.1 Content Analysis

Table 4-5 presents examples of search patterns, internet slang, and emoticons used to develop the queries. (A full list of detailed concept queries can be found in Appendix 4-1). The results of the automated content analysis can be found in Table 4-6. Resources were the most commonly tweeted content (66%). Personal experiences and opinions were less common (3.5%, 2.9%, respectively). Questions were the most common qualifier (13.24%) followed by concern (8.7%). 0.9% of tweets were categorized as misinformation.
### Table 4-5: Descriptions and Examples of Automated Coding Search Patterns

<table>
<thead>
<tr>
<th>Concept</th>
<th>Example Keywords</th>
<th>Example Emoticons</th>
<th>Example Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humour/Sarcasm</td>
<td>Lol, haha, j/k</td>
<td>:P</td>
<td>“when pigs fly”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“just kiddin”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“bacon flu”</td>
</tr>
<tr>
<td>Concern</td>
<td>Omg, afraid, worried</td>
<td>:(</td>
<td>“freaking out”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“uh oh”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“stay away”</td>
</tr>
<tr>
<td>Frustration</td>
<td>annoy, stupid, angry</td>
<td>&gt;:]</td>
<td>“swine flu sucks”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“hate swine flu”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“so sick of”</td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>overblown, hype, hysteria</td>
<td></td>
<td>“calm down”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“not a big deal”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“forget about swine flu”</td>
</tr>
<tr>
<td>Relief</td>
<td>whew, grateful, thankful</td>
<td></td>
<td>“thank God”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“glad to hear”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“feeling better”</td>
</tr>
<tr>
<td>Misinformation</td>
<td>conspiracy, toxin, autism</td>
<td></td>
<td>“mind control”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“chemical warfare”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“real story”</td>
</tr>
<tr>
<td>Personal Experiences</td>
<td>my (mom, co-worker, classmate, teacher, etc.)</td>
<td></td>
<td>“went to get my swine flu shot”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“feeling sick”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“going to a clinic”</td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>imho, heard, reading</td>
<td></td>
<td>“I believe”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“government should”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“I feel that”</td>
</tr>
<tr>
<td>Resources</td>
<td>http://, https://</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4-6: Content and Qualifiers of Automatically Coded H1N1 Tweets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resources</td>
<td>8,903 (53)</td>
<td>3,687 (51)</td>
<td>3,576 (41)</td>
<td>3,666 (52)</td>
<td>7,010 (61)</td>
<td>6,172 (60)</td>
<td>9,000 (49)</td>
<td>5,590 (59)</td>
<td>3,307 (66)</td>
<td>50,911 (54)</td>
</tr>
<tr>
<td>Personal Experiences</td>
<td>375 (2)</td>
<td>237 (3)</td>
<td>357 (4)</td>
<td>266 (4)</td>
<td>307 (3)</td>
<td>381 (4)</td>
<td>853 (5)</td>
<td>350 (4)</td>
<td>170 (3)</td>
<td>3,296 (4)</td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>486 (3)</td>
<td>179 (3)</td>
<td>350 (4)</td>
<td>233 (3)</td>
<td>286 (3)</td>
<td>280 (3)</td>
<td>640 (4)</td>
<td>214 (2)</td>
<td>90 (2)</td>
<td>2,758 (3)</td>
</tr>
<tr>
<td><strong>Qualifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humour</td>
<td>1,418 (9)</td>
<td>582 (8)</td>
<td>725 (8)</td>
<td>443 (6)</td>
<td>581 (5)</td>
<td>634 (6)</td>
<td>916 (5)</td>
<td>484 (5)</td>
<td>243 (7)</td>
<td>6,026 (6)</td>
</tr>
<tr>
<td>Relief</td>
<td>67 (0)</td>
<td>33 (1)</td>
<td>87 (1)</td>
<td>47 (1)</td>
<td>34 (0)</td>
<td>63 (1)</td>
<td>111 (1)</td>
<td>38 (0)</td>
<td>15 (0)</td>
<td>495 (1)</td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>224 (3)</td>
<td>56 (2)</td>
<td>59 (2)</td>
<td>54 (3)</td>
<td>78 (1)</td>
<td>84 (1)</td>
<td>187 (3)</td>
<td>53 (2)</td>
<td>60 (0)</td>
<td>858 (1)</td>
</tr>
<tr>
<td>Concern</td>
<td>964 (14)</td>
<td>630 (11)</td>
<td>1,546 (16)</td>
<td>685 (13)</td>
<td>1,144 (12)</td>
<td>872 (12)</td>
<td>1,522 (15)</td>
<td>594 (13)</td>
<td>252 (12)</td>
<td>8,209 (9)</td>
</tr>
<tr>
<td>Frustration</td>
<td>498 (3)</td>
<td>253 (4)</td>
<td>295 (4)</td>
<td>195 (3)</td>
<td>305 (3)</td>
<td>309 (3)</td>
<td>648 (4)</td>
<td>249 (3)</td>
<td>109 (2)</td>
<td>2,858 (3)</td>
</tr>
<tr>
<td>Misinformation</td>
<td>116 (1)</td>
<td>25 (0)</td>
<td>39 (1)</td>
<td>100 (1)</td>
<td>84 (1)</td>
<td>151 (1)</td>
<td>209 (1)</td>
<td>73 (1)</td>
<td>47 (1)</td>
<td>844 (1)</td>
</tr>
<tr>
<td>Question</td>
<td>2,253 (14)</td>
<td>808 (11)</td>
<td>1365 (16)</td>
<td>934 (13)</td>
<td>1388 (12)</td>
<td>1172 (12)</td>
<td>2759 (15)</td>
<td>1184 (13)</td>
<td>585 (12)</td>
<td>12,448 (13)</td>
</tr>
</tbody>
</table>

4.3.2 Automated Coding Trends

Chi-square tests for trend found that all 3 content concepts and 4 out of 7 qualifier concepts displayed significant linear trends over our timeframe (Table 4-7). The content categories all trended in the same direction as in the manual coding. Tweets that had humour/sarcasm or downplayed the risk of H1N1 also had the same downward trends as in the manual analysis. Trends for misinformation and concern were unique to the automated coding. Although a downward trend for frustration was found in the manual coding, no such trend was found in the automated analysis.
Table 4-7: Automated Coding Trends over Time

<table>
<thead>
<tr>
<th>Concept</th>
<th>$\chi^2$ test for trend (df =1)</th>
<th>p value</th>
<th>Automated coding trend over time</th>
<th>Manual coding trend over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources</td>
<td>252.98</td>
<td>&lt;.001*</td>
<td>➤</td>
<td>➤</td>
</tr>
<tr>
<td>Personal Experience</td>
<td>73.83</td>
<td>&lt;.001*</td>
<td>➤</td>
<td>➤</td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>6.42</td>
<td>.011*</td>
<td>➯</td>
<td>➯</td>
</tr>
<tr>
<td>Humour &amp; Sarcasm</td>
<td>292.54</td>
<td>&lt;.001*</td>
<td>➯</td>
<td>➯</td>
</tr>
<tr>
<td>Relief</td>
<td>.41</td>
<td>.522</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>7.85</td>
<td>.005*</td>
<td>➯</td>
<td>➯</td>
</tr>
<tr>
<td>Frustration</td>
<td>2.64</td>
<td>.104</td>
<td></td>
<td>➯</td>
</tr>
<tr>
<td>Concern</td>
<td>25.47</td>
<td>&lt;.001*</td>
<td>➯</td>
<td>➯</td>
</tr>
<tr>
<td>Misinformation</td>
<td>13.66</td>
<td>&lt;.001*</td>
<td>➤</td>
<td>➤</td>
</tr>
<tr>
<td>Question</td>
<td>.51</td>
<td>.477</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* chi-square test for trend significant at $p < 0.05$

4.3.3 Validation

7 out of 10 automated queries were found to correlate significantly well with the results of the manual coding (Table 4-8). Correlations were particularly high for personal experiences ($r = 0.91$), concern ($r = 0.87$), and personal opinion/interest ($r = 0.86$) (Figure 4-3). Queries for relief, frustration, and downplayed risk were not significantly correlated to the manual coding.

Table 4-8: Correlations between Manual and Automated Coding

<table>
<thead>
<tr>
<th>Concept</th>
<th>Pearson Correlation ($r$)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources</td>
<td>.83</td>
<td>.006*</td>
</tr>
<tr>
<td>Personal Experiences</td>
<td>.91</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>.86</td>
<td>.003*</td>
</tr>
<tr>
<td>Humour/Sarcasm</td>
<td>.70</td>
<td>.037*</td>
</tr>
<tr>
<td>Relief</td>
<td>-.38</td>
<td>.312</td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>.08</td>
<td>.845</td>
</tr>
<tr>
<td>Concern</td>
<td>.87</td>
<td>.002*</td>
</tr>
<tr>
<td>Frustration</td>
<td>.45</td>
<td>.228</td>
</tr>
<tr>
<td>Misinformation</td>
<td>.815</td>
<td>.007*</td>
</tr>
<tr>
<td>Question</td>
<td>.83</td>
<td>.005*</td>
</tr>
</tbody>
</table>

N = 9, * significant at $p < 0.05$
4.4 Public Attention & Sentiment on Twitter

The figures presented in this section are screenshots taken from Infovigil. For clarity, the largest peak in each longitudinal graph was scaled to 100 on the y-axis and all other peaks were plotted relative to that peak. In other words, values on the y-axis are not percentages. The denominator
for each day was the total number of tweets with our given keywords (“H1N1”, “swine flu”, and “swineflu”). Sharp increases in absolute H1N1-related tweet volume coincided with major H1N1 news events, for example, a large peak on June 11 (Figure 4-1) corresponded to the WHO’s Pandemic Level 6 announcement on that same day (WHO, 2009b). The volume of humorous tweets also decreased on this day (drop A in Figure 4-4) and the number of frustrated tweets increased (peak A in Figure 4-5).

**Figure 4-4: The proportion of tweets expressing humour**

![Figure 4-4](image)

A = June 11: WHO pandemic level 6 announcement  
1 = July 5: Harry Potter actor Rupert Grint has H1N1

**Figure 4-5: The proportion of tweets expressing frustration**

![Figure 4-5](image)

A = June 11: WHO pandemic level 6 announcement  
1 = Aug 25, Sept 2, Sept 9: increases in profanity directed at H1N1 and spam campaigns

As seen in Figure 4-6, which depicts the proportion of tweets sharing personal experiences, the October to November peak directly coincides with the second wave of H1N1 in North America (CDC, 2010b). Similarly, when personal experiences were further broken down into sub-concepts for more detail, tweet volume of vaccination experiences (e.g., “I got my H1N1 shot today”) increased rapidly following the arrival of H1N1 vaccinations in the United States on October 6 (CNN, 2009).
Figure 4-6: The proportion of tweets sharing personal experiences

A = June 11: WHO pandemic level 6 announcement
1 = October 6: H1N1 vaccinations arrive in the US
Subconcepts: Red = indirect experience; Yellow = personal/direct experience; Blue = vaccination experience

Tweets expressing concern had one outstanding peak on July 5 (peak 1 in Figure 4-7). This spike coincided with a news story that one of the actors from the “Harry Potter” movies was recovering from H1N1 (CBC, 2009a). The amount of humour (peak 1 in Figure 4-4) and relief (peak 1 in Figure 4-8) also increased in response to this story.

Figure 4-7: The proportion of tweets expressing concern

A = June 11: WHO pandemic level 6 announcement
1 = July 5: Harry Potter actor Rupert Grint has H1N1
Subconcepts: Blue = concern for others; Red = concern for self; Yellow = concerned emoticons; Green = general concern.
Figure 4-8: The proportion of tweets expressing relief

A = June 11: WHO pandemic level 6 announcement
1 = July 5: Harry Potter actor Rupert Grint has H1N1
2 = Oct 15: “Swine flu 6 months later: relief, but winter looms” AP article
3 = Oct 22: “Swept by swine flu? NFL policy gives teams relief” AP article
4 = Nov 4: Spam campaign “Family-created iPhone game raises $$ for CDC's H1N1 relief fund!”
5 = Nov 25-27: #iamthankfulfor campaign

Misinformation had several large peaks in our timeframe (Figure 4-9). The largest peak (peak 2) appeared from September 18-21 with the circulation of a story listing the “ten swine flu lies told by the mainstream media” (Adams, 2009). Another large peak (peak 3) appeared on November 27 due to a Danish conspiracy news story that claimed the WHO and drug companies were in collusion regarding H1N1 (Financial Armageddon, 2009). Several other peaks (peaks 1 and 4) were not the result of true misinformation or speculation, rather the popular news stories on those days had keywords associated with the query for misinformation (Times of India, 2009; Age of Autism, 2009). Headline-influenced peaks occurred for relief as well (Figure 4-8, see peaks 2 and 3).

Figure 4-9: The proportion of tweets expressing misinformation

A = June 11: WHO pandemic level 6 announcement
1 = Aug 2: CBS reports on parental concerns about H1N1
2 = Sept 18-21: Ten swine flu lies told by the mainstream media
3 = Nov 27: WHO and drug companies in collusion
4 = Dec 25: Carbon monoxide poisoning can create same symptoms as H1N1
Viral dissemination of campaigns on Twitter resulted in several large spikes. One campaign comparing the perceived need for facemasks versus condoms (“90 people get swine flu & everybody wants to wear a mask. A million people have AIDS & nobody wants to wear a condom”) was responsible for two large peaks in downplayed risk on July 20 and December 1 (peak 1 and 2 in Figure 4-10). The “#iamthankfulfor” campaign, taking place between November 25-27 (American Thanksgiving weekend), resulted in the largest peak of tweets expressing relief (peak 5 in Figure 4-8). In this campaign, users posted items they were thankful for, which in our data was related to getting the H1N1 vaccine and not becoming infected.

Another notable campaign was the “#oink” movement on August 16 to support the pork industry and farmers by urging the media and public to utilize “H1N1” terminology over “swine flu” (Pork Magazine, 2009). This event resulted in the number tweets using H1N1 increasing and those using swine flu decreasing. In one case, viral dissemination of new information caused a large activity spike of tweets (peak 1 in Figure 4-11). On September 8, Twitter was used to report the discovery of the first confirmed H1N1 case at a videogame convention in Seattle and urged symptomatic attendees to seek medical advice (Mastrapa, 2009).

**Figure 4-10: The proportion of tweets expressing downplayed risk**

A = June 11: WHO pandemic level 6 announcement
1 = July 20: Viral dissemination of the “face mask (H1N1) versus condom (AIDS) comparison”
2 = Dec 1: Viral dissemination of the “face mask (H1N1) versus condom (AIDS) comparison”
Figure 4-11: The proportion of tweets sharing personal opinions & interest

A = June 11: WHO pandemic level 6 announcement
1 = Sept 8: Case of H1N1 confirmed at PAX videogame convention in Seattle
2 = Oct 25: Unexplained peak

The largest volume of questions posted on Twitter was in reference to the WHO pandemic level 6 announcement (peak A), the “Harry Potter” actor illness (peak 1), and the facemask versus condom campaign (peak 2) (Figure 4-12). An unexplained significant drop in questions occurred on August 5 (drop 3).

Figure 4-12: The proportion of tweets sharing questions

A = June 11: WHO pandemic level 6 announcement
1 = July 5: Harry Potter actor Rupert Grint has H1N1
2 = July 20: Viral dissemination of the “face mask (H1N1) versus condom (AIDS) comparison”
3 = August 5: Unexplained drop in questions

Within resources, no major peaks could be detected. However, there is a consistent upward trend of tweeted resources from May to December (Figure 4-13). This trend was also present in the manual coding.
4.5 Retweet Analysis

A total of 642 RTs were coded. The expected total of 648 was short by 6 tweets as not enough RTs were found on September 28. The manual coding of RTs found that the proportion of tweets sharing personal experiences was significantly less compared to nonRTs ($\chi^2(1) = 11.45, p = .001$) (Figure 4-14). No other significant differences in the aggregate data were found. Aggregate comparisons between RT and nonRTs for content, qualifiers, and links are displayed in Figure 4-14, 4-15, and 4-16, respectively.

Figure 4-14: Comparison of content proportions between RTs and nonRTs

$* = P < 0.05$ using $\chi^2$ test for trend
Only three significant trends were detected within RTs over the 9 selected days (Table 4-9). Chi-square test for trends found significant downward trends for jokes ($\chi^2(1) = 6.83, p = .009$)
and humour ($\chi^2(1) = 6.46, p = .011$), matching the non-RT trends for these categories. The only other trend found was an upward trend in links to government or public health websites ($\chi^2(1) = 11.77, p = .001$). This trend was unique to RTs.

**Table 4-9: RT Trends over Time**

<table>
<thead>
<tr>
<th>Concept</th>
<th>$\chi^2$ test for trend (df =1)</th>
<th>p value</th>
<th>RT trend over time</th>
<th>Non-RT trend over time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resources</td>
<td>1.05</td>
<td>.305</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Experience</td>
<td>.08</td>
<td>.779</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Opinions</td>
<td>.22</td>
<td>.639</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jokes/Parody</td>
<td>6.83</td>
<td>.009*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing</td>
<td>.81</td>
<td>.369</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spam</td>
<td>.36</td>
<td>.551</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Qualifiers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humour &amp; Sarcasm</td>
<td>6.46</td>
<td>.011*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relief</td>
<td>.16</td>
<td>.694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downplayed Risk</td>
<td>3.49</td>
<td>.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frustration</td>
<td>.29</td>
<td>.589</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concern</td>
<td>.03</td>
<td>.873</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misinformation</td>
<td>3.38</td>
<td>.066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>3.69</td>
<td>.055</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Links</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News Websites</td>
<td>.64</td>
<td>.424</td>
<td></td>
<td></td>
</tr>
<tr>
<td>News Blogs, Feeds, or Niche News</td>
<td>3.29</td>
<td>.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government or Public Health</td>
<td>11.77</td>
<td>.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Blogs</td>
<td>.45</td>
<td>.504</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Networks</td>
<td>.50</td>
<td>.481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online Stores</td>
<td>.76</td>
<td>.382</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>.08</td>
<td>.781</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Reference Provided</td>
<td>.79</td>
<td>.374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Accessible</td>
<td>1.21</td>
<td>.271</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparisons between longitudinal graphs of RT and nonRT results for automated queries found only minor tweet volume changes in a few concepts. When RTs were included, tweet activity spikes in personal opinions (peak 1 in Figure 4-9), downplayed risk (peak 2 in Figure 4-10), and
misinformation (peak 1, 3, and 4 in Figure 4-11) increased in volume when RTs were included. An audit of these peaks found that the RTs propagated the most popular tweet content of the day (see respective figure legends).
CHAPTER 5: DISCUSSION & CONCLUSION

5.1 Overview of Principle Findings

The results of our content analysis indicate that from May to December 2009, H1N1-related content, attitudes, and information seeking behaviour shared on Twitter varied over time. H1N1 resources were the most commonly tweeted content (52.6%) and the majority received their information from mainstream or local media sources (23.2%). Misinformation was not commonly shared (4.5%) and government authorities were not directly referred to by the public (1.5%). The majority of our automated analyses correlated well with our manual results and found that Twitter activity spikes were influenced by news events.

5.2 Knowledge Translation: H1N1 versus Swine Flu Terminology

The proportion of tweets using the term “H1N1” increased compared to the relative usage of “swine flu”, demonstrating the gradual adoption of WHO-recommended terminology by the public and media on Twitter. Although these findings may indicate that the knowledge translation strategy for proper terminology use was effective overall, it is likely that the media’s and not the public’s adoption of “H1N1” was the primary reason for this trend as the amount of tweeted resources and news articles increased over time while the amount of jokes and personal comments decreased. In other words, the volume of public or personal posting in regards to H1N1 decreased as the outbreak continued and the media were producing the majority of H1N1-related content. As such, Twitter appeared to serve as mainly a news re-broadcasting tool. However, the importance of the media’s terminology choice cannot be underestimated as they hold much influence as major information transmitters and word choice can be key in encouraging or discouraging certain risk behaviours (Kallan et al., 1997). In one notable example, the public and the pork industry worked together to increase awareness and utilization of “H1N1” terminology via the #oink campaign. This viral technique had a significant impact on H1N1 posting activity on that day with the use of “H1N1” increasing sharply and “swine flu” decreasing rapidly.
5.3 Tweet Content & Public Attention

In our manual coding we found that news and information were the most commonly tweeted H1N1-related material (52.6%). This proportion is much higher than the average, as it has been previously estimated that only 3.6% of all tweets share news stories (Pear Analytics, 2009). However, there are significant differences between our sample and the one previously studied; in our case H1N1-related tweets are a subset of the greater tweet population and our topic was under intense media coverage while the other study looked at general tweets. Our results do correspond with a study of Twitter use during Hurricane Gustav and Ike, where roughly half of all hurricane-related tweets contained URLs, which can be interpreted as the amount of web resources shared (Hughes & Palen, 2009). Additionally, a content analysis of forum posts during the 2008 Sichuan Earthquake found that the majority of posts were information-related and these posts were the most commonly viewed and commented on (Qu et al., 2000). Collectively, our findings highlight the information dissemination role that Twitter and other social networking tools can play in rapid, widespread communication in emergencies.

The change of tweet content over our timeframe was not unexpected. It seems reasonable that public posting behaviour would change in response to external events such as the increasing incidence rate of H1N1 and vaccination rollout. H1N1 surveys that reported longitudinal results using traditional methods found that public behaviour and attitudes varied over the course of the epidemic. In these studies, public concern and engagement in protective behaviours increased when the threat of the outbreak increased and decreasing when the perceived risk declined (Sypsa et al., 2009; Wong & Sam, 2010b; Van et al., 2010). Longitudinal studies conducted during the SARS outbreak in 2003 also had similar findings (Leung et al., 2005; Lau et al., 2003). Similarly, in our data we found that personal accounts of H1N1 increased over time, while the amount of humorous comments decreased, possibly due to the increasing perceived seriousness of the situation and/or the declining popularity of the subject.

More minute changes (e.g., peaks emerging over the course of a few days) were also observed and were highly influenced by the media and external events. Examples of this included the large spike in tweets that resulted from the WHO pandemic level 6 announcement on June 11,
2009 and the two peaks in personal experiences that coincide with the first and second wave of H1N1 in North America. Similarly, Starbird et al. (2010) found that tweet volume related to the Red River flooding in 2009 increased when the threat was largest and Hughes & Palen (2009) reported similar findings for Hurricane Gustov and Ike-related Twitter activity. These results indicate that perceived severity of the threat and intense news coverage are likely factors that dictate tweet dissemination and posting activity.

Our longitudinal automated analysis also demonstrated the potential to qualitatively examine tweet content to see what story has captured the online public’s attention and what sentiments those stories evoke. For example, in our results the illness of one of the “Harry Potter” actors was widely circulated on Twitter and was found to elicit feelings of concern, humour, and relief. However, the interpretation of peaks is not always clear, particularly when there are significant decreases, such as on August 5, 2009 where a large unexplained dip in question-asking tweets was found.

Lastly, similar to media stories, both viral tweet dissemination and specific twitter campaigns were found to have a considerable effect on tweet volume and posting behaviour in several cases (i.e., the #oink campaign showed immediate and potentially lasting impact on terminology use). Viral marketing techniques have been piloted and evaluated in other internet applications (e.g., Gosselin & Poitras, 2008), but have not been studied on Twitter. These techniques and methods used on Twitter may have potential applications in public health and should be considered as an area of future research.

5.4 Retweets

Retweets were unexpectedly very similar in content to non-retweets. We had expected to see an increase in retweeted marketing and spam volume; however this was not the case. Instead, our retweet analysis found that the only significant difference in posting was the amount of shared personal experiences. Original tweets contained significantly more tweets with personal experiences compared to retweets. This finding indicates that users are not likely to repost another user’s personal updates en mass, potentially because there is little interest or perceived
benefit in reposting second-hand or unrelated personal information. Instead, RTs appeared to be used mainly as a news re-broadcasting tool. As such, the spread of anecdotal H1N1 stories on Twitter may be low, which concurs with other studies which have shown that tweets must either have broad appeal or provide detailed information of local utility in order to be commonly retweeted (Starbird & Palen, 2010). Consequently, the tweet spikes that decreased when retweets were removed from the automated data provide a likely indicator of stories that had these qualities.

5.5 Information Demands & Sources

During the outbreak, a variety of traditional media sources speculated that misinformation was rampant in social media (Sutter, 2009). However, we classified only 4.5% of manually coded tweets as possible misinformation or speculation. Although this amount ranged from 2.2 to 9.2% across our 9 time points, increasing amounts of misinformation did not occur until August, months after initial media reports. It has been previously observed that social media and the internet have great potential to spread health myths and rumours, particularly in regards to vaccinations, although the overall percentage of related content on the internet is low (Zimmerman et al., 2005; Keelan et al., 2007; Keelan et al., 2010). While we did not observe large amounts of misinformation in our data, the effect of any amount of misinformation available to the public is unknown. In any case, tweeted misinformation and questions are potentially useful for public health agencies to address the information needs of the public. While not explored in this study, automatic and qualitative analysis of the most frequently tweeted misconceptions, questions, and sources of information on specific days could allow agencies to become quickly aware of public concerns, issues and myths and address them before they become problematic. This information could be potentially helpful for directing both online and offline health education initiatives and campaigns. Media monitoring has been used by the CDC to inform risk communication strategies in previous emergencies (Prue et al., 2003).

It is noteworthy that 90.2% of tweets provided references to the information that they were providing, allowing others to determine for themselves the trustworthiness of the material should they choose to follow the URL. While the majority of these tweets linked to mainstream or local
news websites, it is not clear if this is indicative of users being discerning about choosing credible information sources or is simply the result of information supply and demand (more information from major news providers when the popularity of the topic is high and vice versa). However, because the proportion of links to secondary news sites (news blogs/feeds/niches, social networks, and other web pages) increased over time, it would seem to indicate the latter. Although discouraging, the lack of critical assessment and evaluation of online health information by consumers is a well-documented problem (Eysenbach & Kohler, 2002). Similarly, it is also disconcerting that public health and government authorities such as the CDC, WHO, and PHAC were often times not referenced directly by users (1.5% of links). While mentionings of these governing bodies were higher in our data due to the proliferation of news headlines quoting or referring to them, direct linking to the authority and its resources was rare. However, this finding may be expected as the public tend to look for information synthesis, and this is a service that the media provides. An analysis of retweets also found this trend, although there was a significant upward trend in linking to authorities over time that was not found in original tweets. This unique trend may indicate that users began to recognize the utility of official resources over time and started to recommend them to others by retweeting them.

Studies on traditional information sources used by the public during emergencies have also found that government resources were less frequently used than mainstream media (Pollard, 2003). It may be the case that the public turns to the media in emergencies as typical emergency preparedness and disaster resources are challenging to read and not culturally appropriate (Friedman et al., 2008). However, mainstream media are tasked with providing a different focus than government agencies; they look to cover multiple newsworthy aspects of an event, not just public health, (Mebane et al., 2003) and typically do not provide enough information for consumers to act upon (Tanner et al., 2009). Our findings appear to conform to prior information studies and raise questions regarding the role that public health authorities have in social networking and the effectiveness of their current Twitter presence.
5.6 Automated Analysis

The majority of our automated queries performed well and correlated with the results of our manual coding. We see this as an encouraging first step to applying more intricate and technological approaches in the future. However, several of our qualifiers did not achieve positive results. Without the assistance of natural language processing or more advanced techniques, creating searches that could pick up specific sentiments or emotions at the exclusion of others was difficult as we were relying only on key search terms that were predominantly unique to one concept. Queries that did not perform well (relief, downplayed risk, and frustration) had less defined vocabularies than others and were more difficult to associate with particular expressions. In addition, our results show that spam and popular news articles that contain key phrases can influence search results and create peaks in activity that may not be reflective of the concept. Lastly, our queries are limited to keywords found in the manual coding and variants that the authors could anticipate and likely do not encompass the entire vocabulary used on Twitter. This may be indicated as the proportion of positive results returned by the automated analysis for certain concepts was low but high in the manual coding (e.g., personal experiences). These issues emphasize the importance of analysing the overall content of the tweet and the intricacies of building a substantial search vocabulary. While full semantic analysis is necessary in these cases, Infovigil currently does not have the capability. In the future, it may be possible to use natural language processing or other sophisticated semantic processing methods to achieve higher precision and accuracy.

5.7 Methodological Limitations & Advantages

Public attitudes, perceptions, and behaviours during the pandemic have also been reported by other studies using traditional survey methods (e.g., Lau et al, 2009a; Rubin et al., 2009; Jones & Salathe, 2009). Thus far we have not formally attempted to correlate Twitter data with these reports but intend on addressing this research question in the future. However, there may be some practical limitations to directly comparing our results and methodology to those of other authors’. The largest limitation to our approach in this respect is the lack of a well-defined study population. With the anonymity that the internet offers, it is difficult to determine the true
identity of any online persona. While our database allows us to link a user with any given tweet, it is beyond the scope of our study to retrieve every user profile in order to determine the demographics of our sample. Several marketing research companies have developed methodologies for measuring Twitter’s demographic user base. According to a study conducted by Cheng et al. (2009), the service is predominantly used by Americans and they account for 50.8% of all users. Pew Internet reported that approximately 19% of all online American adults are using Twitter or a similar application (Lenhart & Fox, 2009). And as of April 10, 2010, it is estimated that in the United States, 55% of users are female, 45% are between the ages of 18-34, 69% are Caucasian, 49% have less than a college degree, and 58% make over $60K a year (Quantcast, 2010). These demographics may give us some insight as to whom our population consists of; however, those who tweet about H1N1 may not necessarily be representative of the greater Twitter population at large. Findings from a recent study on general Twitter usage by the Harvard School of Business suggest that this is likely the case. Their analysis of 300,000 tweets found that a disproportionate number of tweets (90%) were produced by the top 10% of power users (Heil & Piskorski, 2009). In addition, because we are potentially sampling across the globe, it would be difficult to narrow the study context and compare the results with studies that report on a certain geographic region (e.g., Lau et al., 2009; Seale et al., 2009). This methodological issue is present even in traditional studies that attempt to corroborate their results with papers from different cities or countries (Balkhy et al., 2010). In the future it may be possible to take advantage of tweet geocoding to address this problem and sort tweets based on location. Secondly, certain questions posed to survey respondents may not be completely translatable to a query concept or category, even if numerous search patterns are used. For example, measuring vaccination intent and behaviour may be more easily defined in a query due to a more constrained and distinct vocabulary compared to measuring the public’s perceived efficacy of H1N1 prevention methods.

In regards to our sampling, no existing validated sampling method for Twitter has been documented in the literature and the decisions made in our study may not be optimal in all cases. It is unknown whether our manual sample was representative of all H1N1 tweets, although the results of our automated analysis do indicate there is a significant relationship between the two. However, proportions and trends of some content and qualifiers were different between
automated and manual results. This lack of correspondence may indicate that the manual sample was not adequate or that search queries were not optimized. We recognize that it is likely that not all relevant tweets are represented in our tweet database as some tweets may not have included our keywords (H1N1, swine flu, swineflu) and used their own terminology to refer to H1N1. In addition, some H1N1 tweets may be more conversational with back-and-forth discussion between users. While the context in these cases may be referring to H1N1 the tweets may not include any keywords.

As mentioned previously, automated analysis with Infovigil is feasible, but still under development. In its current form it does not have advanced language-processing techniques and as such, it only captures results based on groups of keywords or concepts. This method is very sensitive to widely circulated news stories using any of the keywords and is not particularly specific. It also allows for the double counting of tweets if multiple keywords are present in a tweet. While problematic, these issues can likely be solved in future iterations of the system.

Our validation process involved taking results from the manual and automated coding and using Pearson’s correlations to find the relationship between the two. However, the automated dataset included tweets from the manual coding dataset. Therefore, it is possible that the significant correlations achieved were because Infovigil was returning tweets coded in the manual analysis. Yet it is necessary for us to include these tweets in both the manual and automated datasets in order to determine if Infovigil was indeed detecting tweets that were known to contain our concepts. Our results indicated that Infovigil did indeed do this and also detected relevant tweets not in the manual coding dataset, as seen by the large amount of tweets returned in the searches.

The significant correlations that were achieved between manual and automated coding are likely influenced by the more robust vocabulary that certain concepts had. Concepts that did not correlate well were harder to define and consequently had fewer keywords. With a greater number of keywords there is a higher probability of correlation and as such, correlations for concepts could potentially change from project to project depending on the robustness of the vocabulary available in each situation.
We did not observe large amounts of misinformation in our data, but this may be a conservative estimate as we did not code humourous or confusing posts as misinformation. It is possible that misleading H1N1 parodies and jokes do contribute to public confusion and concern. And while our estimates were low, we do not know the effect of any amount of misinformation that exists on the internet, particularly when internet sources are archived and indexed in search engines. Similarly, this work is also limited by the assumption that all tweets are equal in terms of impact or impressions. In other words, we treated a tweet written by a user with 5 followers the same as a tweet written by actor Ashton Kutcher, who has over 5 million followers and no weighting was applied. As such, the tweets of high-profile users will have more “ripple effect” and broadcasting reach than those of low-key individuals. While this assumption may not have any effect on the proportions of tweet types that we found, it will likely have an effect on the spread or impact of tweets. However, message impact was not a goal of this research project, but may be a worthwhile phenomenon to study in the future.

Our analysis of retweets found that retweets linking to government or public health websites were significantly increasing over time. However, the interpretation of this trend should be taken with caution as we did not analyze tweet authorship and this trend may simply indicate that official Twitter feeds started to produce more tweets linking to their own websites and resources as the pandemic continued.

Despite these limitations, there are numerous advantages to using infoveillance. Because our method of data collection is continuous and ongoing, the length of our study time frame likely has no survey-based equivalent. Thus far, the existing H1N1 pandemic studies have collected data anywhere from a span of one day (Woien & Tonsberg, 2009) to four months (Sypsa et al., 2009). Those with shorter time frames have reported their results in aggregate, and only a handful has presented longitudinal results of selected questions (e.g., Sypsa et al., 2009; Van et al., 2010). While relevant to the study period, trends reported to be present (or absent) in these papers may be more complicated to interpret within the larger scope of the pandemic, especially if the time frame is short and if the context of external events is ignored. Even in our own results, popular news stories were shown to have a significant effect on public sentiment and attention. Although our manual coding was limited to 9 time points of analysis, Infovigil is
continuously collecting and analysing data, thereby creating a significant database that captures both large and small shifts in user posting and puts them into perspective within the overall pandemic picture.

Although this initial study focused on a biological threat, the same methodology could be applied to other contexts and emergencies. Our current vocabulary reflects the biological nature of H1N1, but a threat-specific vocabulary for different crises could be built using data from previous or emergent situations. Additionally, the archival of tweets allows for qualitative exploration and explanation of these changes. This methodology may offer complementary insight to traditional survey methods at a more rapid and less costly rate.

5.8 Practical Implications for Public Health

The findings of this study may have practical implications for public health agencies wishing to use automated infoveillance approaches. The time needed to produce the codebook and keywords for automated analysis of Twitter data for any given crisis may take approximately one week of preliminary data collection and analysis. From this data, common phrases and vocabulary can be transformed into search queries and allow Infovigil to analyze Twitter data automatically. From our experience, it is probable that new keywords will have to be added as the language of the emergency changes over time.

By monitoring Twitter in real-time, agencies can understanding and be aware of questions or misconceptions that the public as they emerge and allow officials to address specific concerns and tailor their communication and education strategies accordingly. For example, using Twitter, automated official responses with links to resources could be scripted to reply to questions that users ask on Twitter. If users enable geocoding on their account, more precise answers could be given, particularly if users are concerned about infection in their area or where to receive vaccinations or aid. In other cases, monitoring misinformation may lead to individual websites that broadcast these reports and may allow agencies to sanction or alert the public to be wary of these resources.

Viewing the most commonly tweeted resources may also help agencies to understand what information the public is most interested in and where their attention is focused. This may allow agencies to gain insight into what stories are being read (knowledge translation) and what media sites are the most commonly accessed for news on the emergency. Knowledge translation of specific terms and facts can
also be assessed by analyzing the frequency of use longitudinally, as we showed in this study using the terms “H1N1” and “swine flu”.

5.9 Research Objectives Revisited

This infodemiology project had 3 main research goals. Here we summarize our findings grouped accordingly.

1) Track the use of H1N1 versus swine flu terminology. A linear regression of the proportion of H1N1 tweets to swine flu tweets found that the use of H1N1 in tweets increased over time, suggesting that efforts to encourage the use H1N1 terminology were successful on Twitter.

2) Determine what Twitter users are communicating on Twitter, how they are expressing themselves, what information sources they are using, and if this content is changing over time. We found that the majority of tweets were resource-related. Tweets most often expressed humour, concern, and questions and linked to news websites. Posting activity of several content, qualifier and link categories changed over time, indicating that Twitter activity is not constant and likely fluctuates due to external events.

3) Explore the use of Twitter as a real-time content, sentiment, and public attention trend analysis and tracking tool. We developed automated search queries for several content and qualifier categories based on the manual codebook. We found that the results of the majority of search queries correlated well with the manual coding results using Pearson’s correlation. Qualitative analysis of tweet spikes revealed news events or topics that lead to the increased posting activity. This methodology may serve as a means of measuring public attention.

This study is the first known content analysis of tweets concerning a global public health emergency or the 2009 H1N1 pandemic. The methodology and results yield unique and important data regarding the use of Twitter during such crises by the public and may provide insight for social media use by authorities during future emergency situations and researchers wishing to study public health applications of social media. The novel infodemiology methodology described in this paper to code tweets automatically and in real time demonstrates
the feasibility of conducting content and sentiment analysis rapidly using data from Twitter. This time and resource-saving development in content analysis methodology allows for rapid investigation of large quantities of data and may compliment data collected by traditional survey methods.

5.10 Future Directions

The work detailed within this report is the first step in a larger infodemiology project that examines the burgeoning role of social media in public health emergencies. As such, there are many additional research questions to be addressed.

1) A detailed analysis of public information needs would be informative to health officials for future communication and education strategies. Content that was coded as “misinformation” or as “questions” can be qualitatively explored for popularly discussed myths or uncertainties. Emergent patterns can be quantified to determine the frequency of occurrences and how information needs changed over time.

2) The confusion and hesitation surrounding H1N1 vaccinations was detrimental to public vaccination campaigns. An examination of the vaccination discourse on Twitter may provide helpful insight as to what concerns, myths, and information sources may have hindered national campaigns and determine when and what events may have caused public sentiments or attitudes to change.

3) A qualitative and quantitative analysis of Twitter activity by government accounts may yield important recommendations for future official communication strategies. Messages can be qualitatively described for content and type of risk communication strategy. The effectiveness or popularity of these accounts can be measured by the number of times messages were retweeted (specific content popularity) and by the number of followers (general popularity).

4) As mentioned previously, we will attempt to correlate our results with previously existing studies on public attitudes, perceptions, and behaviours during the H1N1 pandemic. Search queries to match different types of survey questions will have to be created to capture the concept being measured, although this may not be possible for all types of
questions. This research question would represent a new methodological challenge for infoveillance, but an important one.

5) Further analysis of specific Twitter notations (e.g., RT, follow @username) may provide further insight as to how Twitter users use the tool. As RTs are an informal recommendation tool (Starbird & Palen, 2010), describing and quantifying the most retweeted content would allow us to see which news stories were the most influential or pervasive on Twitter. Longitudinal analysis of retweets would allow us to track these RTs over time to study their lifespan and reach of their readership. A study on follow@ tweets may provide details on the relationships and power dynamics between information transmitters and receivers in the information dissemination process.

6) As we are unsure of our population at the current time, devising a way to gather more demographic information on tweet authors (i.e., their location) would aid in interpretive validity. Similarly, it would be worthwhile to determine which accounts are tweeting and how many tweets each account is producing. Both pieces would help us to understand the generalizability of our results. In addition, these details may inform future Twitter sampling methodologies.

7) Infovigil and the automated analyses that were described in this study are still in their infancy. More advanced language and sentiment processing techniques will need to be applied for more accurate, specific, and finite results.

8) Formative research on the language used in different health emergencies can be undertaken in order to build threat-specific vocabulary sets for Infovigil to analyze tweets in future emergencies.

9) Message impact as a function of followers would help to understand what messages were the most widely read and could be an important consideration for future infodemiology studies using social media.

5.11 Conclusion

This study illustrates the potential and feasibility of using social media to conduct “infodemiology” studies for public health. 2009 H1N1 pandemic-related tweets on Twitter were primarily used to disseminate information from credible sources to the public, but were also a
rich source of opinions and experiences. These tweets can be used for near real-time content, sentiment, and public attention analysis, knowledge translation research, and potentially as a syndromic surveillance tool, allowing health authorities to become aware of and respond to real or perceived concerns (i.e., misinformation) raised by the public and complement traditional public polling studies. While the first step in this study included manual classifications and preliminary automated analyses, more advanced semantic processing tools may be used in the future to classify tweets with more precision and accuracy.
REFERENCES


34. Effler, P., Carcione, D., Ciele, C., Dowse, G., Goggin, L., & Mak, D. (2010). Household responses to pandemic (H1N1) 2009-related school closures, Perth, Western Australia. Emerging Infectious Diseases, 16(2), 205-211.


49. Frost, J., & Massagli, M. (2008). Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. *Journal of Medical Internet Research, 10*(3), e15.


42nd Hawaii International Conference on System Sciences, Waikoloa, Big Island, Hawaii, USA.


177. YouTube. (2010). At five years, two billion views per day and counting. *Official YouTube Blog*. Retrieved from [http://www.webcitation.org/5qaU0lWuQ](http://www.webcitation.org/5qaU0lWuQ)


Appendix 2-1: Protocol for Meta-review of H1N1 Public Perceptions, Attitudes, and Behaviour

**Research Question:** Meta-review of public perceptions, attitudes, and behaviour during the H1N1 pandemic

**Inclusion Criteria:**
- Studies of public attitudes, beliefs, behaviours, intents, perceptions, or opinions specific to the 2009/10 influenza A(H1N1) outbreak and ensuing vaccination strategy
- Study population: health consumers (general public)
- Written in English
- Published peer-reviewed studies
- Interview or survey methodology
- Cross-sectional designs
- Panel designs

**Exclusion Criteria:**
- Study population includes only healthcare workers, professionals, administrators and/or organizations, not the general public/consumer
- Rapid opinion polls
- Qualitative studies
- Written in languages other than English

**Databases:**
- PubMed

**Search Strategy:**
(h1n1 OR "swine flu" OR "swine-flu") AND (attitud* OR perception* OR behavio* OR intent* OR belief* OR opinion*) AND (survey* OR poll* OR interview* OR questionnaire*)
Appendix 2-2: Flowchart of Study Selection for H1N1 Meta-review

Citations retrieved with search terms \( (n=52) \)

- Studies excluded by title \( (n=25) \)

Abstracts retrieved \( (n=27) \)

- Studies excluded by abstract \( (n=5) \)

Potentially appropriate studies to be included in the meta-analysis \( (n=22) \)

- Studies excluded (inaccessible, foreign language) \( (n=2) \)

Studies included in meta-analysis \( (n=20) \)
## Appendix 2-3: Description of Studies Selected for H1N1 Meta-review

<table>
<thead>
<tr>
<th>First Author, Year</th>
<th>N</th>
<th>Sampling Method</th>
<th>Response Rate</th>
<th>Geographic Focus</th>
<th>Participant Demographics</th>
<th>Data Collection Period</th>
<th>Outcomes</th>
<th>Data Collection Method</th>
<th>Longitudinal or Aggregated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lau, 2009a</td>
<td>550</td>
<td>Randomly selected from telephone directory</td>
<td>61.9</td>
<td>Hong Kong</td>
<td>41.3</td>
<td>50-60 (29.5%)</td>
<td>Chinese Hong Kong</td>
<td>Telephone survey</td>
<td>Knowledge Risk perception Behaviour Emotional Response Att. to Authorities</td>
</tr>
<tr>
<td>Lau 2009b</td>
<td>301</td>
<td>Random selected from telephone directory</td>
<td>80%</td>
<td>Hong Kong</td>
<td>45</td>
<td>40-49 (29%)</td>
<td>Chinese Hong Kong</td>
<td>Telephone survey</td>
<td>Vaccination Knowledge Risk perception Att. to Authorities</td>
</tr>
<tr>
<td>Maurer, 2009</td>
<td>2067</td>
<td>National representative internet panel</td>
<td>54%</td>
<td>USA</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Internet Survey</td>
<td>Vaccination</td>
</tr>
<tr>
<td>Jones, 2009</td>
<td>6249</td>
<td>Convenience, Stanford subject pool</td>
<td>N/A</td>
<td>USA</td>
<td>53.4</td>
<td>37.6</td>
<td>N/A</td>
<td>Internet survey via subject pool &amp; social networking</td>
<td>Risk perception Emotional Response Information sources Behaviour</td>
</tr>
<tr>
<td>Quinn, 2009</td>
<td>1543</td>
<td>Online research panel - representative sample of US, oversample African Americans &amp; Hispanics</td>
<td>62%</td>
<td>USA</td>
<td>48.2</td>
<td>35-64 (57%)</td>
<td>White (88.8)</td>
<td>Internet survey via subject pool</td>
<td>Vaccination Risk Perception Att. to Authorities</td>
</tr>
</tbody>
</table>
### Appendix 2-3: Description of Studies Selected for H1N1 Meta-review (continued)

<table>
<thead>
<tr>
<th>First Author, Year</th>
<th>N</th>
<th>Sampling Method</th>
<th>Response Rate</th>
<th>Geographic Focus</th>
<th>Participant Demographics</th>
<th>Data Collection Method</th>
<th>Outcomes</th>
<th>Data Collection Period</th>
<th>Longitudinal or Aggregated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wong, 2010a</td>
<td>1050</td>
<td>Cross-sectional, random from telephone directory</td>
<td>60%</td>
<td>Kuala Lumpur, Malaysia</td>
<td>36, 18-39 (51.1%) Malay, 50.8, N/A</td>
<td>Telephone survey</td>
<td>Knowledge Risk Perception Behaviour Emotional Response Information Sources</td>
<td>July 11 – Sept 12, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Wong, 2010b</td>
<td>1050</td>
<td>Cross-sectional, random from telephone directory</td>
<td>60%</td>
<td>Kuala Lumpur, Malaysia</td>
<td>N/A, N/A, N/A, N/A, N/A, N/A</td>
<td>Telephone survey</td>
<td>Emotional Response Behaviour</td>
<td>July 11 – Sept 12, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Goodwin, 2009</td>
<td>328</td>
<td>Convenience</td>
<td>90% for paper survey</td>
<td>Europe, Malaysia</td>
<td>38, 31.2 Malaysian (54.8), UK (36) N/A</td>
<td>Paper (Malaysia), internet (Europe)</td>
<td>Behaviour Emotional Response Risk Perception Knowledge</td>
<td>April 30 – May 6, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Woen, 2009</td>
<td>506</td>
<td>Random from telephone directory, with representative weighting</td>
<td>37%</td>
<td>Norway</td>
<td>N/A, N/A, N/A, N/A, N/A, N/A</td>
<td>Telephone survey</td>
<td>Risk Perception Information Sources Emotional Responses Att. to Authorities</td>
<td>April 29, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Rubin, 2009</td>
<td>997</td>
<td>Random digit dialing, proportional quota sampling</td>
<td>6.97%</td>
<td>United Kingdom</td>
<td>47, 35-54 (35.2%) White (92.6) 37.6 59.4 (includes all work)</td>
<td>Telephone survey</td>
<td>Behaviour Emotional Response Risk Perception Att. to Authorities Information Sources</td>
<td>May 8-12, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Schwarzinger, 2010</td>
<td>2253</td>
<td>Stratified, Random selection from online representative research panel, oversampled low response populations</td>
<td>12%</td>
<td>France</td>
<td>50.9, 35-54 (46.2%) French, 10.6 70</td>
<td>Online survey sent via e-mail</td>
<td>Intent to Vaccinate Risk Perception</td>
<td>Nov 17-25, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Sypsa, 2009</td>
<td>1000</td>
<td>Cross-sectional, proportional quota sampling</td>
<td>N/A</td>
<td>Greece</td>
<td>34.2, 51.9 Greek, 31.3, N/A</td>
<td>Telephone survey</td>
<td>Vaccination Risk perception</td>
<td>Aug 23 – Oct 31, 2009</td>
<td>Longitudinal &amp; aggregate</td>
</tr>
<tr>
<td>Goodwin, 2010</td>
<td>186</td>
<td>Online</td>
<td>N/A</td>
<td>Europe</td>
<td>32, 33.77 Portuguese (43) 63, N/A</td>
<td>Internet survey</td>
<td>Behaviour Emotional Response</td>
<td>April 29 – June 11, 2009</td>
<td>Aggregate</td>
</tr>
<tr>
<td>First Author, Year</td>
<td>N</td>
<td>Sampling Method</td>
<td>Response Rate</td>
<td>Geographic Focus</td>
<td>% Male</td>
<td>Mean Age or Mode</td>
<td>% Ethnic Majority</td>
<td>% College degree or higher</td>
<td>% Full time Employed</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>--------</td>
<td>------------------</td>
<td>-------------------</td>
<td>---------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Balkhy, 2010</td>
<td>1548</td>
<td>Proportional quota sampling from mall patrons</td>
<td>97%</td>
<td>Saudi Arabia (Riyadh &amp; Jeddah)</td>
<td>53.5</td>
<td>25-39 (53.2%)</td>
<td>Saudi</td>
<td>61.9</td>
<td>62 (includes all work)</td>
</tr>
<tr>
<td>Kamate, 2010</td>
<td>791</td>
<td>Stated to be random, but unclear</td>
<td>95%</td>
<td>Udaipur, India</td>
<td>57</td>
<td>30-39 (32.9)</td>
<td>Indian</td>
<td>36.4</td>
<td>26.2 (includes all work)</td>
</tr>
<tr>
<td>Seale, 2009</td>
<td>620</td>
<td>Convenience, unknown for e-mail survey</td>
<td>85.4% paper 61.4% e-mail</td>
<td>Sydney</td>
<td>42.9</td>
<td>25-34 (33.7%)</td>
<td>White (75.3)</td>
<td>64</td>
<td>53.4</td>
</tr>
<tr>
<td>Effler, 2010</td>
<td>233</td>
<td>Questionnaire sent to all parents of 3 schools affected by closure</td>
<td>58%</td>
<td>Perth, Australia</td>
<td>N/A</td>
<td>11 median (range 5-13)</td>
<td>Australian</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Eastwood, 2010</td>
<td>830</td>
<td>Random telephone sample, representative quota</td>
<td>72%</td>
<td>Australia</td>
<td>38</td>
<td>41-60 (43%)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Van, 2010</td>
<td>2882</td>
<td>UNSW staff, students</td>
<td>5.7%</td>
<td>University of South Wales, Sydney, Australia</td>
<td>45.3</td>
<td>18-24 (51.6%)</td>
<td>Australian (51.8)</td>
<td>N/A</td>
<td>33.6</td>
</tr>
<tr>
<td>Seale, 2010</td>
<td>627</td>
<td>Convenience mall sampling, every 5th person for 2 hours</td>
<td>47%</td>
<td>Sydney, Australia</td>
<td>40.7</td>
<td>25-34 (26%)</td>
<td>White (67)</td>
<td>73.2</td>
<td>79.3</td>
</tr>
</tbody>
</table>
Appendix 4.1. SQL Queries for Automated Tweet Coding & Analysis

Content Concepts

<table>
<thead>
<tr>
<th>Resources</th>
<th>Subconcept: “Personal/direct H1N1 experience”</th>
</tr>
</thead>
</table>
| SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.website_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(T.title) LIKE "%swine flu%" OR LOWER(T.title) LIKE "%swineflu%" OR LOWER(T.title) LIKE "%h1n1%")) AND ((LOWER(T.title) NOT LIKE "%rt @%" AND LOWER(T.title) NOT LIKE "%rt@%")) AND ((LOWER(T.title) NOT LIKE "%http://%" AND LOWER(T.title) NOT LIKE "%https://%")) GROUP by title ORDER by publishdate ASC

<table>
<thead>
<tr>
<th>Personal Experiences</th>
<th>Subconcept: “Indirect Experience”</th>
</tr>
</thead>
</table>
| SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.website_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(T.title) LIKE "%swine flu%" OR LOWER(T.title) LIKE "%swineflu%" OR LOWER(T.title) LIKE "%h1n1%")) AND ((LOWER(T.title) NOT LIKE "%rt @%" AND LOWER(T.title) NOT LIKE "%rt@%")) AND ((LOWER(T.title) NOT LIKE "%http://%" AND LOWER(T.title) NOT LIKE "%https://%")) GROUP by title ORDER by publishdate ASC

-OR-

Subconcept: “Personal/direct H1N1 experience”

Subconcept: “Indirect Experience”

SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.website_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(T.title) LIKE "%swine flu%" OR LOWER(T.title) LIKE "%swineflu%" OR LOWER(T.title) LIKE "%h1n1%"))
AND ((LOWER('title') NOT LIKE "%rt @%" AND LOWER('title') NOT LIKE "%rt@%") AND (LOWER(title) LIKE "%my mom%") OR (LOWER(title) LIKE "%my mum%") OR (LOWER(title) LIKE "%my mother%") OR (LOWER(title) LIKE "%my mom%") OR (LOWER(title) LIKE "%my dad%") OR (LOWER(title) LIKE "%my father%") OR (LOWER(title) LIKE "%my bro%") OR (LOWER(title) LIKE "%my sis%") OR (LOWER(title) LIKE "%my uncle%") OR (LOWER(title) LIKE "%my aunt%") OR (LOWER(title) LIKE "%my grandm%") OR (LOWER(title) LIKE "%my grandpa%") OR (LOWER(title) LIKE "%my grandfather%") OR (LOWER(title) LIKE "%my cousin%") OR (LOWER(title) LIKE "%my niece%") OR (LOWER(title) LIKE "%my nephew%") OR (LOWER(title) LIKE "%my friend%") OR (LOWER(title) LIKE "%my classmate%") OR (LOWER(title) LIKE "%my roommate%") OR (LOWER(title) LIKE "%my neighbour%") OR (LOWER(title) LIKE "%my my boyfriend%") OR (LOWER(title) LIKE "%my girlfriend%") OR (LOWER(title) LIKE "%my bf%") OR (LOWER(title) LIKE "%my gf%") OR (LOWER(title) LIKE "%my wife%") OR (LOWER(title) LIKE "%my husband%") OR (LOWER(title) LIKE "%my kid%") OR (LOWER(title) LIKE "%my son%") OR (LOWER(title) LIKE "%my daughter%") OR (LOWER(title) LIKE "%my baby%") OR (LOWER(title) LIKE "%my doctor%") OR (LOWER(title) LIKE "%my co-worker%") OR (LOWER(title) LIKE "%my coworker%") OR (LOWER(title) LIKE "%my class%") OR (LOWER(title) LIKE "%my school%") OR (LOWER(title) LIKE "%my university%") OR (LOWER(title) LIKE "%my church%") OR (LOWER(title) LIKE "%my city%") OR (LOWER(title) LIKE "%my town%") OR (LOWER(title) LIKE "%my dorm%") OR (LOWER(title) LIKE "%my rez%") OR (LOWER(title) LIKE "%my campus%") OR (LOWER(title) LIKE "%my home%") OR (LOWER(title) LIKE "%my house%") OR (LOWER(title) LIKE "%my office%") OR (LOWER(title) LIKE "%my country%")) GROUP by title ORDER BY publishdate ASC

-OR-

Subconcept: “Vaccination Experience”

SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url, COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER('title') LIKE "%swine flu%" OR LOWER('title') LIKE "%swineflu%" OR LOWER('title') LIKE "%h1n1%") AND ((LOWER('title') NOT LIKE "%rt @%" AND LOWER('title') NOT LIKE "%rt@%")) AND ((LOWER('title') NOT LIKE "%got my shot%") OR (LOWER('title') LIKE "%got my h1n1 shot%") OR (LOWER('title') LIKE "%got my swine flu shot%") OR (LOWER('title') LIKE "%got vaccinated %") OR (LOWER('title') LIKE "%i'm vaccinated%") OR (LOWER('title') LIKE "%got the swine flu shot%") OR (LOWER('title') LIKE "%got the h1n1 shot%") OR (LOWER('title') LIKE "%got the swine flu jab%") OR (LOWER('title') LIKE "%got my h1n1 jab%") OR (LOWER('title') LIKE "%got my swine flu jab%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got the swine flu vaccine%") OR (LOWER('title') LIKE "%got my swine flu vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOWER('title') LIKE "%got my h1n1 vaccine%") OR (LOW...
immuni%") OR (LOWER(title) LIKE "%got my swine flu immuni%") OR (LOWER(title) LIKE "%went to get the vaccine%") OR (LOWER(title) LIKE "%went to get vaccinated%") OR (LOWER(title) LIKE "%went to get the swine flu shot%") OR (LOWER(title) LIKE "%went to get the swine flu vaccine%") OR (LOWER(title) LIKE "%went to get my swine flu jab%") OR (LOWER(title) LIKE "%went to get my swine flu vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 shot%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%went to get my swine flu shot%") OR (LOWER(title) LIKE "%went to get my swine flu vaccine%") OR (LOWER(title) LIKE "%went to get my swine flu jab%") OR (LOWER(title) LIKE "%went to get my swine flu vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 shot%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%went to get my swine flu shot%") OR (LOWER(title) LIKE "%went to get my swine flu vaccine%") OR (LOWER(title) LIKE "%went to get my swine flu jab%") OR (LOWER(title) LIKE "%went to get my swine flu vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 shot%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%went to get the h1n1 vaccine%") OR (LOWER(title) LIKE "%went to get the h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%") OR (LOWER(title) LIKE "%is immuni%") OR (LOWER(title) LIKE "%went to get my h1n1 shot%") OR (LOWER(title) LIKE "%went to get my h1n1 jab%")

SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url, COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(`title`) LIKE "%swine flu%" OR LOWER(`title`) LIKE "%swineflu%" OR LOWER(`title`) LIKE "%h1n1%") AND ((LOWER('title') NOT LIKE "%rt @%" AND LOWER('title') NOT LIKE "%rt@%")) AND ((LOWER(title) LIKE "%in my opinion%") OR (LOWER(title) LIKE "%i think%") OR (LOWER(title) LIKE "%imho%") OR (LOWER(title) LIKE "%government should%") OR (LOWER(title) LIKE "%gov't should%") OR (LOWER(title) LIKE "%doctors should%") OR (LOWER(title) LIKE "%Obama should%") OR (LOWER(title) LIKE "%media should%") OR (LOWER(title) LIKE "%cdc should%") OR (LOWER(title) LIKE "%i believe%") OR (LOWER(title) LIKE "%schools should%") OR (LOWER(title) LIKE "%teachers should%") OR (LOWER(title) LIKE "%hospitals should%") OR (LOWER(title) LIKE "%people should%") OR (LOWER(title) LIKE "%my stance%") OR (LOWER(title) LIKE "%my take%") OR (LOWER(title) LIKE "%my view%") OR (LOWER(title) LIKE "%my feeling%") OR (LOWER(title) LIKE "%my impression%") OR (LOWER(title) LIKE "%my theory%") OR (LOWER(title) LIKE "%my thought%") OR (LOWER(title) LIKE "% pov %") OR (LOWER(title) LIKE "%my opinion%") OR (LOWER(title) LIKE "%i recommend%") OR (LOWER(title) LIKE "%i suggest%") OR (LOWER(title) LIKE "%my suggestion%") OR (LOWER(title) LIKE "%did you hear%") OR (LOWER(title) LIKE "%reading%" AND NOT LOWER(title) LIKE "%spreading%") OR (LOWER(title) LIKE "%interest%") OR (LOWER(title) LIKE "%looking up%" AND NOT LOWER(title) LIKE "%things are looking up%" AND NOT LOWER(title) LIKE "%it\'s looking up%") OR (LOWER(title) LIKE "%researching%") OR (LOWER(title) LIKE "%heard %") OR (LOWER(title) LIKE "%read %" AND NOT LOWER(title) LIKE "%spread %") OR (LOWER(title) LIKE "%I read %") OR (LOWER(title) LIKE "%i hear %") OR (LOWER(title) LIKE "%i feel that%") GROUP by title ORDER BY publishdate ASC
### Qualifiers Concepts

<table>
<thead>
<tr>
<th>Humour/Sarcasm</th>
<th>Relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>`SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN &quot;2009-05-01 00:00:00&quot; AND &quot;2009-12-31 23:59:59&quot; AND ((LOWER(title) LIKE &quot;%swine flu%&quot; OR LOWER(title) LIKE &quot;%swineflu%&quot; OR LOWER(title) LIKE &quot;%h1n1%&quot;)) AND ((LOWER(title) NOT LIKE &quot;%rt @%&quot; AND LOWER(title) NOT LIKE &quot;%rt@%&quot;)) AND ((LOWER(title) LIKE &quot;%lol%&quot;) OR (LOWER(title) LIKE &quot;%lMAO%&quot;) OR (LOWER(title) LIKE &quot;%haha%&quot;) OR (LOWER(title) LIKE &quot;%hehe%&quot;) OR (LOWER(title) LIKE &quot;%hilarious%&quot;) OR (LOWER(title) LIKE &quot;%funny%&quot;) OR (LOWER(title) LIKE &quot;%ode to tamiflu%&quot;) OR (LOWER(title) LIKE &quot;%j/k%&quot;) OR (LOWER(title) LIKE &quot;% XD %&quot;) OR (LOWER(title) LIKE &quot;%:P%&quot;) OR (LOWER(title) LIKE &quot;%=P%&quot;) OR (LOWER(title) LIKE &quot;%ROFL%&quot;) OR (LOWER(title) LIKE &quot;%:c%&quot;) OR (LOWER(title) LIKE &quot;%=%&quot;) OR (LOWER(title) LIKE &quot;% jk %&quot;) OR (LOWER(title) LIKE &quot;% XP %&quot;) OR (LOWER(title) LIKE &quot;%=D%&quot;) OR (LOWER(title) LIKE &quot;%:D%&quot;) OR (LOWER(title) LIKE &quot;% ha ha %&quot;) OR (LOWER(title) LIKE &quot;%just kiddin%&quot;) OR (LOWER(title) LIKE &quot;%piggy%&quot;) OR (LOWER(title) LIKE &quot;%oink%&quot;) OR (LOWER(title) LIKE &quot;%bacon flu%&quot;) OR (LOWER(title) LIKE &quot;%when pigs fly%&quot;) OR (LOWER(title) LIKE &quot;%he he %&quot;) OR (LOWER(title) LIKE &quot;%heh%&quot;) OR (LOWER(title) LIKE &quot;%unlikelysequels%&quot;) OR (LOWER(title) LIKE &quot;%bacon fever%&quot;) OR (LOWER(title) LIKE &quot;%joke%&quot; AND NOT LOWER(title) LIKE &quot;%what a joke%&quot;) OR (LOWER(title) LIKE &quot;%hiney%&quot;) OR (LOWER(title) LIKE &quot;%heimie%&quot;)) GROUP by title ORDER BY publishdate ASC</td>
<td></td>
</tr>
<tr>
<td>`SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN &quot;2009-05-01 00:00:00&quot; AND &quot;2009-12-31 23:59:59&quot; AND ((LOWER(title) LIKE &quot;%swine flu%&quot; OR LOWER(title) LIKE &quot;%swineflu%&quot; OR LOWER(title) LIKE &quot;%h1n1%&quot;)) AND ((LOWER(title) NOT LIKE &quot;%RT@%&quot; AND &quot;%RT @%&quot;)) AND ((LOWER(title) LIKE &quot;%relieved%&quot;) OR (LOWER(title) LIKE &quot;%thank God%&quot;) OR (LOWER(title) LIKE &quot;%thankful%&quot;) OR (LOWER(title) LIKE &quot;%whew%&quot;) OR (LOWER(title) LIKE &quot;%i am ok%&quot;) OR (LOWER(title) LIKE &quot;%im ok%&quot;) OR (LOWER(title) LIKE &quot;%all better%&quot;) OR (LOWER(title) LIKE &quot;%feeling % better%&quot;) AND NOT LOWER(title) LIKE &quot;%not feeling % better%&quot;) OR (LOWER(title) LIKE &quot;%i’m recovering%&quot;) OR (LOWER(title) LIKE &quot;%i just recovered%&quot;) OR (LOWER(title) LIKE &quot;%i’ve recovered%&quot;) OR (LOWER(title) LIKE &quot;%back at school%&quot;) OR (LOWER(title) LIKE &quot;%back at work%&quot;) OR (LOWER(title) LIKE &quot;%glad that %&quot;) OR (LOWER(title) LIKE &quot;%happy to hear%&quot;) OR (LOWER(title) LIKE &quot;%glad to hear%&quot;) OR (LOWER(title) LIKE &quot;%i’m safe%&quot;) OR (LOWER(title) LIKE &quot;%grateful%&quot; AND NOT LOWER(title) LIKE &quot;%ungrateful%&quot;) OR (LOWER(title) LIKE &quot;%good to hear%&quot;) OR (LOWER(title) LIKE &quot;%happy that%&quot;) OR (LOWER(title) LIKE &quot;%good that%&quot;) OR (LOWER(title) LIKE &quot;%survived%&quot;) OR (LOWER(title) LIKE &quot;%bounced back%&quot;) OR (LOWER(title) LIKE &quot;%recuperate%&quot;) OR</td>
<td></td>
</tr>
<tr>
<td>Subconcept</td>
<td>SQL Query</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Concern for Others</td>
<td>SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN &quot;2009-05-01 00:00:00&quot; AND &quot;2009-12-31 23:59:59&quot; AND ((LOWER(title) LIKE &quot;%swine flu%&quot; OR LOWER(title) LIKE &quot;%swineflu%&quot;) OR LOWER(title) LIKE &quot;%h1n1%&quot;)) AND ((LOWER(title) NOT LIKE &quot;%RT@%&quot; AND &quot;%RT @%&quot;)) AND (LOWER(title) LIKE &quot;%get better%&quot;) OR (LOWER(title) LIKE &quot;%get well%&quot;) OR (LOWER(title) LIKE &quot;%are you ok%&quot;) OR (LOWER(title) LIKE &quot;%is it swine flu%&quot;) OR (LOWER(title) LIKE &quot;%is it h1n1%&quot;) OR (LOWER(title) LIKE &quot;%is it the swine flu%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%hope % is ok%&quot;) OR (LOWER(title) LIKE &quot;%hope you don't get %&quot;) OR (LOWER(title) LIKE &quot;%hope you're not sick%&quot;) OR (LOWER(title) LIKE &quot;%hope you aren't sick%&quot;) OR (LOWER(title) LIKE &quot;%hope u r ok%&quot;) OR (LOWER(title) LIKE &quot;%hope u don't get %&quot;) OR (LOWER(title) LIKE &quot;%hope u aren't infected%&quot;) OR (LOWER(title) LIKE &quot;%hope you aren't %&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;)</td>
</tr>
<tr>
<td>Concerned Emoticons</td>
<td>SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN &quot;2009-05-01 00:00:00&quot; AND &quot;2009-12-31 23:59:59&quot; AND ((LOWER(title) LIKE &quot;%swine flu%&quot; OR LOWER(title) LIKE &quot;%swineflu%&quot;) OR LOWER(title) LIKE &quot;%h1n1%&quot;)) AND ((LOWER(title) NOT LIKE &quot;%RT@%&quot; AND &quot;%RT @%&quot;)) AND (LOWER(title) LIKE &quot;%:-(%&quot;) OR (LOWER(title) LIKE &quot;%=-(&quot;)) OR (LOWER(title) LIKE &quot;%:%(%&quot;) OR (LOWER(title) LIKE &quot;%:%(%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;) OR (LOWER(title) LIKE &quot;%:=%&quot;)</td>
</tr>
<tr>
<td>General Concern</td>
<td>SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN &quot;2009-05-01 00:00:00&quot; AND &quot;2009-12-31 23:59:59&quot; AND ((LOWER(title) LIKE &quot;%swine flu%&quot; OR LOWER(title) LIKE &quot;%swineflu%&quot;) OR LOWER(title) LIKE &quot;%h1n1%&quot;)) AND ((LOWER(title) NOT LIKE &quot;%RT@%&quot; AND &quot;%RT @%&quot;)) AND (LOWER(title) LIKE &quot;%get better%&quot;) OR (LOWER(title) LIKE &quot;%get well%&quot;) OR (LOWER(title) LIKE &quot;%are you ok%&quot;) OR (LOWER(title) LIKE &quot;%is it swine flu%&quot;) OR (LOWER(title) LIKE &quot;%is it h1n1%&quot;) OR (LOWER(title) LIKE &quot;%is it the swine flu%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;) OR (LOWER(title) LIKE &quot;%r u ok%&quot;)</td>
</tr>
</tbody>
</table>
W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(title) LIKE "%swine flu%" OR LOWER(title) LIKE "%swineflu%" OR LOWER(title) LIKE "%h1n1%") AND ((LOWER(title) NOT LIKE "%RT%@%" AND "%RT @%")) AND ((LOWER(title) LIKE "%omg%") OR (LOWER(title) LIKE "%oh my god%") OR (LOWER(title) LIKE "%uh oh%") OR (LOWER(title) LIKE "%oh no%") OR (LOWER(title) LIKE "%worried%" AND NOT LOWER(title) LIKE "%not worried%" AND NOT LOWER(title) LIKE "%unworried%") OR (LOWER(title) LIKE "%scared%" AND NOT LOWER(title) LIKE "%not scared%" AND NOT LOWER(title) LIKE "%don\'t bescared%" AND NOT LOWER(title) LIKE "%dont be scared%") OR (LOWER(title) LIKE "%stay away%") OR (LOWER(title) LIKE "%dangerous%" AND NOT LOWER(title) LIKE "%not dangerous%" AND NOT LOWER(title) LIKE "%no more dangerous%" AND NOT LOWER(title) LIKE "%less dangerous%") OR (LOWER(title) LIKE "%terrified%") OR (LOWER(title) LIKE "%afraid%" AND NOT LOWER(title) LIKE "%not afraid%" AND NOT LOWER(title) LIKE "%don't be afraid%" AND NOT LOWER(title) LIKE "%stop being afraid%" AND NOT LOWER(title) LIKE "%unafraid%" AND NOT LOWER(title) LIKE "%dont be afraid%") OR (LOWER(title) LIKE "%frightened%" AND NOT LOWER(title) LIKE "%not frightened%" AND NOT LOWER(title) LIKE "%stop freaking out%" AND NOT LOWER(title) LIKE "%not freaking out%" AND NOT LOWER(title) LIKE "%afraid%" AND NOT LOWER(title) LIKE "%not scared%" AND NOT LOWER(title) LIKE "%confused%" OR (LOWER(title) LIKE "%confusing%") OR (LOWER(title) LIKE "%yikes%") OR (LOWER(title) LIKE "%uneasy%") OR (LOWER(title) LIKE "%I worry%") OR (LOWER(title) LIKE "%worrisome%" AND NOT LOWER(title) LIKE "%not worrisome%" OR LOWER(title) LIKE "%dread%") OR (LOWER(title) LIKE "%concern%" AND NOT LOWER(title) LIKE "%unconcern%" AND NOT LOWER(title) LIKE "%not concern%" AND NOT LOWER(title) LIKE "%don't be concern%" AND NOT LOWER(title) LIKE "%dont be concern%" AND NOT LOWER(title) LIKE "%im concern%" OR LOWER(title) LIKE "%i am concern%" OR LOWER(title) LIKE "%i\m concern%" OR LOWER(title) LIKE "%kind of concern%" OR LOWER(title) LIKE "%kinda concern%" OR LOWER(title) LIKE "%sort of concern%" OR LOWER(title) LIKE "%sorta concern%" OR LOWER(title) LIKE "%really concern%") GROUP by title ORDER BY publishdate ASC

Subconcept: “Concern for Self”
SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updateated, T.authorname, T.authoruri, W.website_id, W.long_url,COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid =
| Downplayed Risk | SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.Updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url, COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND (LOWER(T.title) LIKE "%swine flu%" OR LOWER(T.title) LIKE "%swineflu%" OR LOWER(T.title) LIKE "%h1n1%") AND (LOWER(T.title) NOT LIKE "%RT@%" AND "%RT @%") AND ((LOWER(T.title) LIKE "%I might have h1n1%") OR (LOWER(T.title) LIKE "%I might have swine flu%") OR (LOWER(T.title) LIKE "%I think I have the swine flu") OR (LOWER(T.title) LIKE "%I have a fever") OR (LOWER(T.title) LIKE "%I have symptoms") OR (LOWER(T.title) LIKE "%I feel sick") OR (LOWER(T.title) LIKE "%I'm sick") OR (LOWER(T.title) LIKE "%I don't want to die") OR (LOWER(T.title) LIKE "%I don't have symptoms") OR (LOWER(T.title) LIKE "%I'm not feeling well") OR (LOWER(T.title) LIKE "%feeling sick") OR (LOWER(T.title) LIKE "%I don't care") GROUP by T.title ORDER BY publishdate ASC |
OR (LOWER(title) LIKE "%not worried%") OR (LOWER(title) LIKE "%dont care%") OR (LOWER(title) LIKE "%who cares%") OR (LOWER(title) LIKE "%forget swine flu%" AND NOT LOWER(title) LIKE "%dont forget swine flu%" AND NOT LOWER(title) LIKE "%d0t forget swine flu%" AND NOT LOWER(title) LIKE "%do not forget swine flu%" AND NOT LOWER(title) LIKE "%won't forget swine flu%") OR (LOWER(title) LIKE "%forget h1n1%" AND NOT LOWER(title) LIKE "%don't forget h1n1%" AND NOT LOWER(title) LIKE "%d0nt forget h1n1%" AND NOT LOWER(title) LIKE "%do not forget h1n1%" OR (LOWER(title) LIKE "%swine flu is nothing%") OR (LOWER(title) LIKE "%like normal flu%") OR (LOWER(title) LIKE "%like regular flu%") OR (LOWER(title) LIKE "%like seasonal flu%") GROUP by title ORDER BY publishdate ASC

Frustration
SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url, COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-12-31 23:59:59" AND ((LOWER(`title`) LIKE "%swine flu%" OR LOWER(`title`) LIKE "%swineflu%" OR LOWER(`title`) LIKE "%h1n1%") AND ((LOWER(`title`) NOT LIKE "%rt @%" AND LOWER(`title`) NOT LIKE "%@") AND (LOWER(title) LIKE "%sir %") OR (LOWER(title) LIKE "%grt%") OR (LOWER(title) LIKE "%wtf%") OR (LOWER(title) LIKE "%annoy%") OR (LOWER(title) LIKE "%irritat%") OR (LOWER(title) LIKE "%pathetic%") OR (LOWER(title) LIKE "%pissed%") OR (LOWER(title) LIKE "%freakin swine flu%") OR (LOWER(title) LIKE "%FML%") OR (LOWER(title) LIKE "%fuck%") OR (LOWER(title) LIKE "%shit%") OR (LOWER(title) LIKE  "%:(@%") OR (LOWER(title) LIKE "%:;%") OR (LOWER(title) LIKE="%:;(")) OR (LOWER(title) LIKE "%shut(up%") OR (LOWER(title) LIKE "%friggin%") OR (LOWER(title) LIKE "%getting out of hand%") OR (LOWER(title) LIKE "%stupid%") OR (LOWER(title) LIKE "%hate swine flu%") OR (LOWER(title) LIKE "%hate h1n1%") OR (LOWER(title) LIKE "%hate being sick%") OR (LOWER(title) LIKE "%hate it%") OR (LOWER(title) LIKE "%h1n1 sucks%") OR (LOWER(title) LIKE "%damn%") OR (LOWER(title) LIKE "%effing%") OR (LOWER(title) LIKE "%freaking swine flu%") OR (LOWER(title) LIKE "%frick%") OR (LOWER(title) LIKE "%freaking h1n1%") OR (LOWER(title) LIKE "%freakin h1n1%") OR (LOWER(title) LIKE "%i can\'t believe%") OR (LOWER(title) LIKE "%they better not%") OR (LOWER(title) LIKE "%i am mad %" OR LOWER(title) LIKE "%i'm mad %" OR LOWER(title) LIKE "%Im mad %" OR LOWER(title) LIKE "%is mad %" OR LOWER(title) LIKE "%so mad %") OR (LOWER(title) LIKE "%frustrated%") OR (LOWER(title) LIKE "%angry%") OR (LOWER(title) LIKE "%outrage%") OR (LOWER(title) LIKE "%cranky%") OR (LOWER(title) LIKE "%peved%") OR (LOWER(title) LIKE "%furious%") OR (LOWER(title) LIKE "%bitter%") OR (LOWER(title) LIKE "% irk%") OR (LOWER(title) LIKE "%crushed%") OR (LOWER(title) LIKE "%so sick of%") OR (LOWER(title) LIKE "%if i hear%"))) GROUP by title ORDER BY publishdate ASC

Misinformation
SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content, T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url, COUNT(T.title) AS number_hits FROM tweets T LEFT JOIN webcite W ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-
12-31 23:59:59"AND ((LOWER('title') LIKE "%swine flu%" OR
LOWER('title') LIKE "%swineflu%" OR LOWER('title') LIKE "%h1n1%"))
AND ((LOWER('title') NOT LIKE "%rt @%" AND LOWER('title') NOT LIKE
"%rt@%") AND ((LOWER(title) LIKE "%hidden%") OR (LOWER(title) LIKE
"%secret %") OR (LOWER(title) LIKE "%conspiracy%" AND NOT
LOWER(title) LIKE "%not a conspiracy%") OR (LOWER(title) LIKE
"%conspirator%") OR (LOWER(title) LIKE "%autism%") OR (LOWER(title)
LIKE "%h0ax%") OR (LOWER(title) LIKE "%apocalypse%") OR
(LOWER(title) LIKE "%armageddon%") OR (LOWER(title) LIKE "%real
story%") OR (LOWER(title) LIKE "%poison%") OR (LOWER(title) LIKE
"%guillain-barre%") OR (LOWER(title) LIKE "%eugenics%") OR
(LOWER(title) LIKE "%plot%") OR (LOWER(title) LIKE "%mind control%")
OR (LOWER(title) LIKE "%weapon%") OR (LOWER(title) LIKE
"%bioterrorism%") OR (LOWER(title) LIKE "%chemical warfare%") OR
(LOWER(title) LIKE "%toxin%") OR (LOWER(title) LIKE "%toxic%") OR
(LOWER(title) LIKE "% lie %") OR (LOWER(title) LIKE "% lies %") OR
(LOWER(title) LIKE "%mindcontrol%") OR (LOWER(title) LIKE "%mind-
control%")) GROUP by title ORDER BY publishdate ASC

Question
SELECT DISTINCT T.title, T.tweetid, T.id, T.publishdate, T.link, T.content,
T.updatedate, T.authorname, T.authoruri, W.webcite_id, W.long_url,COUNT(
T.title ) AS number_hits FROM tweets T LEFT JOIN webcite W
ON T.tweetid = W.tweetid WHERE publishdate BETWEEN "2009-05-01 00:00:00" AND "2009-
12-31 23:59:59" AND ((LOWER('title') LIKE "%swine flu%" OR
LOWER('title') LIKE "%swineflu%" OR LOWER('title') LIKE "%h1n1%"))
AND ((LOWER('title') NOT LIKE "%rt @%" AND LOWER('title') NOT LIKE
"%rt@%") OR (LOWER(title) LIKE "%?%") GROUP by title ORDER BY
publishdate ASC