AN ONTOLOGY-BASED FRAMEWORK FOR FORMULATING SPATIO-TEMPORAL INFLUENZA (FLU) OUTBREAKS FROM TWITTER

Maddumage Udaya Kumara Jayawardhana

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

August 2016

Committee:

Peter Gorsevski, Advisor

Jeffrey Snyder

Sheila J. Roberts

ABSTRACT

Peter Gorsevski, Advisor

Early detection and locating of influenza outbreaks is one of the key priorities on a national level for preparedness and planning. This study presents the design and implementation of a web-based prototype software framework (Fluwitter) for pseudo real-time detection of influenza outbreaks from Twitter in space and time. Harnessing social media to track real-time influenza outbreaks can provide different perspectives in battling the spread of infectious diseases and lowering the cost of existing assessment methods. Specifically, Fluwitter follows a three-tier architecture system with a thin web client and a resourceful server environment. The server side system is composed of a PostGIS spatial database, a GeoServer instance, a web application for visualizing influenza maps and daemon applications for tweet streaming, pre-processing of data, semantic information extraction based on DBpediaSpotlight and WS4J, and geo-processing. The collected geo-tagged tweets are processed by semantic NLP techniques for detecting and extracting influenza related tweets. The synsets from the extracted influenza related tweets are tagged and ontology based semantic similarity scores produced by WUP and RES algorithms were derived for subsequent information extraction. To ensure better detection, the information extraction was calibrated by different rules produced by the semantic similarity scores. The optimized rule produced a final *F*-measure value of 0.72 and accuracy (ACC) value of 94.4%. The Twitter generated influenza cases were validated by weekly influenza related hospitalization records issued by ODH. The validation that was based on Pearson's correlations suggested existence of moderate correlations for the Southeast region ($r = 0.52$), the

Northwestern region ($r = 0.38$), and the Central region ($r = 0.33$). Although, additional work is needed, the potential strengths and benefits of the prototype are shown through a case study in Ohio that enables spatio-temporal assessment and visualization of influenza spread across the state.

This is dedicated to Dr. Kala Melchior for the true inspiration and vision I received from you.

ACKNOWLEDGMENTS

I gratefully acknowledge my advisor Dr. Peter Gorsevski who gave me the greatest support, encouragement, mentorship, and motivation throughout the research as well as last two years of my stay at BGSU. Your work and vision inspired me to pursue this master's degree in Applied Geospatial Science and it also opened several new opportunities for me to work in the field of geo-informatics. The Committee members, Dr. Jeff Snyder, and Dr. Sheila J. Roberts are also greatly acknowledged for their support. I also convey my gratitude to the faculty and staff of the School of Earth, Environment and Society, BGSU, especially Bill and Pat.

I would like to pass my special thanks to the former doctoral student Hideki Shima in Carnegie Mellon University for developing WS4J tool which I greatly used in my work. Further, I also like to thank the open source developers of the tools I used including DBpediaSpotlight, WordNet, GeoTools, and Geobricks. I appreciate and thank Dr. Ruwan Weerasinghe (University of Colombo School of Computing, Sri Lanka) for his excellent theoretical and practical teaching in NLP which I certainly benefited in working with two master's research. Further, I am grateful to my former colleagues, Chamil Madusanka and Prabath Abesekara for the invaluable discussions we had regarding ontologies, semantic web, and rules based NLP methods in developing the CargoSpotter™ NLP based software solution and other projects at attune Consulting™, Sri Lanka.

Finally, I would like to thank my loving wife, parents and sister for always being the strength in my life and tolerating my unavailability in many occasions throughout these two years.

TABLE OF CONTENTS

INTRODUCTION

Seasonal influenza is a viral disease which causes severe health issues and mortality in high risk population groups and spreads from person to person (Fauci 2006, Wikramaratna and Gupta 2009). The seasonal influenza circulates globally and affects all ages of people causing different symptoms such as high fever, cough, headache, pain, sore throat, and runny nose. Seasonal influenza differs from influenza pandemic which is caused by the emergence of new non-existent viruses as people have neither natural resistance nor there are readily-available vaccines (Smith *et al.* 2009). Major pandemics such as the "Spanish flu" (N1 subtype) have killed over 50 million people world-wide in 1918-1919 while a different subtype strain of the same influenza (N2) caused a total of 69,800 and 33,800 deaths in 1957-1958 and 1968-1969 respectively in United States alone. Other example of a recent influenza threat such as the swine flu (H1N1) caused a world-wide pandemic in 2009 and currently is a human seasonal flu that also circulates in pigs. The avian influenza (H5N1 and H7N9) which is another highly pathogenic disease caused by domestic poultry has been reported since 2003. The highly pathogenic avian influenza has caused number of concerns because of high mortality among confirmed human cases and spread from birds to other mammals globally (Meltzer *et al.* 1999, Yoldascan *et al.* 2010, About the Flu | Flu.gov 2015). However not all influenza subtypes will mutate into highly pathogenic forms that will cause severe illnesses or deaths (Wikramaratna and Gupta 2009).

Significant outbreaks of influenza such as at the "World Youth Day 2008 Mass Gathering" in Sydney, Australia (Blyth *et al.* 2010) are infrequent while seasonal small scale outbreaks are the most common form (Gault *et al.* 2009). Apart from the health risks and other unpredictable effects on communities, influenza epidemics can have a great indirect impact on

the annual economy through absenteeism caused by closure of work places, businesses, schools and other infrastructure (Meltzer *et al.* 1999, Keogh-Brown *et al.* 2010, Yoldascan *et al.* 2010). Some risk estimates and predictions suggest that a potential severe influenza epidemic in the United States alone will cause 89,000 to 207,000 deaths and economic impact between $71.3 to $166.5 billion (Meltzer *et al.* 1999) while estimates for United Kingdom suggest a loss of 6% annual GDP (Keogh-Brown, Smith, *et al.* 2010). Because of the unpredictable nature associated with potential influenza epidemics or outbreak and limited response timeframe before a major event, an early detection of influenza and potential spread is one of the key priorities on a national level (Aramaki *et al.* 2011) .

Influenza preparedness planning varies by individual countries and most of the countries maintain their own guidelines for assessing and tracking outbreaks and interventions through their national institutes and programs such as the Center for Disease Control and Prevention (CDC) in the United States. Such institutes maintain different programs including the influenza-like Illness Surveillance Net-work (ILINet) (ILINet, 2015) in United States, the European Influenza Surveillance Scheme (EISS), and the Japanese Infection Disease Surveillance Center (IDSC) which employ traditional virology and clinical data (Griffin *et al.* 2009).

Some of the shortcomings are that current assessment practices are costly and inefficient for prompted reporting of the spread and consequent management and containment (Culotta 2010). For instance, current assessment practices that are used for early influenza detection include telephone triage service data (Espino *et al.* 2003) and telephone or internet based voluntary influenza reporting data from such health institutions/programs and hospitals (Rutter *et al.* 2014). Alternatively, over-the-counter pharmaceutical sales have been used for making early warnings of disease outbreaks (Magruder 2003), but certainly it will be ineffective in countries

where anti-influenza drugs are not issued over the counter. Other assessments are based on school absenteeism to detect possible school-based outbreaks (Mann *et al.* 2011) as well as other evidence-based techniques that are able to detect school closures during influenza outbreaks (Sasaki *et al.* 2009). However, the drawback with the current assessment approaches is that they have a lag time in data collection and delayed processing time which complicates preparedness and real time response.

In recent years, online social networking has revolutionized interpersonal communication and becomes ubiquitous and important tool that initiates new assessment strategies (Crooks *et al.* 2013). Among different social media outlets, information generated by microblogging platforms like Twitter (https://twitter.com) becomes more important for understanding dynamic trends that can support real-time assessment and consequent decision making because they are concise and tend to be updated more frequently. Microblogging is a form of blogging generated by crowdsourced information that has a small content in terms of actual and aggregated file size. For instance, Twitter limits its content to 140 characters, while the daily volume of tweets exceeds more than 500 million by an approximately more than 284 million active users (Oussalah et al. 2013; Cheng, Caverlee, and Lee 2010). The geographically referenced or geo-tagged nature of the tweets provides a real-time and cost-free data stream for space-time analytics. Such framework has influenced the development of many different applications in areas such as disaster management (Lucas 2012, Crooks *et al.* 2013), emergency response (Gelernter and Mushegian 2011), regional event detection (Lee *et al.* 2011), road hazard detection (Kumar *et al.* 2014), disease spread detection (Chen *et al.* 2010, Culotta 2010, Aramaki *et al.* 2011, Signorini *et al.* 2011, Kostkova *et al.* 2014), predicting future events (Bermingham and Smeaton 2011, Rao and Srivastava 2012, Vu *et al.* 2012, Kostkova *et al.* 2014, Ceron *et al.*

2015) and analyzing the effects of a past event (Chew and Eysenbach 2010, Vieweg *et al.* 2010, Miyabe *et al.* 2012).

The use of Tweeter to track influenza and other epidemic research has also been implemented through Twitter Application Programming Interfaces (APIs) (Aramaki *et al.* 2011, Lee *et al.* 2013). Although the streaming APIs provide real-time tweets, most of the early work used locally archived tweets that were coupled with post-processing (i.e., non-real-time) approaches. Techniques such as archived data based classifiers (implemented using machine learning) (Aramaki *et al.* 2011, Lamb *et al.* 2013, Bodnar *et al.* 2014, Kostkova *et al.* 2014), construction of regression and other predictive statistical models (Chen *et al.* 2010, Arias *et al.* 2013), and development of information ranking algorithms (Stewart and Diaz 2012) are just a few of the approaches used for processing tweets. Applications that validate occurrences and distribution of seasonal influenza form tweets using conventional sources such as government health agencies have been also attempted (Lampos and Cristianini 2010, Lamb *et al.* 2013). Although there is some success with those non real-time approaches, the main drawback is their inability to provide status of influenza in real-time.

New application developments that are focused on different methodological ideas for real-time influenza tracking have been also reported by multiple researchers (Achrekar *et al.* 2011, Aramaki *et al.* 2011, Lee *et al.* 2013). For instance, Lee et al. (2013) used fixed tag word frequency analysis in flu tweets (i.e., keyword "flu") for tracking cases of influenza in real-time. Achrekar et al (2011) used text mining and autoregression with exogenous inputs (ARX) model where past time-series ILINet data from CDC represent the autoregression portion of the model while tweets serve as exogenous inputs (i.e., external component). The intention of the model is to relate tweets to the time-series where one would like to explain the extent of ILI cases

reported with high accuracy. A support vector machine based classifier was employed by Aramaki *et al*. 2011 for real-time extraction of influenza tweets. A support vector machine (SVM) is a learning machine approach that is used for two-group classification tasks where input vectors are non-linearly mapped to a very high dimension feature space (Cortes and Vapnik 1995). Aramaki *et al*. 2011 used SVM classification for extracting only tweets mentioning actual influenza patients from all retrieved tweets.

The most critical part of the twitter based research is the information extraction (IE). Tweets contain semi-structured/unstructured, non-standard, and ill-formed text such as user hash tags (Oussalah *et al.* 2013), hyperlinks (Vakali *et al.* 2012, Kostkova *et al.* 2014), image links, geo-location tags (Lee *et al.* 2011, Tao *et al.* 2012, Guo and Chen 2014) and symbolic emotions. While user created tags and symbols do not always provide dependable information, links provide a little or no information (Chang *et al.* 2013). Additionally, user profile information, location of origin, number of times a tweet has been re-posted (retweet) and posted time might also be available as metadata along with each tweet depending on the user account settings (Chew and Eysenbach 2010, Walther and Kaisser 2013).

Different approaches have been proposed for IE from unstructured tweets. Some of the approaches depend on hash tags and URLs based IE, while other approaches use entire tweet content (Chew and Eysenbach 2010, Walther and Kaisser 2013, Kostkova *et al.* 2014). In the current literature, the two major approaches used for IE are rule based (hand crafted rules) and statistical approaches (Hua *et al.* 2012). The automated handcrafted rule based implementations have faster executions, but they are difficult to construct with all the possible rules for a real world scenario (Hua *et al.* 2012). Instead, approaches such as machine learning are capable of generating rules and/or models on their own, and can be implemented to different sets of data

using supervised or unsupervised techniques. Hence, such approaches become more practical and have been widely adapted for influenza research using tweets (Aramaki *et al.* 2011, Signorini *et al.* 2011, Lamb *et al.* 2013).

The shortcomings with typical rule based or statistical techniques are that they are purely driven on keyword searches and lack semantic capabilities. The semantics in computer sciences is a growing field that focuses on capturing relationships and meaning between signifiers, like words, phrases, signs, and symbols, and contextual meaning as inherent in the larger text blocks or narratives (Pustejovsky and Boguraev 1996). In social media content, the intent of semantic IE is to enable computers to understand the meanings of human expressions and concepts that are specified by content (Grassi *et al.* 2011, Bontcheva and Rout 2014). In order to improve the computer-understanding of those relationships, often ontologies are used which represent conceptualization of specific domains of interest for organizing the concepts. Ontology is a semantic structural framework for understanding human expressions through description, classification, and reasoning of spatial (and non-spatial) data. Ontology represents knowledge thought a set of concepts which is used to describe relations that exist within the structural framework of the domain. They are comprised of classes and properties where classes represent a concept or a physical entity in the domain of interest while the properties link the relations between multiple classes. Ontologies used for IE are knowledge based, often developed to model relationships through descriptions consisting of classes and properties (Guarino 1998).

Such ontologies are used in IE attempts to increase understanding of natural language text in the tweets (Hedden 2008). DBpedia (Lehmann *et al.* n.d.) is one of the best open-domain ontologies, which uses knowledge (articles) contained in the open encyclopedia "wikipedia.org" (Nebhi 2012). DBpedia Spotlight tool (Mendes *et al.* 2011) which is an extended product of

DBpedia is currently available and it can be used for extracting the meaning and relationships of words/terms in tweets. Thus, there is a need for additional research that improves current real-time approaches by focusing on the true meaning (semantics) of tweet contents. Further, application of spatio-temporal analysis techniques for flu tweets will bring additional information and enhanced decision support capabilities. This study proposes to develop a real-time approach that integrates Twitter Streaming APIs for real time data retrieval with the purpose of formulating spatio-temporal influenza outbreaks using tweets.

The proposed objectives for this study are 1) to harness real-time information from georeferenced Twitter messages that relate to the spread of influenza; 2) to develop a rule-based information extraction system that links semantic descriptions to ontological properties of the influenza related tweets; 3) to calibrate the rule-based information extraction and validate the influenza tweets using data from ILINet; and 4) to build a web-based prototype for visualization and analytics of tweets. The first objective intends to acquire and prepossess real-time Twitter messages that have a geolocation for the state of Ohio. The goal of the second objective is to implement information extraction from the Tweets by tagging and queries of custom built dictionary. The next objective will calibrate the rule-based information extraction by exploring different thresholds applied to a training dataset. In addition, this objective will attempt to validate scenario from real-time tweets against standard ILINet influenza data. The aim of the final objective is to develop an interactive web application for visualization of spatio-temporal distribution of influenza tweets.

CHAPTER I. DATASET AND METHODS

1.1. Dataset

Study area is restricted by geo-tagged tweets which have geographic coordinates and originate within the state of Ohio. The area is defined by the following east-west extent between -80.52°W and -84.81°W longitude and a north-south extent between 41.99°N and 38.40°N. The geo-tagged tweets are most likely generated by mobile devices such as smart phones and tablets, but personal computers are also used to generate considerable amount of tweets. In the state of Ohio, nearly 100,000 geo-tagged tweets are generated each day, but this number can be significantly increased during holidays or a popular gathering events. The Pew Research Center study from 2013 (Social Media Update 2013 | Pew Research Center 2016), suggests that 50% of USA Twitter users are between 18 and 50 of age. Considering the similar age groups in Ohio, 30.4% of the population is between 18 - 40 age and 46% of the population is between 18 - 50 age. Furthermore, the Northeastern Ohio shows the highest regional mean of the median age of its counties which is 37.74, while the Southeast Ohio has the lowest regional mean of the median age of its counties of 35.35 (Ohio QuickFacts from the US Census Bureau 2016) .

Tweet messages are streamed nearly real time from Twitter data center servers using Twitter Streaming APIs. The geographic region of interest for streaming tweets is provided as a filter with the values of the bounding box coordinates. The individual messages streamed by this API are JSON (JavaScript Object Notation) encoded which is an open-standard format that uses attribute/value pairs represented by readable text. A tweet object is composed of varying number of attributes depending on the content such as location of origin, source device, a unique tweet id, message text, message source (i.e. web, i-phone, android and etc.), longitude, latitude, place

polygon, and created time (Tweets | Twitter Developers 2015). A set of attributes that are used in this study are shown in Table 1. For example, the "text" attribute in the table contains the twitter message which is the content information used to identify a topic of interest such as flu or influenza. The contextual information such as spatial information (originating location) is stored in "longitude", "latitude" fields, while the "place_polygon" attribute stores the geo-coordinates of the polygon representing the originating location which is normally a bounding box or boundary of a municipality or other administrative district. Depending on the user's privacy and device settings, exact originating location could be missing from the "longitude" and the "latitude" fields. In such cases, the information comes in "place_polygon" attribute which is used to capture the spatial information. Temporal information is extracted from the "created_at" timestamp attribute.

1.2. System Architecture

The proposed client-server system has been designed following the three tier architecture including: data tier, a logic tier (application tier) and a presentation tier (Fig 1). The data tier stores both unprocessed and processed tweets and they are accessed by the logic tier components. The data tier combines PostgreSQL database server with the PostGIS spatial extension that is used to accommodate efficient implementation of spatial data queries that follow the simple features for SQL specification from the Open Geospatial Consortium (OGC). The PostgreSQL open source software component is an object-relational database management system (ORDBMS) with an emphasis on extensibility for allowing users to define internal functions in many programming languages (i.e., Java, Python, PHP, C++). The database server can handle different workloads which support small single-machine applications or large distributed and concurrent applications.

The logic tier contains the Twitter streaming client, pre-processor, semantic tagger, semantic similarity calculator and geographic data processor components. The fundamental tasks performed by the logic tier include real-time collection of tweets using Streaming API, storing tweets in a spatial database, preprocessing and normalizing, semantic tagging, calculating semantic similarity between words, and other analytics which support visualization and mapping. The logic tier makes concurrent uses of the PostgreSQL/PostGIS database in the data tier which is the centric place for storing unprocessed twitter messages and retrieving twitter messages for further processing. The individual components in the logic tier have no direct link between them but they are connected through the database module. The components represent independent daemon applications developed in Java and they are hosted as system services in a Linux (Ubuntu) server environment. However, to meet the real time processing requirements some of the compute-intensive components such as the pre-processor, the semantic tagger and the semantic similarity calculator services were designed to perform parallel processing.

The data-intensive tweet streaming client has been designed to efficiently read and store tweets real-time with no lags, while the rest of the compute-intensive components have been designed to efficiently perform the pseudo-real time Twitter data processing in the backend. The Twitter streaming client maintains a streaming connection based on Hypertext Transfer Protocol (HTTP) to keep the connection with Twitter data servers. Both the semantic tagger that connects with the DBpedia Spotlight web service and the geographic data processor that connects with the GeoServer web services, use HTTP based Representational State Transfer (REST) style communication.

Finally, the presentation tier consists of the custom built web application and the GeoServer geo-spatial data sharing server. GeoServer stores the maps generated by the

geographic data processor component and delivers them to the web application on request. The web application is developed in Java, HTML and JavaScript and it renders maps received from the GeoServer on an interactive web browser interface.

1.3. System Components and Major Tasks

1.3.1. Tweet Streaming Client

The Twitter Streaming APIs collect real-time tweets from public streams. The collection process from the public streaming endpoint uses a module called Twitter4J which is an open-source Java library for the Twitter API, released under the Apache License 2.0 (Yamamoto 2010). To make the collection process more robust, the implementation constraints the language of the content to English and the geographic area to the bounding box of Ohio. Then those constraints are sent as the streaming filter criteria called a Twitter "POST statuses / filter" along with the Twitter streaming request. The Twitter Streaming API returns the matching tweets in either XML (EXtensible Markup Language) or JSON format which is used in the prototype due to its simplicity and compactness. In addition, to avoid disruption in the collection of the streaming data, the individual JSON tweets are saved first to the database before separate background process is used to parse the tweets into components which are organized and converted to Java objects. The text message content of individual tweets are stored in a PostgreSQL spatial database along with their originating geo-location and timestamp.

1.3.2. Pre-Processor

The main goal of this module is to pre-process the raw data which contains a substantial amount of noise. The noise contains repetitive and decorative character sequences, slang words, URLs, and emoticons, misspelled words, and abbreviated phrases which require further preprocessing of the tweets. Figure 2 shows the preprocessing steps which include tokenizing,

standardizing, word noise removing, and word spell correction. At the pre-processing stage, the original tweets are retrieved from the database and then they are passed for text processing using the chained processing sequences described above. The noise free pre-processed tweets were stored back in the database using a separate table.

The removal of emotions and unnecessary punctuations is the first step that filters tokens such as decorative and/or repetitive character sequences or other emoticons characters using the ArkTweet NLP (Owoputi *et al.* 2012) library. The ArkTweet NLP methodology implements an unsupervised hierarchical clustering technique and supports a broad range of Unicode. In addition, the lists of tokens extracted by ArkTweet NLP were further standardized by using a custom built list of frequently found anomalous words/terms such as abbreviations. The word look-up list included an anomalous word list from the GATE Twitter part-of-speech tagger (Derczynski *et al.* 2013) and a selection of frequent anomalous words that were found from collected tweets. Additional pre-processing included, corrections of misspelled words by using predefined rules and patterns from a standard dictionary, removal of high-frequency words (i.e. "the", "is", "at", "which", "on" and etc.) using a “stop word dictionary” , and removal of URLs or other user profiles tagged with "@" symbol. On the other hand "#" symbols in hashtags were striped from meaningful words and kept for further processing.

1.3.3. Semantic Tagger

The semantic tagging process intends to extract the meaning of a subject contained in a tweet. Semantic tagging implies meaning of a human concept that a computer can be programmed to understand. The ontologies are the building blocks that define the relationships and respective meanings between the words. For instance, the sentence "I have got the flu, it's the last day of school" could be tagged with "flu" and "school" tags referring them to the matching

ontology classes of "http://dbpedia.org/page/Influenza" and "http://dbpedia.org /page/School" respectively. Semantic taggers are available as programmatically accessible software tools which tag words using a given ontology. For instance, the DBpedia Spotlight semantic tagger which maps words to DBpedia ontology implements natural language text tagging and generates annotations in formats such as HTML, RDFa, XML, and JSON (Fig. 3).

The DBpedia Spotlight semantic tagger is used for tagging of the individual tweets and ontological URIs references are extracted and stored in a database as tags. The statistical implementation of the DBpediaSpotlight was hosted as a standalone web service that is implemented on a local server and accessed by a web service client request. Such a web service request consists of a set of parameters including values for support, confidence, annotation policy and entity types (optional) that are accompanied by the tweet text. For instance, the confidence and the support parameters govern the criteria for tagging and for specific ontology classes, where different confidence and support values can results in different number of tagged outcomes. In this application the values for these parameters were kept constant using confidence and support of 0.2 and 20 respectively. The values were determined through a calibration process that used a subset of true positives (i.e., correctly identified tweets) and false positives (i.e., incorrectly identified tweets). Each request to the web service generates a set of semantic tags and a set of attributes for each tag including support and similarity score (degree of similarity between the tagging term and the related ontology class).

A wide variety of semantic tags with different subject areas are generated for each individual tweet. The processing of semantic tags in this work is focused only on the semantic tags which relate to potential influenza subject areas. To accomplish this, a list composed of influenza related ontological entity types such as biology, health, medicine, and their sub types

so-called “white-list policy” was used to accelerate the processing and enhance the extraction of influenza related tweets. Since the speed of received tweets at the streaming client exceeds the speed of semantic tagging processing, a parallel-processing component was implemented to increase the performance. Distributed and parallel-processing of data was achieved by implementing virtual partitioning on the pre-processed tweet records in the database based on the primary key value. However, additional customization of the implemented design allows for further improvements such as distributed processing over multiple servers.

1.3.4. Semantic Similarity Calculator

Semantic similarity or semantic relatedness represents a measure defined over a pair of terms that reflect relationship or the likeness of their meaning. For instance, when a tweet contains an influenza case, semantics of words/terms in the tweet reflect the similarity between words/terms with "influenza" as a measure of strength. The semantic similarity in this implementation was calculated by two different similarity scores including scores from semantic tags and from the words in the pre-processed tweets. Specifically, the semantic similarity can be calculated by different techniques such as WUP (Wu and Palmer 1994), RES (Resnik 1995) and JCN (Jiang and Conrath 1997). WUP computes semantic similarity between terms based on the number of nodes between the hierarchies of terms which represent their lowest common subsume and the root. Both RES and JCN determine the similarity of two terms based on their distances to the closest common ancestor term and/or the annotation statistics of their common ancestor terms. The WordNet that was used here is a semantic network database for English language which was developed by University Princeton. WordNet similarity for Java (WS4J) (Shima 2013) is a tool that implements WUP and RES semantic similarity algorithms in Java (Miller 1995). The implementation of the algorithms is based on the path length and information

content methods. While the path length method calculates number of nodes or relation between nodes in taxonomy, the information content method is based on frequency counts of concepts as found in a corpus of text. In this work, WUP and RES algorithms in WS4J were used to measure the semantic similarity for each tweet.

The WUP algorithm estimates relatedness based on depths of two synsets (i.e., a set of synonym of a concept) in the WordNet database and in the direction of depth of the *lcs* (least common subsume). For instance, synsets can be related based on semantic relation that holds between two words that can (in a given context) express opposite meaning (antonymy), semantic relation of being superordinate or belonging to a higher rank or class (hypernymy), or semantic relation that holds between a part and the whole (meronymy). Equation 1 takes two concepts c_1 and c_2 (two ontology elements) and returns the semantic similarity between them which is a score between 0 and 1. A score of 0 indicates no relationship between the synsets and 1 indicates that the synsets are identical. In case of an error a score of -1 is returned by the algorithm. The *lcs* in the equation represents the deepest "shared parent" of two nodes where the depth is defined as the separation from the root concept in terms of amount of nodes. Thus, the deeper the *lcs* is, the more similar the concepts are. For example, the *lcs* of "cough" and "fever" is "symptom" as the closest relationship between them is being symptoms of a disease.

$$sim_{WP}(c_1, c_2) = \frac{2 * depth(lcs(c_1,\ c_2))}{len(c_1, lcs(c_1, c_2)) + \ len(c_2, lcs(c_1, c_2)) + \ 2 * depth(lcs(c_1,\ c_2))} \qquad \text{Eq.(1)}$$

The Resnik's measure (RES) of semantic similarity between synsets is based on the information content (IC) that uses the term probability. The RES algorithm takes two concepts c_1 and c_2 (two words or terms) and returns the semantic similarity which is a score between 0 and

positive infinity, or returns an error score of -1 in case of an error. The Resnik's measure for comparing the synsets is defined as follows:

$$sim_{RS}(c_1, c_2) = IC(lcs(c_1,\ c_2)) \qquad \text{Eq.(2)}$$

$$IC(c) = -ln(p(c))$$

where sim(c_1, c_2) is the set of common ancestors of terms c_1 and c_2 in the ontology. One of the drawbacks with the Resnik measure is the coarseness because many different pairs of concepts may share the same *lcs*.

The semantic similarity scores used in the prototype included sets that were generated from semantic tags and sets that were generated from synsets (words/terms). The extracted tags and synsets from each tweet were evaluated for semantic similarity with “influenza” by WUP and RES separately. In addition to WUP and RES scores, overall scores of semantic similarities were calculated by summing the scores generated by WUP and RES for each tweet. The similarity scores from both algorithms produced new database columns including "Tag_WUP", "Tag_RES”, "Word_WUP" and "Word_RES” which were stored for each individual tweet. The usefulness of the overall similarity scores is that signify the degree of relation between tweets and “influenza” as a function of multiple tags or synsets. However, it should be clear that the effectiveness of those semantic similarity scores depends upon the strength of the semantic tagger and initial specification of confidence and support parameters.

Because ontologies are conceptualization of a domain of interest that is associated with representation of knowledge, the influenza related concepts are represented by a concise amount of classes (i.e., tags or synsets) or a finite set of elements in the WordNet environment. The property that links two or more classes in the domain of interest (i.e., cough and fever associated

with influenza) signifies that those classes are members of the same set. Members of the same set have similar structural hierarchy which allows for development of custom rules that can place focus on specified content of interest. For that reason, in the proposed prototype similarity scores were pre-calculated for the most relevant set of influenza related concepts. The pre-calculated similarity scores were generated by WS4J which were later used as a quick reference (i.e., lookup file). The reason for pre-calculated similarity scores was to reduce the computational requirements that are required for real time calculation of dynamic scores. However, the approach described here is completely applicable to other domains but additional similarity scores need to be recalculated or introduced dynamically.

The additional efforts that were considered in the prototype were focused on the detection of similar relationships used from different domains. For instance, the information extraction process was further improved to capture precise semantics of synonyms such as fever, flu, and chills that are also used in domains such as sport and entertainment. This task was performed by random selection of 2000 tweets that were queried for containment of influenza related words. Each tweet was flagged as Boolean true or false and consequently labeled with the mostly appropriate domain representation. The next step implemented a word frequency analysis to depict domain labels and to elicit the occurrence of non-representative domains used in the influenza context. Identified domains were mapped and their respective ontological classes were listed for the exclusion of such relationships (i.e., exclusion semantics list or rules). The exclusion list was then used to calculate the semantic similarity scores for tags and synsets that were associated with non-influenza related domains. The overall scores produced by WUP and RES algorithms produced two new measurements called “UnlikeFlu_WUP" and "UnlikeFlu_RES" that were added into the database. All semantic similarity scores

("Tag_WUP", "Tag_RES", "Word_WUP", "Word_RES", " UnlikeFlu_WUP" and "UnlikeFlu_RES") were used in the calibration process for formulating rule(s) for identifying influenza related tweets.

1.3.5. Calibration

The calibration used a randomly selected subset of 1400 tweets, which were determined by influenza related keywords and had pre-calculated semantic similarity scores. The subset of tweets were individually interpreted for influenza related content and the final result yielded a total of 107 influenza related tweets (i.e. positives) and 1293 non-influenza related tweets (i.e. negatives). The influenza related tweets were manually processed for inclusion of symptoms such as fever, chills, sore throat and sickness induced incidents (i.e. school or workplace absence). However, it should be clear that ambiguity associated with this complex subject area presents a significant challenge and limitation when it comes to optimized calibration for semantic language processing (i.e. accounting for all possible word meanings) and automated computational knowledge acquisition. Therefore, the presented calibration may not be optimal but represents an attempt for finding a better understanding of the influenza related meaning in tweets. Manually processed tweets were flagged using Boolean true or false notation and stored in a new table field called "HasFlu".

To improve the prediction of influenza, the randomly selected subset was used to generate rules that optimize the detection based on the similarity scores. The following four similarity scores including "Tag_WUP", "Tag_RES", "Word_WUP" and "Word_RES" were tested separately for a threshold value that optimizes the number of correctly identified influenza tweets (i.e. true positives) and correctly identified non-influenza tweets (i.e. true negatives). Fig. 4 shows a confusion matrix also called contingency table that is used for a binary classification.

Across the top are the prediction outcomes (i.e. predicted) and down the side are the actual values (i.e. observed). The confusion matrix displays the number of correct and incorrect predictions made by a model compared with the actual classifications in the test data. In this work, the quality of the classification from the threshold value was evaluated from a confusion matrix by different measurements shown in Equations 3 to 6. The recall (Eq. 3) represents a measure of the ability of the system to present all relevant items, while precision (Eq. 4) represents a measure of the ability of the system to present only relevant items. The *F*-measure that combines precision and recall is the harmonic mean and represents a weighted average of the precision and recall, where a score of 1 is the best score and 0 is the worst score. The accuracy (Eq. 6) is a statistical measure of how well a binary classification test performed in terms proportion of true results (both true positives and true negatives) among the total number of items. The accuracy is expressed as a percentage and 100% indicates the best accuracy level.

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq.(3)}$$

where TP is the number of true positives and FN is the number of false negatives. The number of relevant items retrieved is the TP while the number of relevant items in the collection is the TP + FN.

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq.(4)}$$

where TP is the number of true positives and FP is the number of false positives. The number of relevant items retrieved is the TP while the total number of items retrieved is the TP + FP.

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \quad \text{Eq.(5)}$$

$$Accuracy\ (ACC) = \frac{(TP + TN)}{P + N} * 100 \quad \text{Eq.(6)}$$

where TP is the number of true positives, TN is the number of true negatives, P is the number of predicted positives and N is the number of predicted negatives.

Fig. 5 shows the calibration best outcome curve which is associated with the "Word_WUP" semantic similarity scores. The *x*-axis in the figure represents the threshold values that were used for derivation of confusion matrices. For instance, the lowest overall score used as a threshold value on the *x*-axis is 0.05 while the highest score is 1.7. The *y*-axis represents the *F*-measure values that were derived from the confusion matrices generated by different threshold values. The figure shows that a threshold value of 0.93 generates an *F*-measure value of 0.65. The accuracy (ACC) for the "Word_WUP" threshold value (0.93) produced accuracy of 90.1%.

Additional enhancements for improving the *F*-measurement value and the accuracy (ACC) included a development of a rule that excluded similar influenza relationships from other domains. This was accomplished by a logical condition that tested the relationship, if influenza related semantic similarity scores ("Word_RES") are greater than the exclusion semantics similarity scores ("UnlikeFlu_RES"). The final rule that was implemented in this research was formulated as "**Word_WUP >= 0.93 AND Word_RES > UnlikeFlu_RES**". The implementation of this rule produced a final *F*-measure value of 0.72 and accuracy (ACC) value of 94.4% where a total of 99 out of 107 influenza tweets were correctly identified in the calibration dataset.

1.3.6. Geographic Data Processor

Initial geographical data contained in tweets are points and polygons, but polygon data cannot be directly used to make an interpolated raster. Therefore, all polygon data are converted to their respective centroids points. Then, all overlapping data points are represented by one point and the number of total overlapped points is represented as the value for that point. Using those extracted data, GeoTools (About GeoTools — GeoTools 2015) library is employed to generate a shapefile for each day within the temporary file storage. The gdal_grid functionality of GDAL (Geographic Data Abstraction Library) (Open Source Geospatial Foundation 2015) was used to interpolate the influenza spread throughout the entire state of Ohio. By default, Inverse Distance Weighting (IDW) Interpolation is used, but other GDAL supporting interpolation algorithm can also be used. Since the resulting interpolated raster is a rectangular area extends beyond the boundary of Ohio, it was clipped by the boundary of Ohio using the gdal_wrap functionality. All GDAL functionalities are accessed through the Geobricks (Barbaglia and Murzilli 2011) library which is an open source Java wrapper for GDAL.

Generated rasters for each day are uploaded to a predefined workspace of the GeoServer (GeoServer 2015) instance with necessary spatial reference information using the GeoServerManager REST client library (GeoSolutions 2015). Uploaded rasters are available in the GeoServer as web map layers for web applications. Daily influenza maps which consist of number of influenza related tweets captured are automatically updated by hour reflecting latest available data.

1.3.7. Web Application

The influenza maps shown in Fig. 6 are displayed on a simple web application. The web application is hosted in an Apache Tomcat web server and integrates HTML (Hypertext Markup Language), Cascading Style Sheets (CSS) and Java Script. In addition, jQuery that is a set of JavaScript libraries designed to simplify HTML document traversing, animation, event handling, and Ajax interactions was used for the development of the client side. Daily influenza map layers stored in GeoServer are programmatically accessed through the Web Mapping Service (WMS) of GeoServer using the OpenLayers (OpenLayers 2015) JavaScript mapping library. Loaded daily maps are animated along the time to better visualize the variation of spatial and temporal spread of influenza. Further, generated statistical results including daily histograms and weekly influenza spread line charts are displayed along with the animated maps.

CHAPTER II. RESULTS AND DISCUSSIONS

The results generated by the prototype software framework that quantify influenza related tweets in Ohio were generated for a total of 21 week period. The results were used for comparisons purposes against real influenza cases that came from hospitalizations reports issued by Ohio Department of Health (ODH) and morbidity and mortality weekly reports issued by CDC. Although the morbidity and mortality weekly reports show pneumonia related deaths, the weekly reports do not link pneumonia deaths directly to influenza cases. Often, the weekly reports are characterized by a time leg between initial reports of influenza symptoms and consequent hospitalization treatment or potential deaths. Thus, the comparison between real influenza cases and tweeter generated influenza cases is aimed at understanding possible similarities in the trends generated by both datasets and to show possible widespread influenza activity in time and space where the tweets originated.

In the state of Ohio, the Ohio Department of Health (ODH) generates influenza related laboratory surveillance and influenza related hospitalizations reports. The influenza laboratory surveillance reports are perhaps the most reliable because they represent actual influenza cases. However, the availability of influenza laboratory surveillance reports is limited to few larger cities in Ohio. On the other hand, the influenza related hospitalizations are reported weekly, and they lack spatial detail because multiple counties in Ohio are aggregated into regions. Figure 7 shows a total of seven regions labeled as: Northwest, Northeast, West Central, Central, East Central, Southwest, and Southeast. In this work, the comparison is based on this regional level of influenza that reports cases that are counted weekly.

The time-series plot in Figure 8 shows the influenza cases generated by the ontology-based twitter prototype and the ODH influenza related hospitalizations for the seven regions. The

x-axis shows the time period comparison for each region, starting from the 40th week of 2015 and ending by the 13th week of 2016 (October 4, 2015 to April 02, 2016). The *y*-axis shows the number of cases generated by the twitter prototype and the hospitalizations reports. The scale of the *y*-axis varies because of differences in population, demographics, and perhaps tweeter users in rural or urban areas. It is interesting that all plots show higher influenza cases generated by the twitter prototype in the first part of the time-series (i.e. dashed vertical line week 8th), followed by abrupt jump associated with higher influenza cases from hospitalizations reports . For instance, the Central Ohio plot which is associated with a large student population and potential twitter users from Ohio State University shows the highest volume of tweets, while the West Central plot shows the lowest volume of tweets. The Pearson's correlations between the twitter prototype identified influenza cases and the hospitalizations reports are present for the beginning period of the validation. For instance, the highest correlations in the beginning period (i.e. left from dashed vertical line) are associated with the Southeast plot ($r = 0.52$), the Northwester plot ($r = 0.38$), and the Central plot ($r = 0.33$). The West Central plot has the weakest correlation ($r = -0.03$). Although there are no correlations in the last period of the validation in the figure, it is noticeable that there is a significant spike for all regions that shows similar pattern for the influenza cases generated from hospitalizations reports. However, at this point the cause that generates those differences that coincides with the end of the winter period is unclear. One possible cause based on data from the US National Center for Health Statistics (CDC - http://www.cdc.gov/nchs/deaths.htm) suggest that in the US the months of December, January, February and March are associated with the highest mortality. According to CDC (J *et al.* 2016) reports most of the influenza related deaths are recorded for very young (under 1 year) and older age groups (over 65 years) which are the most likely non-twitter users. In addition, the

uncertainties associated with natural language processing in this prototype require additional improvements and testing with new techniques such as artificial inelegance, pattern recognition, and data mining techniques.

Also, the correlations reported here are maybe deceptive for number of reasons. For instance, the twitter dataset could be further processed to discard retweets and successive posts by one user that is forwarded by another user. Such, retweets do not indicate new cases but represent a noise that can significantly increase the amount of cases reported with influenza. Also, an individual user may have multiple encounters with a single episode which can cause duplication of reporting influenza that relates to the same case. Thus, filtering tweets based on time constraint could be used to eliminate reports from the same users. Another consideration is that the Twitter data generates a real-time assessment of potential influenza cases while the aggregated ODH data has time lag in the actual reporting of the data. On the positive side, the strength of the proposed methodology is that it allows a real-time assessment in space and time that can be used for preparedness of potential spread of illnesses such as influenza at finer time and space scales.

The relationship between the twitter generated influenza cases and influenza related cases from hospitalizations was further scrutinized for possible lagged correlations. Lagged correlation is the correlation between two time-series where one of the series exhibits shift due to delayed responses in time relative to the other series. For example, there is a delay between first symptoms, consultation with a physician, time of diagnosis, initiation of treatment and hospitalization that is reported by ODH. To account for potential delays associated with one of the time-series a sample cross correlation function (CCF) is often used to identify the time lags.

For instance, plots from the lagged correlation intend to show two time-series that are shifted in time relative to one another on the *x*-axis.

Figure 9 shows the cross-correlation between different lags that were calculated for the seven regions using the CCF. The horizontal blue lines in the plots represent the upper and the lower 95% confidence levels for significance of the CCF. The confidence interval is computed from a sample size and relies on several assumptions including 1) the time-series are uncorrelated, 2) the processes are not autocorrelated, 3) the populations are normally distributed, and 4) the sample size is large. For a two-tailed test, the approximate 95% confidence interval is $+/- 1.96/\sqrt{N} = +/- 0.4277$ where the sample size is N. A CCF estimates which exceed the confidence interval are considered to be significant with a lag that has an autocorrelation (ACF) that is beyond the dotted line. Also, the significant CCF estimates above the upper ACF line show positive cross-correlations and below the lower ACF line show negative cross-correlations.

Moreover, in Figure 9 cross-correlations significance can be seen for the Central, Southwest, East Central, and West Central regions which are among the most populated regions in Ohio. The cross-correlations significance patterns (correlating lags) are different between the regions. For example, the central regions (Central, East Central, and West Central) have one or more CCF estimates above the ACF which suggest lagged correlation. For example, there are two CCF estimates above the ACF associated with negative values for weeks 6 and 8 which represents correlation between Twitter generated and hospitalizations. Also there are CCF estimates above the ACF associated with the East Central and the West Central with negative lags that ranges between 4 and 8 weeks. The Southwestern region shows a significant cross-correlation between two time series with no lags (i.e. lag is 0) which suggest absence of lagged correlation. The real cause of this delay may suggest that different administrative procedures

may have been in place for different regions and hospitalizations were reported differently. Also differences may have been caused by different parameters such as available health care facilities, mean annual household income and usage of Twitter. For example, the most common Twitter users or 31% of the population are associated with the 18 - 29 age group while 9% of the population are associated with 50 - 64 age group (Social Media Update 2013 | Pew Research Center 2016). Since the demographics varies across the state of Ohio especially in rural and urban areas, the results could be further scrutinized for other critical factors that may affect the information derived from the Twitter.

Figure 10 shows the spatio-temporal distribution of Twitter derived influenza cases in Ohio for a total of 8 weeks period. The maps in the figure are compiled by interpolation of tweets that were aggregated on weekly bases. The period that is shown in the figure starts at the 18th of October, 2015 and ends at the 12th of December, 2015. The weekly data in the figure was processed and aggregated to match the influenza reports issued by ODH. The figure shows that higher number of influenza related tweets are associated with densely populated areas of major Ohio cities such as Columbus, Cincinnati, Cleveland, Toledo, Dayton, Akron and Athens and their suburbs. In particular, the Columbus area in the Central region shows that influenza related tweets are reported almost through the entire 8 week period. On the other hand cities such as Athens show that higher volume of tweets were reported in particular weeks such as weeks 43 and 49. Other cities such as Cleveland show variability through time where higher volume of tweets were reported at weeks 42, 46, and 48. Although, in this research it is difficult to extract the exact causes of influenza patterns, the visualization of the interpolated surfaces allows for examining space-time activity patterns that may be used for subsequent predictions of spared, transmission, and potential influenza infections. The output maps also allow for additional

analysis end exploration of relationships of important factors such as population density, human movement, interacting dimensions which include large gathering events during holidays that may lead to better real-time surveillance and understanding predictive potential of spread.

CHAPTER III. CONCLUSION

The goal of this research was to develop a prototype software framework for formulating spatio-temporal influenza (i.e. flu) outbreaks using Twitter. The proposed framework collected geo-tagged tweets that were generated within Ohio and processed by semantic techniques in Natural Language Processing to extract influenza related tweets. The prototype system was calibrated to maximize the detection of influenza using rules developed by different semantic similarity measurements. The resulting output from the prototype represented a visualization of potential spatio-temporal influenza patterns and spread which is a product of information extraction from the tweets.

Implementation of the prototype software framework used a three-tier system architecture with a thin web client and a resourceful server environment. The server side component comprises of data processing applications, a PostGIS extension that manages the spatial PostgreSQL database, and a web application. The data processing applications enabled daemon parallel computing (i.e. background processing) for meeting pseudo real-time data processing requirements. The DBpediaSpotlight tagger which is based on the DBpedia ontology was used for the semantic tagging, while WS4J module was used for calculating semantic similarity scores. The WS4J Java module which is supported by WordNet ontologies was used to generate different semantic similarity scores from tags and synsets using WUP and RES algorithms. The similarity scores from both algorithms produced different individual and overall scores which were used for subsequent calibration. The calibration used a total of 1400 randomly selected tweets that contained a total of 107 influenza related tweets (i.e. positives) and 1293 non-influenza related tweets (i.e. negatives). The calibration tweets were used to explore different rules from the semantic similarity scores for detecting influenza related tweets and for filtering

out non-influenza related content. The rule that generated the best outcome was associated with a total of 94.4% accuracy (ACC) and *F*-measure value of 0.72. The execution of the rule extracted influenza related geo-tagged tweets that were used for development of influenza maps. The spatial implementation used GDAL and GeoTools open source tools for producing weekly influenza maps. The influenza maps were programmatically published on GeoServer which is a web based geospatial data sharing server that allows map services to be used by standard clients such as web browsers and GIS desktop programs. In this prototype, the web application was developed for retrieval and visualization of influenza maps from GeoServer. The open source JavaScript libraries OpenLayers and GeoExt were used in the web application for displaying the web-based maps in a user-friendly environment.

In summary, this prototype system demonstrated a robust tool for real-time assessment and monitoring of potential influenza cases using Twitter. The prototype can be customized for a specific geographical region and visualization of spatio-temporal distribution of influenza over the web. Also, the system can be customized for other subject areas with improved functionalities such as dynamic manipulation of ontologies or formulating personalized sets of rules. The results of the influenza related content generated by the proposed framework were validated against weekly influenza related hospitalization records reported by ODH. Validation results showed some promising results. Although the volume of the influenza tweets generated by the prototype was consistently higher than influenza reported from hospitalizations, the Pearson's correlations suggested existence of moderate correlations for the Southeast region (r = 0.52), the Northwestern region (r = 0.38), and the Central region (r = 0.33). The Southwestern region shows a significant cross-correlation between influenza related hospitalizations time series and Twitter reported influenza cases time series with no lags (i.e. lag is 0). However, only

a lagged relationship can be seen in other regions. Further, it was discovered that influenza reported over Twitter has a better no-lag cross-correlation in regions which have a higher younger population.

Additional recommendation for improvement of the prototype include more advance processing and pre-processing of tweets, application of different ontological rules, application of better semantic similarity algorithms for the evaluation of similarity scores, and development of robust rules from similarity scores. For example, text mining techniques such as text clustering and sentimental analysis can be further explored for additional improvements of the system. Also, at larger geographical scale (i.e. urban centers) the system can be tested for highly populated cities such as New York or Chicago where the percentage of Twitter users is high and there is access to weekly influenza reports.

REFERENCES

About GeoTools — GeoTools [online], 2015. Available from: http://geotools.org/about.html [Accessed 16 Oct 2015].

About the Flu | Flu.gov [online], 2015. Available from: http://www.flu.gov/about_the_flu/index.html [Accessed 23 Mar 2015].

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B., 2011. Predicting flu trends using twitter data. *In*: *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 702–707.

Aramaki, E., Maskawa, S., and Morita, M., 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *In*: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1576.

Arias, M., Arratia, A., and Xuriguera, R., 2013. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology*, 5 (1), 1–24.

Barbaglia, G. and Murzilli, S., 2011. *Geobricks*.

Bermingham, A. and Smeaton, A.F., 2011. On using twitter to monitor political sentiment and predict election results.

Blyth, C.C., Foo, H., van Hal, S.J., Hurt, A.C., Barr, I.G., McPhie, K., Armstrong, P.K., Rawlinson, W.D., Sheppeard, V., Conaty, S., Staff, M., Dwyer, D.E., and on behalf of the World Youth Day 2008 Influenza Study Group, 2010. Influenza Outbreaks during World Youth Day 2008 Mass Gathering. *Emerging Infectious Diseases*, 16 (5), 809–815.

Bodnar, T., Barclay, V.C., Ram, N., Tucker, C.S., and Salathé, M., 2014. On the ground validation of online diagnosis with Twitter and medical records. *In*: *Proceedings of the*

companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 651–656.

Bontcheva, K. and Rout, D., 2014. Making sense of social media streams through semantics: a survey. *Semantic Web*, 5 (5), 373–403.

Ceron, A., Curini, L., and Iacus, S.M., 2015. Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters--Evidence From the United States and Italy. *Social Science Computer Review*, 33 (1), 3–20.

Chang, Y., Dong, A., Kolari, P., Zhang, R., Inagaki, Y., Diaz, F., Zha, H., and Liu, Y., 2013. Improving recency ranking using twitter data. *ACM Transactions on Intelligent Systems and Technology*, 4 (1), 1–24.

Chen, L., Achrekar, H., Liu, B., and Lazarus, R., 2010. Vision: towards real time epidemic vigilance through online social networks: introducing SNEFT–social network enabled flu trends. *In*: *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*. ACM, 4.

Chew, C. and Eysenbach, G., 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5 (11), e14118.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20 (3), 273–297.

Crooks, A., Croitoru, A., Stefanidis, A., and Radzikowski, J., 2013. #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17 (1), 124–147.

Culotta, A., 2010. Towards detecting influenza epidemics by analyzing Twitter messages. *In*: *Proceedings of the first workshop on social media analytics*. ACM, 115–122.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K., 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *In*: *RANLP*. 198–206.

Espino, J.U., Hogan, W.R., and Wagner, M.M., 2003. Telephone triage: a timely data source for surveillance of influenza-like diseases. *In*: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 215.

Fauci, A.S., 2006. Seasonal and pandemic influenza preparedness: science and countermeasures. *Journal of Infectious Diseases*, 194 (Supplement 2), S73–S76.

Gault, G., Larrieu, S., Durand, C., Josseran, L., Jouves, B., and Filleul, L., 2009. Performance of a syndromic system for influenza based on the activity of general practitioners, France. *Journal of Public Health*, 31 (2), 286–292.

Gelernter, J. and Mushegian, N., 2011. Geo-parsing Messages from Microtext. *Transactions in GIS*, 15 (6), 753–773.

GeoServer [online], 2015. Available from: http://geoserver.org/ [Accessed 23 Mar 2015].

GeoSolutions, 2015. *GeoServer-Manager*.

Grassi, M., Cambria, E., Hussain, A., and Piazza, F., 2011. Sentic Web: A New Paradigm for Managing Social Media Affective Information. *Cognitive Computation*, 3 (3), 480–489.

Griffin, B., Jain, A.K., Davies-Cole, J., Glymph, C., Lum, G., Washington, S.C., and Stoto, M.A., 2009. Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system. *BMC Public Health*, 9 (1), 483.

Guarino, N., 1998. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*. IOS press.

Guo, D. and Chen, C., 2014. Detecting Non-personal and Spam Users on Geo-tagged Twitter Network. *Transactions in GIS*, 18 (3), 370–384.

Hedden, H., 2008. How SEMANTIC TAGGING Increases Findability. *EContent*, 31 (8), 38–43.

Hua, W., Huynh, D.T., Hosseini, S., Lu, J., and Zhou, X., 2012. Information extraction from microblogs: A survey. *Int. J. Soft. and Informatics*, 6 (4), 495–522.

Influenza-like Illness Surveillance Program (ILINet) [online], 2015. Available from: https://www.health.ny.gov/diseases/communicable/influenza/surveillance/ilinet_program/ [Accessed 18 Mar 2015].

J, X., Sl, M., Kd, K., and Ba, B., 2016. Deaths: Final Data for 2013. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 64 (2), 1–119.

Jiang, J.J. and Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Keogh-Brown, M.R., Wren-Lewis, S., Edmunds, W.J., Beutels, P., and Smith, R.D., 2010. The possible macroeconomic impact on the UK of an influenza pandemic. *Health Economics*, 19 (11), 1345–1360.

Kostkova, P., Szomszor, M., and St. Louis, C., 2014. #swineflu: The Use of Twitter as an Early Warning and Risk Communication Tool in the 2009 Swine Flu Pandemic. *ACM Transactions on Management Information Systems*, 5 (2), 1–25.

Kumar, A., Jiang, M., and Fang, Y., 2014. Where not to go?: detecting road hazards using twitter. ACM Press, 1223–1226.

Lamb, A., Paul, M.J., and Dredze, M., 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. *In*: *HLT-NAACL*. 789–795.

Lampos, V. and Cristianini, N., 2010. Tracking the flu pandemic by monitoring the social web. *In*: *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. IEEE, 411–416.

Lee, K., Agrawal, A., and Choudhary, A., 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. *In*: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1474–1477.

Lee, R., Wakamiya, S., and Sumiya, K., 2011. Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs. *World Wide Web*, 14 (4), 321–349.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C., n.d. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*.

Lucas, C., 2012. Multi-criteria modelling and clustering of spatial information. *International Journal of Geographical Information Science*, 26 (10), 1897–1915.

Magruder, S., 2003. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24 (4), 349–53.

Mann, P., O'Connell, E., Zhang, G., Llau, A., Rico, E., and Leguen, F.C., 2011. Alert System to Detect Possible School-based Outbreaks of Influenza-like Illness. *Emerging Infectious Diseases*, 17 (2), 262–264.

Meltzer, M.I., Cox, N.J., Fukuda, K., and others, 1999. The economic impact of pandemic influenza in the United States: priorities for intervention. *Emerging infectious diseases*, 5, 659–671.

Mendes, P.N., Jakob, M., García-Silva, A., and Bizer, C., 2011. DBpedia spotlight: shedding light on the web of documents. *In*: *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 1–8.

Miller, G.A., 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38 (11), 39–41.

Miyabe, M., Miura, A., and Aramaki, E., 2012. Use Trend Analysis of Twitter After the Great East Japan Earthquake. *In*: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*. New York, NY, USA: ACM, 175–178.

Nebhi, K., 2012. Ontology-Based Information Extraction from Twitter.

Ohio QuickFacts from the US Census Bureau [online], 2016. Available from: http://www.census.gov/quickfacts/table/PST045215/39 [Accessed 14 Mar 2016].

Open Source Geospatial Foundation, 2015. *GDAL-Geospatial Data Abstraction Library: version 2.0.0*.

OpenLayers 3 - Welcome [online], 2015. Available from: http://openlayers.org/ [Accessed 23 Mar 2015].

Oussalah, M., Bhat, F., Challis, K., and Schnier, T., 2013. A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems*, 37, 105–120.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N., 2012. Part-of-speech tagging for Twitter: Word clusters and other advances. *School of Computer Science, Carnegie Mellon University, Tech. Rep*.

Pustejovsky, J. and Boguraev, B., 1996. Lexical semantics. *Lexical Semantics*.

Rao, T. and Srivastava, S., 2012. Analyzing stock market movements using twitter sentiment analysis. *In*: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 119–123.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Rutter, P., Mytton, O., Ellis, B., and Donaldson, L., 2014. Access to the NHS by telephone and Internet during an influenza pandemic: an observational study. *BMJ open*, 4 (2), e004174.

Sasaki, A., Hoen, A.G., Ozonoff, A., Suzuki, H., Tanabe, N., Seki, N., Saito, R., and Brownstein, J.S., 2009. Evidence-based tool for triggering school closures during influenza outbreaks, Japan. *Emerging infectious diseases*, 15 (11), 1841.

Shima, H., 2013. *WS4J-WordNet Similarity for Java*.

Signorini, A., Segre, A.M., and Polgreen, P.M., 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 6 (5), e19467.

Smith, R.D., Keogh-Brown, M.R., Barnett, T., and Tait, J., 2009. The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. *BMJ*, 339 (nov19 1), b4571–b4571.

Social Media Update 2013 | Pew Research Center [online], 2016. Available from: http://www.pewinternet.org/2013/12/30/social-media-update-2013/ [Accessed 10 Mar 2016].

Stewart, A. and Diaz, E., 2012. Epidemic Intelligence: For the Crowd, by the Crowd. *In*: M. Brambilla, T. Tokuda, and R. Tolksdorf, eds. *Web Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 504–505.

Tao, K., Abel, F., Hauff, C., and Houben, G.-J., 2012. Twinder: A Search Engine for Twitter Streams. *In*: *Proceedings of the 12th International Conference on Web Engineering*. Berlin, Heidelberg: Springer-Verlag, 153–168.

Tweets | Twitter Developers [online], 2015. Available from: https://dev.twitter.com/overview/api/tweets [Accessed 16 Oct 2015].

Vakali, A., Giatsoglou, M., and Antaris, S., 2012. Social Networking Trends and Dynamics Detection via a Cloud-based Framework Design. *In*: *Proceedings of the 21st*

International Conference Companion on World Wide Web. New York, NY, USA: ACM, 1213–1220.

Vieweg, S., Hughes, A.L., Starbird, K., and Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *In*: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1079–1088.

Vu, T.-T., Chang, S., Ha, Q.T., and Collier, N., 2012. An experiment in integrating sentiment features for tech stock prediction in twitter.

Walther, M. and Kaisser, M., 2013. Geo-spatial event detection in the twitter stream. *In*: *Advances in Information Retrieval*. Springer, 356–367.

Wikramaratna, P.S. and Gupta, S., 2009. Influenza outbreaks. *Cellular Microbiology*, 11 (7), 1016–1024.

Wu, Z. and Palmer, M., 1994. Verbs semantics and lexical selection. *In*: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 133–138.

Yamamoto, Y., 2010. *Twitter4J-A Java Library for the Twitter API*.

Yoldascan, E., Kurtaran, B., Koyuncu, M., and Koyuncu, E., 2010. Modeling the Economic Impact of Pandemic Influenza: A Case Study in Turkey. *Journal of Medical Systems*, 34 (2), 139–145.

APPENDIX A: FIGURES

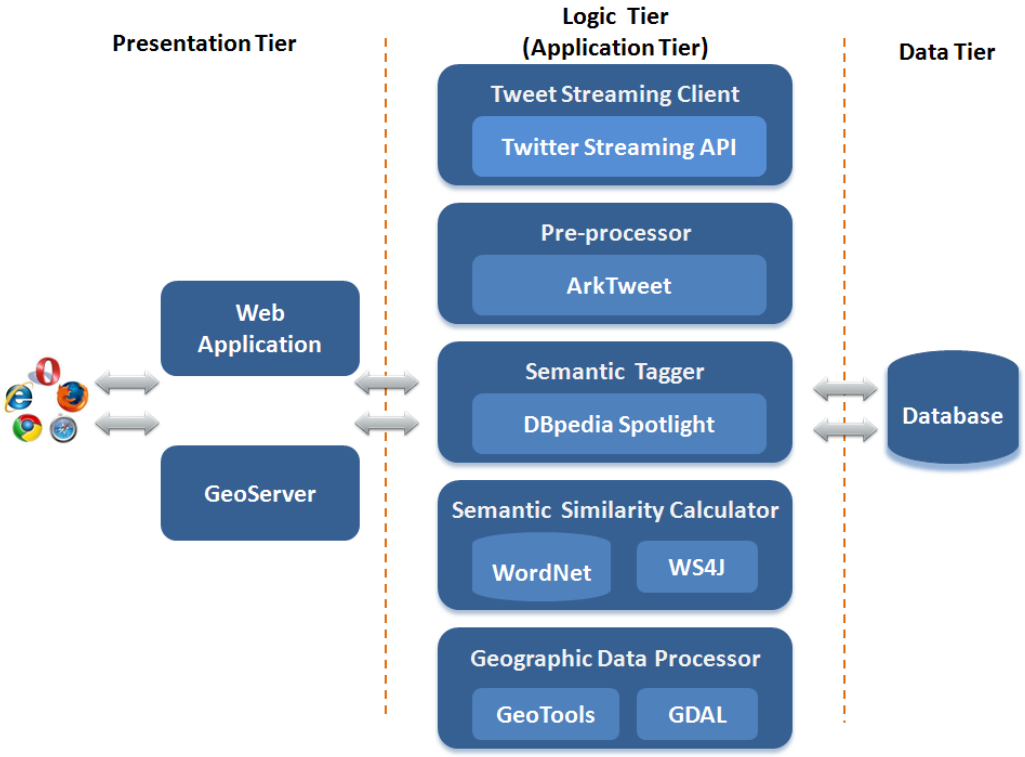

Figure 1. System architecture.

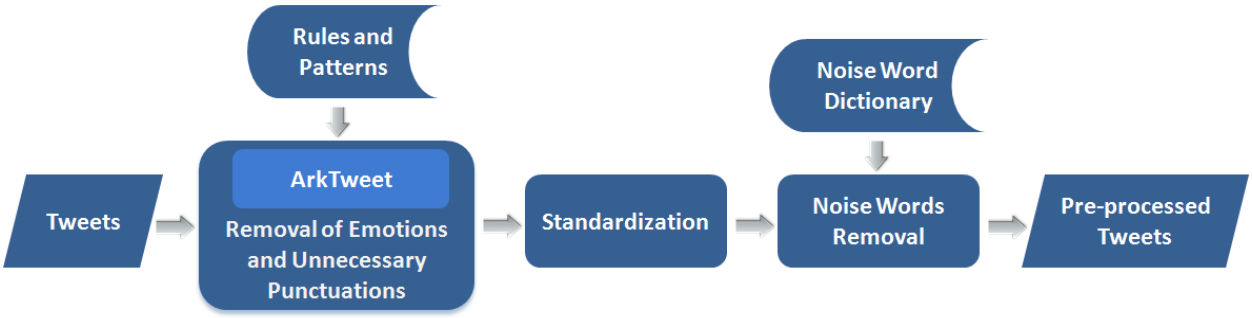

Figure 2. Tweet pre-processing steps

a)

b)

Figure 3. Semantic tagging of a tweet by *a*) DBpediaSpotlight web interface which maps URIs as search queries and *b*) and HTML results from the "Influenza" URI ontology shown as dictionary of descriptive terms.

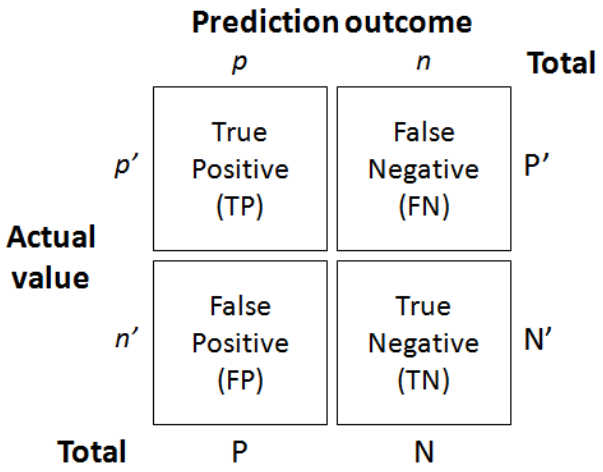

Figure 4**.** An example of a confusion matrix for a binary classifier.

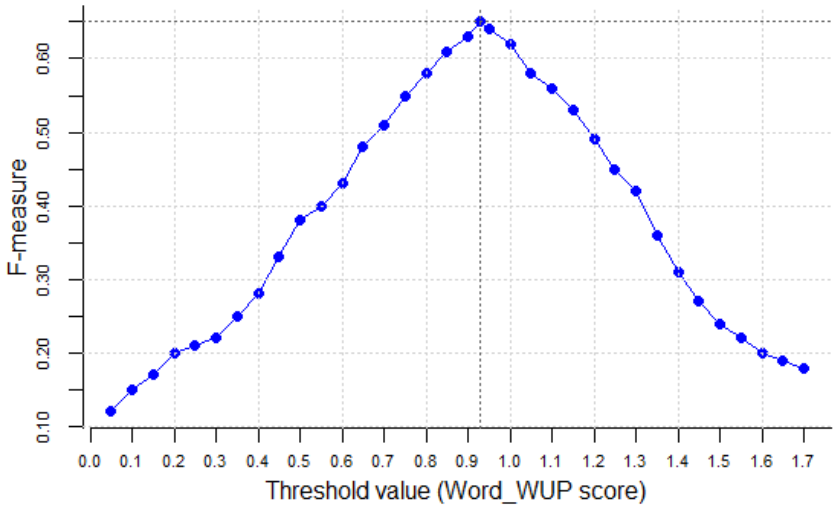

Figure 5. Calibration curve of semantic similarity scores for optimizing threshold "Word_WUP" value using *F*-measures generated by confusion matrices

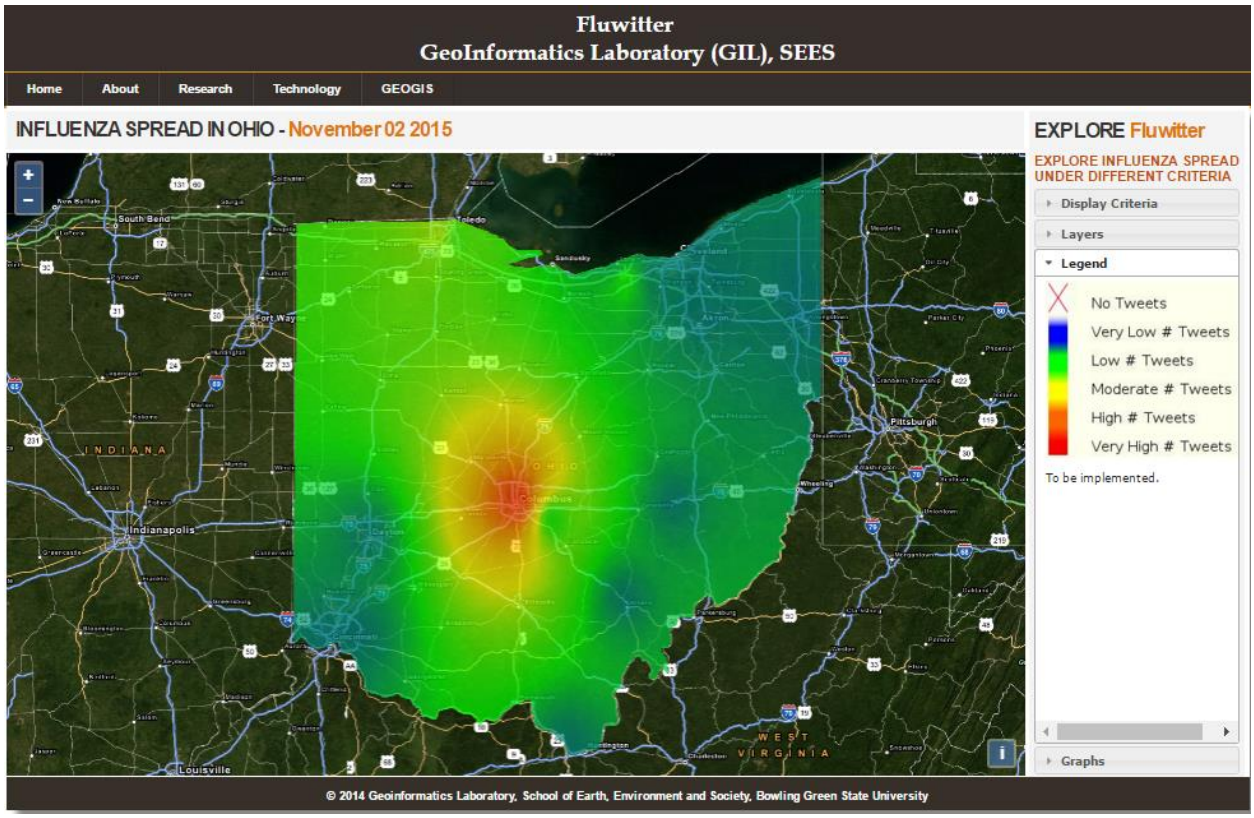

Figure 6. Fluwitter – graphical user interface for visualization of influenza related tweets in Ohio

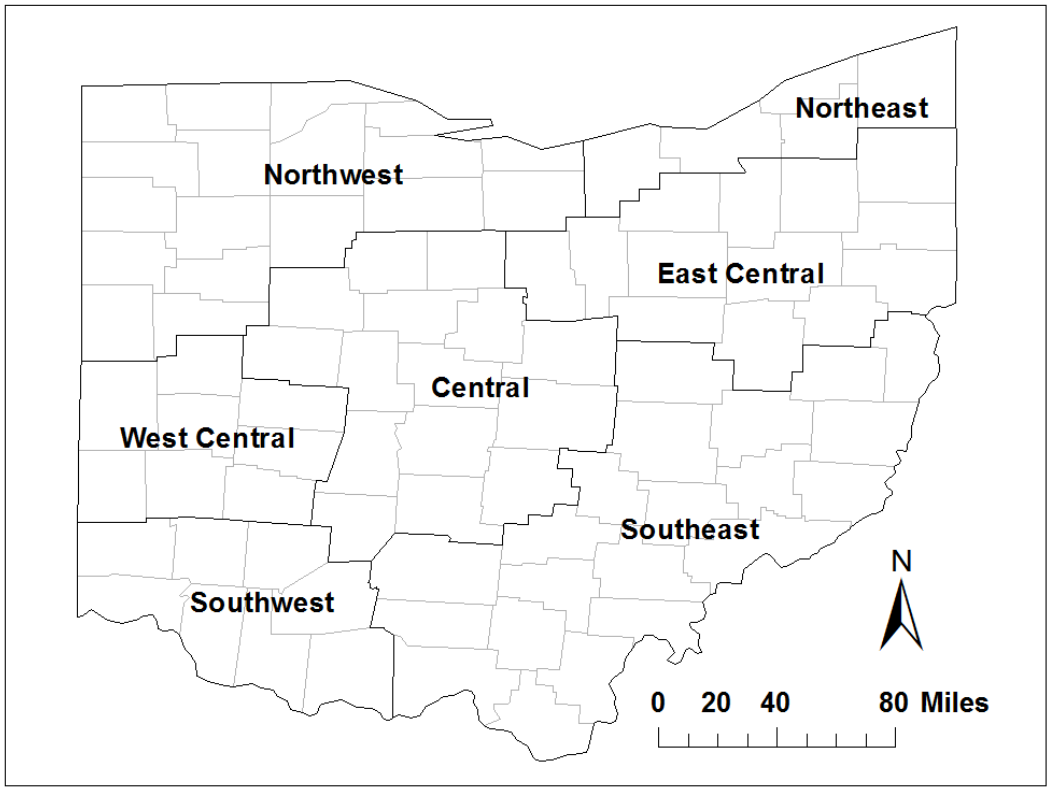

Figure 7. Seasonal influenza reporting regions in Ohio

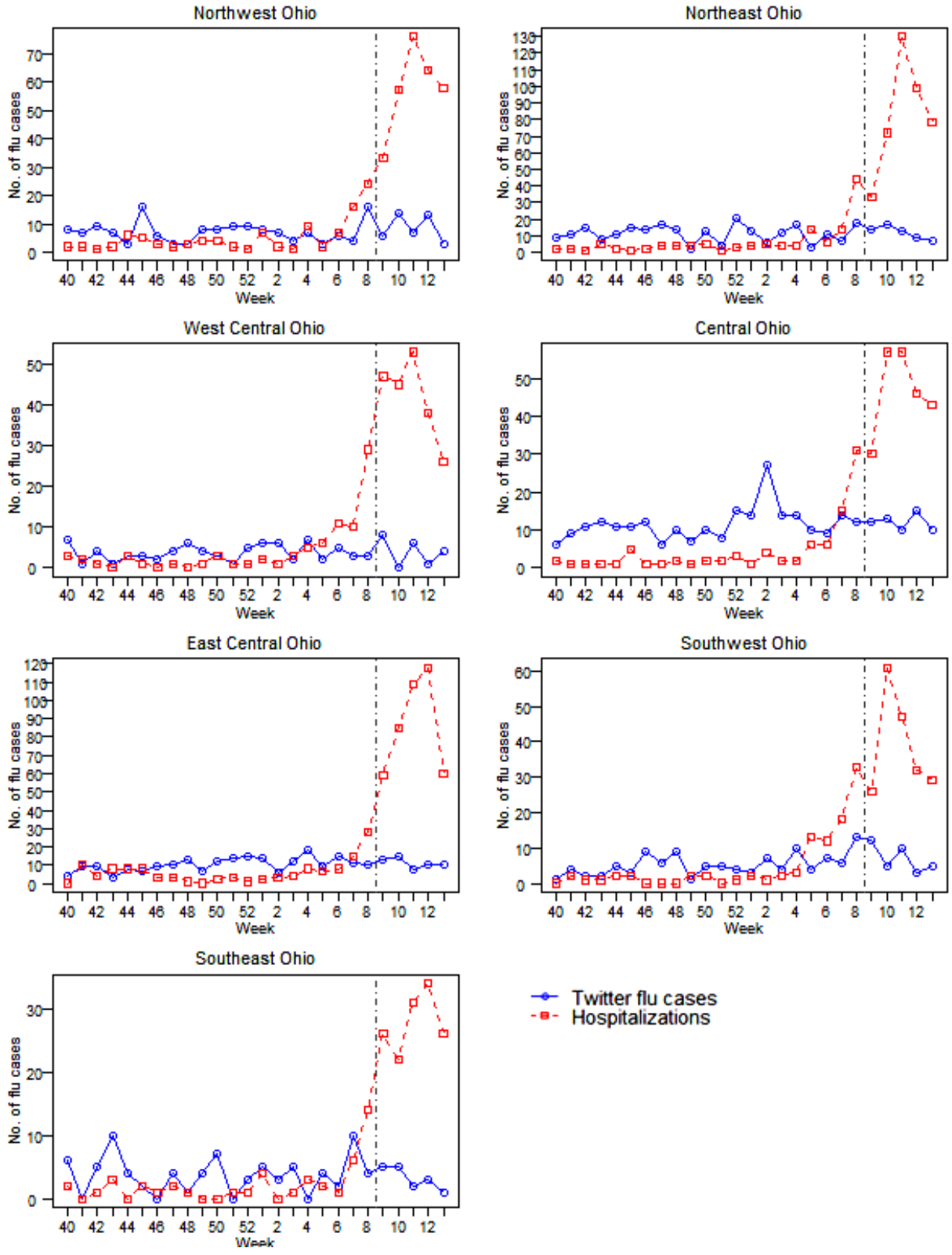

Figure 8. Twitter derived influenza cases and influenza reported hospitalizations using regional weekly summaries in Ohio for the period of October 4, 2015 and April 02, 2016

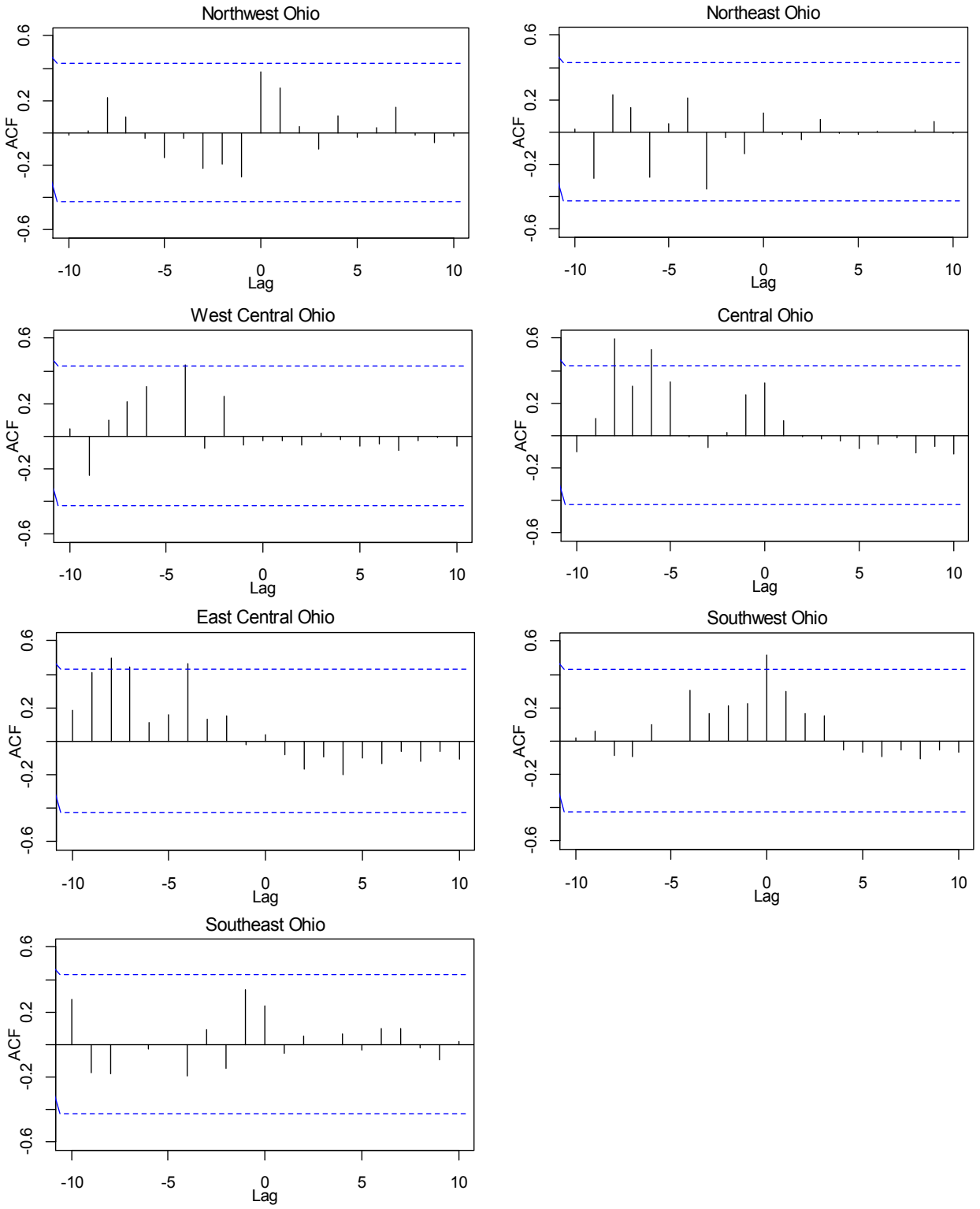

Figure 9**.** Cross correlation between Twitter reported influenza cases and influenza related hospitalizations

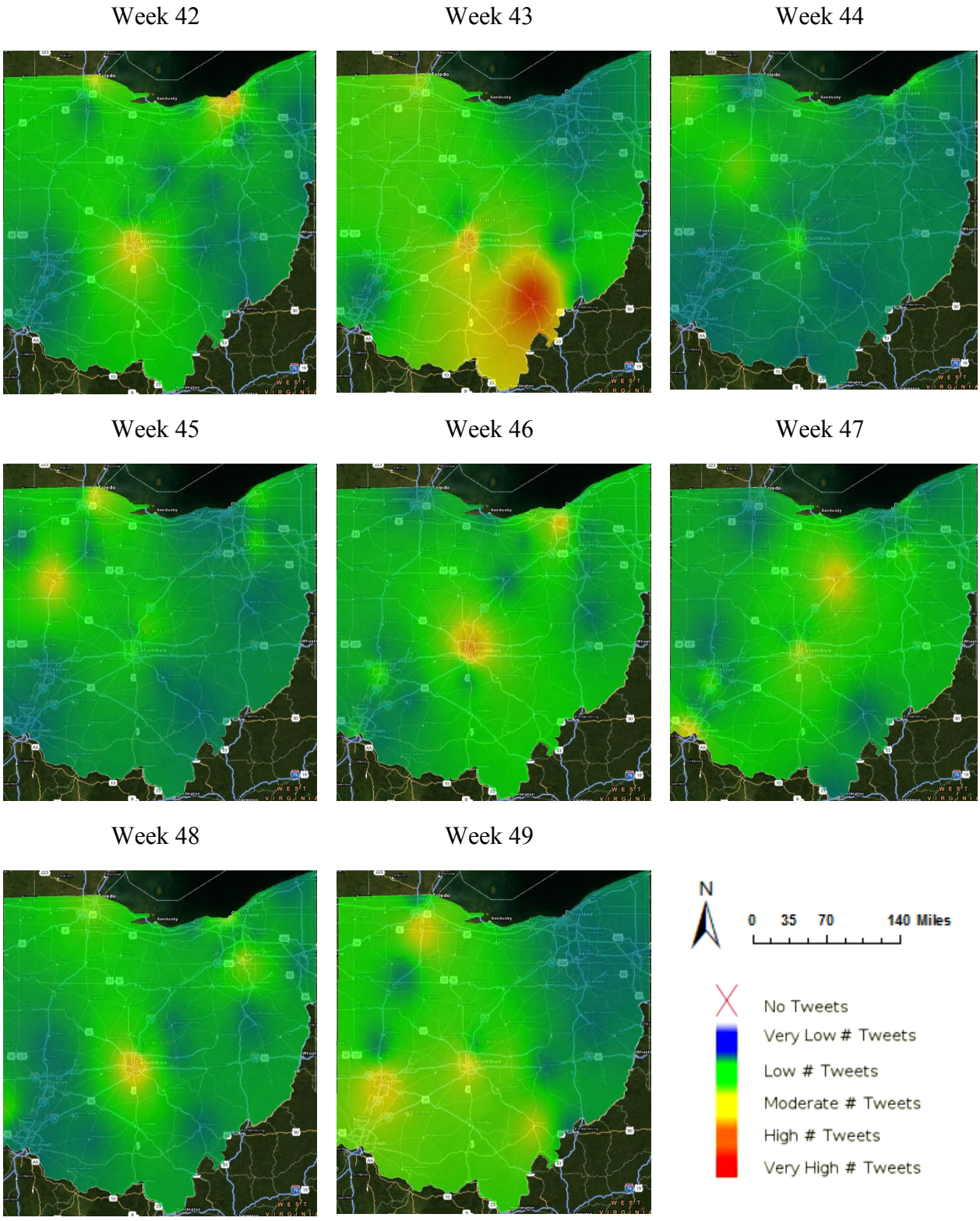

Figure 10. Spatio-temporal pattern of Twitter derived influenza cases in Ohio from week 42 to week 49 (Oct 18 - Dec 12 2015). Very Low : 1 , Low : 2-3, Moderate : 4-5, High: 6-10, Very High : >= 11

APPENDIX B: TABLES

Field Name	Type	Description
id	64 bit integer	Integer representation of unique tweet identifier
text	string	Text content of the tweet
source	string	Utility used to post the tweet. ex:- web, i-phone, android
longitude	float	Longitude of the tweet originated location
latitude	float	Latitude of the tweet originated location
created_at	Date and time	UTC time when the tweet was created.
place_id	string	Unique identifier of the tweet originated place
place_name	string	Name of the tweet originated place
place_type	string	Type of the place tweet originated in
place_polygon	string	Polygon representing the message originated place

Table 1. Attributes extracted from a tweet