

EFFICIENT TEXT CLASSIFICATION WITH LINEAR
REGRESSION USING A COMBINATION OF
PREDICTORS FOR FLU OUTBREAK DETECTION

Ali Al Essa

Under the Supervision of

Dr. Miad Faezipour

DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIRMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOHPY IN COMPUTER SCIENCE

AND ENGINEERING

THE SCHOOL OF ENGINEERING

UNIVERSITY OF BRIDGEPORT

CONNECTICUT

December, 2018

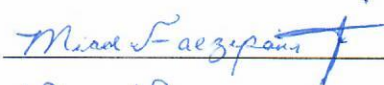

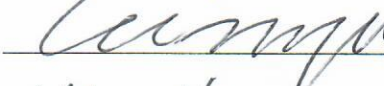
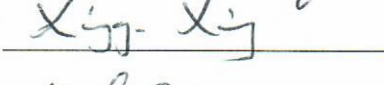
EFFICIENT TEXT CLASSIFICATION WITH LINEAR REGRESSION
USING A COMBINATION OF PREDICTORS FOR FLU OUTBREAK
DETECTION

Ali Al Essa

Under the Supervision of Dr. Miad Faezipour

Approvals

Committee Members

Name	Signature	Date
Dr. Miad Faezipour		<u>Dec. 7, 2018</u>
Dr. Abdel-shakour Abuzneid		<u>Dec. 7, 2018</u>
Dr. Jeongkyu Lee		<u>12/11/18</u>
Dr. Xingguo Xiong		<u>Dec. 7, 2018</u>
Dr. Saeid Moslehpour		<u>Dec 6/18</u>

Ph.D. Program Coordinator

Dr. Khaled M. Elleithy		<u>12/11/2018</u>
------------------------	--	-------------------

Chairman, Computer Science and Engineering Department

Dr. Ausif Mahmood		<u>12-11-2018</u>
-------------------	--	-------------------

Dean, School of Engineering

Dr. Tarek M. Sobh		<u>12/11/2018</u>
-------------------	--	-------------------

EFFICIENT TEXT CLASSIFICATION WITH LINEAR
REGRESSION USING A COMBINATION OF
PREDICTORS FOR FLU OUTBREAK DETECTION

© Copyright by Ali Al Essa 2018

EFFICIENT TEXT CLASSIFICATION WITH LINEAR REGRESSION USING A COMBINATION OF PREDICTORS FOR FLU OUTBREAK DETECTION

ABSTRACT

Early prediction of disease outbreaks and seasonal epidemics such as Influenza may reduce their impact on daily lives. Today, the web can be used for surveillance of diseases. Search engines and Social Networking Sites can be used to track trends of different diseases more quickly than government agencies such as Center of Disease Control and Prevention (CDC). Today, Social Networking Sites (SNS) are widely used by diverse demographic populations. Thus, SNS data can be used effectively to track disease outbreaks and provide necessary warnings. Although the generated data of microblogging sites is valuable for real time analysis and outbreak predictions, the volume is huge. Therefore, one of the main challenges in analyzing this huge volume of data is to find the best approach for accurate analysis in an efficient time. Regardless of the analysis time, many studies show only the accuracy of applying different machine learning approaches. Current SNS-based flu detection and prediction frameworks apply conventional machine learning approaches that require lengthy training and testing, which is not the optimal solution for new outbreaks with new signs and symptoms.

The aim of this study is to propose an efficient and accurate framework that uses SNS data to track disease outbreaks and provide early warnings, even for newest outbreaks accurately. The presented framework of outbreak prediction consists of three main modules: text classification, mapping, and linear regression for weekly flu rate predictions. The text classification module utilizes the features of sentiment analysis and predefined keyword

occurrences. Various classifiers, including FastText and six conventional machine learning algorithms, are evaluated to identify the most efficient and accurate one for the proposed framework. The text classifiers have been trained and tested using a pre-labeled dataset of flu-related and unrelated Twitter postings. The selected text classifier is then used to classify over 8,400,000 tweet documents. The flu-related documents are then mapped on a weekly basis using a mapping module. Lastly, the mapped results are passed together with historical Center for Disease Control and Prevention (CDC) data to a linear regression module for weekly flu rate predictions.

The evaluation of flu tweet classification shows that FastText together with the extracted features, has achieved accurate results with an F -measure value of 89.9% in addition to its efficiency. Therefore, FastText has been chosen to be the classification module to work together with the other modules in the proposed framework, including the linear regression module, for flu trend predictions. The prediction results are compared with the available recent data from CDC as the ground truth and show a strong correlation of 96.2%.

To
My wonderful father and mother
My loving wife
My charming children, Eithar, Albatool, and Hussain
For their unconditional love, trust, and unending inspiration

ACKNOWLEDGEMENTS

I praise God who has provided me this opportunity and has granted me the capabilities to complete my studies successfully. This dissertation comes to its conclusion due to the assistance, guidance and trust of several people. I would like to thank all of them.

I owe my deepest sense of gratitude to my esteemed advisor Dr. Miad Faezipour for her thoughtful guidance, valuable comments, and for the freedom I was granted during the whole period of my Ph.D. researches. I am honored that my work has been supervised by her. She is professional and supportive. I sincerely thank Dr. Faezipour for her time, and help.

I would like to offer my sincere thanks to Dr. Abdel-shakour Abuzneid, Dr. Jeongkyu Lee, Dr. Xingguo Xiong, and Dr. Saeid Moslehpour for serving on my supervisory committee, taking the time to evaluate this dissertation, and providing their valuable feedback and comments.

My thanks are also to Professors Charles Campbell, and Camy Deck from the Tutoring and Learning Center for their proofreading of this document.

Most importantly, my deepest gratitude and sincere thanks are to my wonderful family for their understanding, encouragement, and inspiration. I am especially grateful to my parents, who always believe in me, for their love and trust. My grateful thanks are also to my loving and caring wife Sahar, and my three lovely children Eithar, Albatool, and Husain for their love and everlasting inspiration. Finally, I would like to express my sincere thanks to my sisters and brothers and also my friends for their positive influence.

TABLE OF CONTENTS

ABSTRACT	iv
ACKNOWLEDGEMENTS	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xi
LIST OF FIGURES	xii
ABBREVIATIONS	xiv
CHAPTER 1: INTRODUCTION	1
1.1 Research Problem and Scope	4
1.2 Motivation Behind the Research	5
1.3 Contributions of the Proposed Research	6
CHAPTER 2: BACKGROUND AND LITERATURE SURVEY	8
2.1 Introduction	8
2.2 Article Selection Methodology and Related Work	8
2.3 Methods	10
2.3.1 Text Mining	11
2.3.2 Graph data mining	13
2.3.3 Topic Models	15
2.3.4 Machine Learning Techniques	19
2.3.5 Math/Statistical Based Models	28
2.3.6 Mechanistic disease models	32
2.3.7 Detection Based on Filtered Keywords and Documents	35
2.4 Discussion	37
2.5 Challenges	42
2.5.1 Data Collection	42
2.5.2 Data Size	42
2.5.3 Language	42
2.5.4 Heterogeneity	42
2.5.5 Sampling bias	43

2.5.6	Dataset Consistency	43
2.5.7	User Location	43
2.5.8	Proxy Population	43
2.5.9	Spams	44
2.5.10	Evaluation	44
2.6	Concluding Remarks	44
CHAPTER 3: RESEARCH METHODOLOGY		45
3.1	Data Collection and Preparation	48
3.1.1	Classification Model Data	48
3.1.2	Application Dataset	49
3.1.3	CDC ILINet Data	49
3.1.4	Data of Hospital Emergency Department Syndromic Surveillance (HEDSS) System	50
3.2	Preprocessing	50
3.3	Feature Extraction	51
3.3.1	Textual Features	52
3.3.2	Stylometric features	53
3.3.3	Topic-related keywords based features	53
3.3.4	Sentiment based features	53
3.4	Classification Model Building - Training and Testing	55
3.4.1	FastText	55
3.4.2	Conventional Machine Learning Classifiers	55
3.5	Mapping	58
3.6	Weekly Flu Rate Estimation	60
3.6.1	Linear Regression Model	61
3.6.2	Other Regression Models	61
CHAPTER 4: IMPLEMENTATION AND TESTING		63
4.1	Flu Post Classification	64
4.2	Performance Metrics	65
4.2.1	Text Classification	65
4.2.2	Flu Rate Estimation	66
CHAPTER 5: RESULTS		67
5.1	Classification Results	67
5.2	Weekly Flu Rate Estimation Results	71
CHAPTER 6: DISCUSSION AND VALIDATION		73
6.1	Computational Complexity	73
6.2	Statistical Power Analysis	74
CONCLUSION		78

REFERENCES 80

Appendix A: More Application Data 95

LIST OF TABLES

Table 2.1	Summary of the used data sets in the reviewed studies	39
Table 2.2	Summary of the reviewed methods and techniques	40
Table 2.3	Journal/Conference backgrounds of the reviewed studies	41
Table 5.1	Performance of classifiers	69
Table 5.2	Summary of the reviewed flu posts classifiers (Flu-Relevant / Flu-Irrelevant)	70
Table 5.3	Performance of Flu rate estimator using different regression models .	71
Table 6.1	Summary of the reviewed studies with reported Pearson Correlation .	74

LIST OF FIGURES

Figure 2.1	Articles selection process	11
Figure 2.2	A method to monitor ILI and identify communities in Social Media .	14
Figure 2.3	Health state transition diagram	18
Figure 2.4	A framework for Influenza outbreak detection	25
Figure 2.5	The process of Neural Networks based detection	26
Figure 2.6	SNEFT architecture	31
Figure 3.1	Proposed framework overview	46
Figure 3.2	Methodology for text classification of flu tweets	47
Figure 3.3	Text preprocessing	51
Figure 3.4	General flow of Hadoop MapReduce programming approach	59
Figure 5.1	Performance comparison using ROC	68
Figure 5.2	FastText performance using different sets of features	69
Figure 5.3	Correlation between the proposed framework and CDC ILI rate using different regression models	72
Figure 5.4	Correlation between the proposed framework and CDC ILI rate	72
Figure A.1	Correlation between the proposed framework and CDC ILI rates (Nov. 2018)	95

Figure A.2 Correlation between the proposed framework and CDC ILI rates
using different regression models (Jan.-May, Nov. 2018) 96

ABBREVIATIONS

ACF Autocorrelation function

ARMA Auto regression moving average

ATAM Ailment topic aspect model

BOW Bag-Of-Words

CDC Center for Disease Control and Prevention

CNIC China Nation Influenza Center

ED Emergency Departments

EAKF Ensemble Adjustment Kalman Filter

FN False Negative

FP False Positive

FT FastText

HEDSS Data of Hospital Emergency Department Syndromic Surveillance

HFSTM Hidden Flu-State from Tweet Model

IDF Inverse Document Frequency Weighting

IDSC Infection Disease Surveillance Center

ILINet Influenza Like Illness Surveillance Network

KNN K-Nearest Neighbors

MI Mutual Information

MN Mentions

MR MapReduce

MSE Mean Square Error

NLTK Natural Language Processing Toolkit

RF Random Forest

RMSE Root Mean Squared Error

ROC Receiver Operating Characteristic

RT Retweets

SIRS Susceptible Infections Recovered Susceptible

SNEFT Social Network Enabled Flu Trends

SNS Social Networking Sites

SVM Support Vector Machine

SVR Support Vector Regression

TF Term Frequency Weighting

TF-IDF Term Frequency-Inverse Document Frequency

TN True Negative

TP True Positive

CHAPTER 1: INTRODUCTION

Public health is an important issue. Health care providers must be updated about the public health and disease outbreaks affecting their communities in order to take correct actions at the right time. To produce outbreak reports, typical disease surveillance systems depend on official statistics based on patient visits [1]. In the U.S., these reports are produced by the Center for Disease Control and Prevention (CDC) to inform healthcare providers about certain disease outbreaks such as Influenza outbreaks. CDC publishes flu-related reports using the United States Influenza Like Illness Surveillance Network (ILINet) that gathers flu-related information of outpatients from hundreds of healthcare providers around the U.S. ILINet shows accurate results in detecting flu outbreaks, but it is costly and takes a long time to issue the required reports. It is crucial for any disease surveillance system to collect related data and provide the reports as early as possible to prevent the spread of the disease. To this end, many solutions have been proposed to generate earlier outbreak warnings. Examples include volumes of telephone calls, over-the-counter drug sales [1], search engine logs [2, 3, 4, 5, 6, 7], and SNS data that can be used for real-time analysis for better services [8, 9, 10, 11, 12, 13, 14, 15]. When comparing the different resources used for surveillance, such as search engine logs, SNS data is more descriptive and available to the public. Since SNS provides detailed demographic information, the collected data can be used to simulate the spread of disease outbreaks with temporal analysis.

Social Networking Sites (SNS) are tools that include big data about users and their shared thoughts and ideas, in addition to real-time data of users' conversations and statuses.

The amount of data, aside from the growth of SNS users, represents the important role of SNS in real-time analysis and predictions in many areas [16, 17]. These areas include traffic [18, 19, 20, 21], disaster prediction [22, 23, 24, 25, 26], management [27, 28, 29], networking [30, 31], news [32, 33, 34, 35, 36], and many more. In the public health area, SNS provides an efficient resource to conduct disease surveillance and a communication tool to prevent disease outbreaks [37].

Based on our survey of disease outbreak detection models using social media data, we found that most studies and models were developed to detect Influenza outbreaks from SNS such as seasonal Influenza and the swine Influenza. The developed models can potentially be deployed for other disease outbreak detections and predictions. Although prediction and detection terms are used interchangeably throughout the study, the terms have different definitions. Flu detection refers to the process of discovering flu trends or flu cases that have already occurred. On the other hand, flu prediction collects data to predict flu trends. Furthermore, the term nowcasting refers to the process of predicting flu cases that have happened in real time, which surveillance systems overlook. Because of the surveillance system limitations, the need for new techniques and models, such as Google Flu Trend (GFT), is necessary in order to predict non-reflected flu cases. This nowcasting process is integrated into report revisions before the final reports are issued. Aside from nowcasting, the process of forecasting is used to predict actual flu cases in the future.

In this study, we relied on the Twitter microblog to conduct minute-by-minute analysis in order to track the high frequency of posted messages. We present a framework to track Influenza trends through Twitter postings. The framework includes preprocessing, feature extraction, Twitter documents classification, documents weekly-mapping, and weekly flu rate predictions. The preprocessing phase includes stemming and removal of stop words and ineffective characters, which are non-alphanumeric tokens. Then, the pre-processed data is used to extract features to be passed to a tweet classifier to distinguish

between flu-related tweets and unrelated ones. The flu-related documents are then mapped on a weekly basis. Finally, the mapped results are passed together with historical CDC data to an estimator for flu trend predictions.

The Twitter Microblogging site is used in this study because it is the most widely used Social Networking Site (SNS). It is an efficient resource to track trends for several reasons. First, the high frequency of posted messages helps to perform minute-by-minute analysis. Second, compared with search engine logs, Twitter posts are more descriptive and available for the public. In addition, more analysis can be performed by analyzing the users' profiles such as demographic data and specific details. Third, users of Twitter are of diverse ages, not only young people, but also middle aged, and technology savvy older population [15].

The generated data of SNS is valuable for real-time analysis and outbreak predictions, but its volume is huge. Therefore, one of the main challenges in analyzing this huge volume of data is to find the best approach for accurate analysis in an efficient time. Current Twitter-based flu detection and prediction frameworks apply conventional machine learning approaches that require lengthy training and testing which is not the optimal solution to be used for a new outbreak with new signs and symptoms. Regardless of the analysis time, many studies only report the accuracy of different machine learning approaches. Thus, more efficient solutions are required for accurate results with less processing time. In this study, we demonstrate that using FastText can enhance the efficiency of Twitter-based flu outbreak prediction models. Originally, FastText became an efficient text classifier that was proposed by Facebook. FastText performs more quickly than deep learning classifiers for training and testing procedures and produces comparably accurate results. The FastText classifier can train more than a billion words in about ten minutes and then predict multiple classes within half a million sentences in less than a minute [38].

1.1 Research Problem and Scope

SNS postings can be seen as triggers for different event prediction such as disease outbreaks. Discovering knowledge from the posts for flu surveillance models requires an efficient approach of text processing. It includes gathering the related text (posts) about the disease and then issuing necessary reports at an early stage that is crucial for outbreak prevention. Since the gathered data is unstructured, the first step is to preprocess the unstructured content in order to analyze the data and produce the results in an understandable way. The second step is feature extraction, which is a key to performance enhancement. The third step is knowledge extraction, using machine learning techniques for text classification that includes model training and testing. A post on a microblogging site is then classified into either related or unrelated classes, for example;

Related: *I'm sick, I got flu yesterday.*

Unrelated: *I'm sick of school.*

Our literature survey indicates that most of the existing frameworks use conventional machine learning classifiers [39]. These approaches require long time for the training process. A new outbreak may require retraining the used prediction model with its new signs and symptoms in order to consider the related posts. Thus, such approaches are not optimal solutions for new deadly flu outbreaks.

The proposed framework using FastText classifier together with the extracted features, which have not been previously used for Twitter-based flu surveillance models, aims to extract related posts faster with a comparable accuracy. Thus, it can be used for urgent cases to stop the spread of a new deadly outbreak. Improving the efficiency, along with the accuracy of text classification, is important for text-based surveillance systems for

generating early reports.

The scope of this study is to present an accurate and efficient FastText-based framework to generate Influenza trend predictions from Twitter. In addition to the typical textual features, the proposed framework utilizes the features of text sentiment analysis and the occurrences of predefined topic keywords to distinguish between flu-related tweets and unrelated ones to be passed together with historical CDC data to an estimator module for weekly flu rate predictions.

1.2 Motivation Behind the Research

Seasonal Influenza and flu can be a serious problem that may lead to death. About 250,000 to 500,000 deaths occur worldwide each year because of flu [40]. Public health care providers must be updated about the seasonal flu or any other outbreak to take the required actions for their communities. Getting an early warning will help to prevent the spread of flu in the population. Typically, health care providers take the required action to the public after getting reports of flu from the Center for Disease Control and Prevention (CDC). This center collects data from health care providers to monitor Influenza-Like Illness (ILI) and publishes the reports. This takes one to two weeks' delay, causing the required warning to come late to the provider's attention [40]. The providers need to be warned at the earliest time in order to take the appropriate actions to prevent the spread of flu. Therefore, many solutions have been proposed to provide the warning as early as possible. These include monitoring web search queries like Google Flu Trend, monitoring call volume to advice lines, and monitoring the sales of drugs and flu shots taken by patients. In addition, textual and structural data mining techniques [37] have been used to track the flu activity in Social Networking Sites (SNS). However, the literature survey shows that the existing SNS-based models include conventional techniques of post classification with maximum F -measure of 89.6%. For that reason, it is important to develop model with an

efficient post classifier that is crucial for any SNS based model.

1.3 Contributions of the Proposed Research

Since the SNS-based flu prediction models rely on post classifications, it is still a challenging task that requires more investigation for better predictions. The aim of this research is to propose a framework with an efficient classifier for better Influenza predictions using the data of Social Networking Sites and historical CDC reports as predictors. It classifies flu-related posts using important text features, such as sentiment analysis features, and then the related posts are passed together with historical CDC data to a linear regression module for better weekly flu rate predictions. The contributions of this study include the following:

- Sentiment analysis of the analyzed posts as an additional feature is considered to improve the accuracy of the classification results.
- Simple keywords related to the disease as part of the additional features are also considered to improve the accuracy of the classification results.
- The Term Frequency–Inverse Document Frequency (TF–IDF) weighting technique to weigh textual features is considered to improve the accuracy of flu tweet classifications.
- FastText classifier is fine–tuned in this work to improve the accuracy and efficiency of tweet classification. FastText cuts the required time for classification model training and testing. This is very useful for critical diseases that need immediate action such as Ebola and Corona.
- In addition, six conventional supervised classification methods are evaluated beside FastText to determine the one with better classification accuracy. The evalu-

ated classifiers include Random Forest, Naïve Bayes, SVM, C4.5 Decision Tree, K-nearest neighbors (KNN), and AdaBoost. The preprocessed labeled dataset was used to train and test the classifiers using 10–fold cross validation as the experimental setting.

- A weekly flu rate estimator based on the linear regression model is proposed. It considers a combination of predictors that includes the classification results and the historical ILI rates.

CHAPTER 2: BACKGROUND AND LITERATURE SURVEY

2.1 Introduction

The focus of this chapter is to survey the existing tools, techniques, frameworks, and methods of predicting Influenza trends in social media data. The studied methods evaluate the Twitter posts that have keywords related to Influenza for faster detection in an effort to achieve and maintain healthier communities.

This chapter is organized as follows. The Article Selection Methodology and Related Work Section first presents the method of article selection and evaluation for this review in addition to the related work. The Method Section, then, demonstrates comprehensively different methodologies and techniques of Influenza trends detection from social media data. The Discussion Section presents a discussion and comparison among all the proposed existing methodologies. Then, the Challenges Section discusses the challenges of using social media data for detection processes. Finally, concluding remarks of the literature survey are presented.

2.2 Article Selection Methodology and Related Work

This literature survey aims to review the published work in the past recent years that use social media data such as Twitter to detect Influenza. Relevant articles were collected

from various resources and publishers including IEEE, ACM, BMC, and MDPI. Different keywords were used to collect the relevant articles such as "Influenza trend prediction using social media data." During the collection process the initial number of retrieved articles was 671. The selection process was based on certain criteria such as:

- Being relevant to flu outbreak detection and prediction
- Analyzing social media data in the detection and prediction process
- Being in English Language.

Based on the selection criteria, 602 articles were excluded by reviewing the titles and the abstract of the retrieved articles. Initially, the selected articles were reviewed entirely. Out of 69 of the selected articles, 41 articles satisfied all the criteria. The final number of selected articles that were considered for this review was 27 articles. The other 14 articles were insufficient. Figure 2.1 summarizes the process of the article selection.

Several prediction and detection models that use other web data, such as Google Flu Trend (GFT), have been published in the literature for flu outbreak prediction and detection. Some of these models, such as PROFET, are included in this review to clarify that they can potentially work with the available social media data. Some other publications in the literature present flu surveillance related tools and web applications that don't use social media data for flu detections and predictions. Some of these applications and tools are listed below:

- FluNearYou (<https://flunearyou.org/>): FluNearYou [41] is a web application that uses weekly surveys to collect health status of individuals in addition to the data obtained from CDC and GFT. By using the data from the three sources, the application shows the spread of the disease in the form of maps and charts.
- Influenzanet (<http://www.Influenzanet.eu>): Influenzanet [42] is a web application that collects real-time data about flu epidemics in several European countries through

more than 30,000 contributors of Internet volunteers. Volunteers are asked to report their status weekly.

- FluOutlook (<https://fluoutlook.org/>): FluOutlook [43] is a web application that shows forecasts of the current flu season in North America and Europe in form of maps and charts. Reports are updated weekly using CDC reports. FluOutlook is based on the compartmental epidemic model.
- Columbia Prediction of Infectious Diseases (<http://cpid.iri.columbia.edu/>): Columbia Prediction of Infectious Diseases is a web application that shows forecasts of seasonal flu in curve charts. It also shows the current ILI counts in the US in a map format [44].
- HealthMap (<https://www.healthmap.org/>): HealthMap is an infectious disease monitoring system. It uses unstructured reports of the infectious diseases from multiple sources in the Internet, filters them, classifies, and visualizes information about important identified disease outbreaks [45].

2.3 Methods

There are many ways to discover knowledge and predict flu trends from Twitter data. This Section glances at various existing techniques. The studies for this review were selected to include the existing methods and techniques applied to SNS data for earlier Influenza outbreak prediction. The studied methods and techniques are within the past recent years that fall under one of the main categories of graph data mining, text mining, topic models, machine learning, math/statistical models, or mechanistic models.

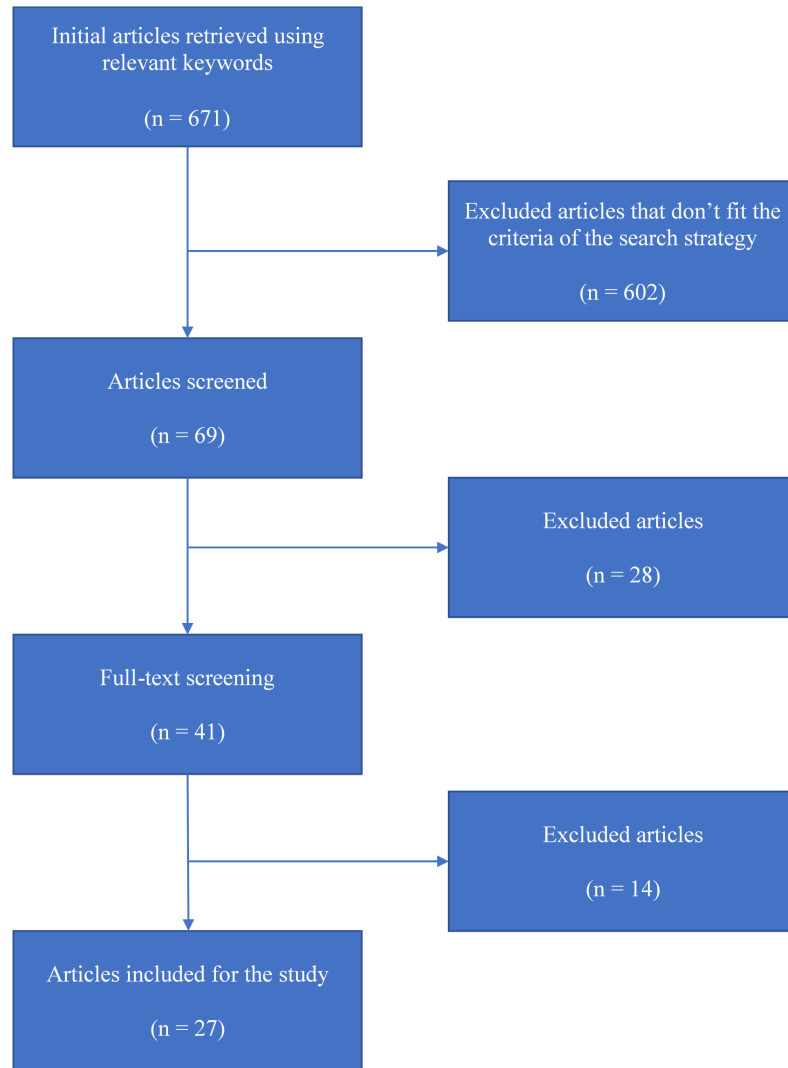


Figure 2.1: Articles selection process

2.3.1 Text Mining

Different studies show that various data mining methods can be employed to extract knowledge and detect different trends from big data such as social media data [46, 47, 1, 48, 49, 50, 51].

Text mining is a process that uses unstructured data (text) to discover intended information. Text mining techniques extract knowledge from unstructured data while data mining extracts data from structured databases. This makes it more difficult than struc-

tured data mining. Text mining can be used to discover Influenza trends from social media data [37].

2.3.1.1 Co-occurrences Analysis

Co-occurrences analysis can be used to discover how frequent certain keywords are used in a document. This analysis helps in finding related social media posts for better flu trend predictions. In addition, more analysis could be conducted using co-occurrences analysis such as medicine misuse analysis. Daniel Scamfeld et al. [52] demonstrated antibiotic misuse analysis using co-occurrences and categorization methods on social media data. Their study has also shown that social networks can be used by patients to share health information. For that reason, these kinds of networks could be used to gather knowledge to explore potential misuse of medicine. This indicates that the co-occurrences and categorization methods, along with the known flu symptoms and treatment can be used to predict flu trends in Social Networking Sites.

2.3.1.2 Historical Pattern Analysis

Since history may repeat itself, future events can be predicted using patterns of historical events such as search queries or social media posts. Kira Radinsky et al. [53] proposed a method named PROFET that predicts future news based on patterns of historical events collected from Google trends services. These services use large number of search queries.

PROFET algorithm extracts information from large number of web resources and analyzes the past events pattern in order to predict future news. It uses Google Hot Trends, which is used to obtain the important events, and Google Related Trends for the related events. It also uses Google Trends Chart to find peaks for an event. PROFET consists of several steps:

- The algorithm identifies a set of all extracted events: $W = \{w_1, w_2, \dots, w_k\}$. For

simplicity, only the important and related events are considered for further processes.

- The algorithm identifies a vector D to represent an ordered set of days: $D = \langle d_1, d_2, \dots, d_n \rangle$.
- The algorithm defines a binary vector for each event w_i : $g(w_i) = \langle d_1^i, d_2^i, \dots, d_n^i \rangle$. This vector is used to indicate that the event w_i appeared when $d_j^i = 1$. The Google Trends Chart is used to find peaks for each event w_i .
- The algorithm predicts the terms or events that may peak in k days.
- The algorithm returns a list of candidate terms with associated weights. The event with a stronger weight is the event with a higher chance of happening in the future within k days.

This algorithm together with the available social media data can help in predicting flu trends in social media. The patterns of the historical social media posts can be used as an extra parameter for any machine learning framework for better predictions.

2.3.2 Graph data mining

This technique is a process of discovering knowledge in structured data using graphical representation and graph theories. Courtney D. Corley et al. showed how graph based data mining can be used to discover flu affected communities and also to detect anomalies for better trend predictions [37].

Corley et al. [37] developed a framework based on text and graph mining. Figure 2.2 shows the general overview of their proposed framework. The framework monitors Influenza–Like Illness (ILI) mentioned in social media. It employs different data mining methods: text mining, link (graphical) mining, and structural data mining methods. The text mining method is used to identify flu trends by extracting information from large

collection of texts from social media web. The link analysis is used to find the targeted communities. A community is represented as a collection of vertices and edges (V, C). The targeted community can be identified using the Girvan–Newman algorithm (GN) that helps to identify clusters of potential communities in the studied social media [37]. The clustering process in this framework is based on content type and publisher (the first responder). The graph–based analysis technique is also used for further detection of possible anomalies (unusual occurrences) and informative substructure that could increase ILI. The results of the proposed framework show high correlation between flu–related posts and CDC weekly reports. The Girvan–Newman algorithm can be applied to any graph for the clustering process. It is composed of several steps that should be iterated to identify clusters as communities. After each iteration, the remaining components in the graph are considered as a cluster/community. Finding targeted communities using this method helps in optimizing the public health responses.

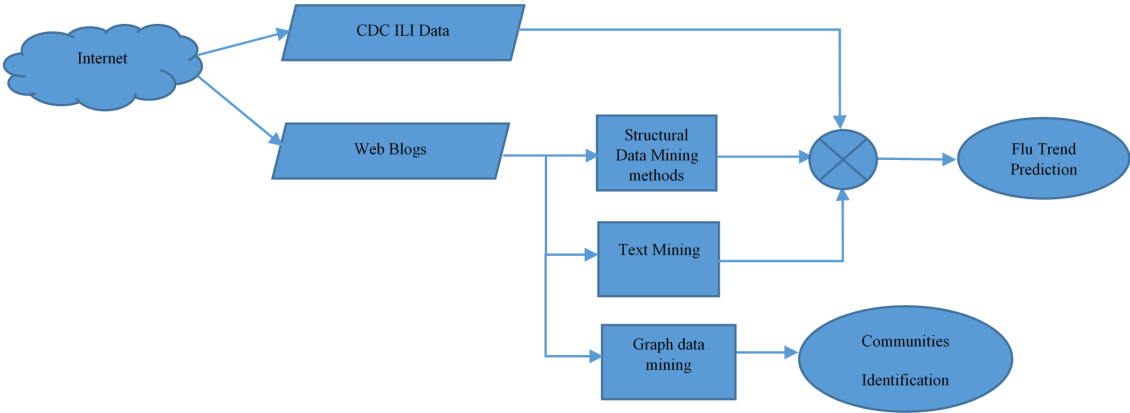


Figure 2.2: A method to monitor ILI and identify communities in Social Media

2.3.3 Topic Models

2.3.3.1 Ailment Topic Aspect Model (ATAM) and Latent Dirichlet Allocation Models

ATAM is a topic model that associates words with their hidden topics. Michael J. Paul et al. [54] showed that the ATAM model can be used to discover health topics posted by users in Twitter. The model is designed to discover more than a single disease. It is based on a probabilistic topic model called LDA (Latent Dirichlet Allocation) that associates words to hidden topics in a text such as a Twitter post and then discovers latent (hidden) structures in the data. Each hidden topic in any document is defined by a multinomial distribution over its words. Applying posterior inference (parameter learning) will return the topics with the words, which frequently co-occur with them. LDA gives topics related to disease, but it doesn't indicate a specific ailment clearly. For example, surgery could be discovered as a treatment, but LDA doesn't identify clearly whether it is for an injury or cancer. In addition to the topic model, the authors developed a structural model that uses symptoms and treatments to discover ailments.

ATAM can be used to associate symptoms, treatments, and general words with an ailment (disease). An ailment comprises of treatments, symptoms and general words. The model could associate a disease with its symptoms and treatment using Social Networking Sites. The authors use 1.6 million tweets to train the model. The model is a low cost alternative to track public health trends. The study [54] has shown that the ATAM model can discover more ailments than LDA. It produces more detailed analysis and tracks disease rate that matches the statistics published by the government (CDC).

2.3.3.2 Enhanced Topic Models (ATAM+)

Paul et al. [55] proposed a variant version of ATAM model called ATAM+. It is an enhanced model that can be used based on what can be learned from Twitter for pub-

lic health to predict specific diseases such as Influenza among other things. The model is improved by using prior knowledge, reports resulting from several new applications, correlating behavioral risk factors with ailments, and analyzing correlation of symptoms and treatments with ailments. The improved process consists of selecting 20 diseases and then collecting articles related to these diseases based on prior knowledge, and in the second step, the words in the articles were paired with the selected diseases. The results of the improved model show high quantitative correlation with government data (CDC) in detecting the flu trend using social media.

The study shows that by using ATAM+, the following could be learned from Twitter:

- **Syndromic Surveillance:** ATAM+ is able to discover and learn several aspects of public health, not only flu or just specific diseases from Twitter. The correlation between the results of the improved model and flu rate produced by CDC is high (0.958).
- **Geographical Behavioral Risk Factor:** This shows how the model can be used to mine public health information based on geographical region. In comparison with the ATAM model, it has been shown that the ailments discovered by the enhanced model (ATAM+) have higher correlation with the risk factors run by CDC. For example, the correlation between cancer and tobacco use is (0.648) using ATAM+ whereas the correlation is (0.320) using ATAM. This demonstrates that the ATAM+ outperforms ATAM.
- **Ailment Tracking over Time and Geography:** ATAM+ model can be used to mine data over time and different locations.
- **Symptoms and Medication Analysis:** The analysis of symptoms and treatment—especially for people who don't go to health care providers—needs a large population sample size. Therefore, SNS is a better alternative to perform symptoms and

treatment analysis using ATAM+. The ATAM+ is able to detect that the headache is the most common ailment treated by pain relievers. Also it shows that Tylenol is the most popular pain reliever on the market.

- Antibiotic usage Analysis: Medicine usage analysis such as antibiotic misuse could be performed using ATAM+.

2.3.3.3 Hidden Flu–State from Tweet Model – HFSTM (Users Health States Transition for Better Prediction)

Liangzhe Chen et al. [56] proposed a model called Hidden Flu–State from Tweet Model (HFSTM) that is able to capture hidden health states of users and the associated transitions by analyzing their tweet posts. The extracted states are used to obtain a better prediction of trends. It aggregates the states of the users in a specific geographical region for better prediction. The proposed model captures not only one tweet post, but also streams of tweet posts of users in order to capture their underlining health status (different health states from tweet posts). The used states for this study are: S (healthy), E (Exposed), I (Infected), and R (Recovered with Immunity).

Most of the other models are coarse–grained because they don't give any understanding of how health states change over time. This model links the social activity models and the epidemiological models. This linkage improves the prediction process. The most common Contagion–based epidemiological models are SI, SIR, SEIS. These models are used here to predict the true flu cases by tracking the health states of a person through the lifecycle of the infection.

Unlike the proposed model, the existing topic models (LDA, ATAM+, Makovian, and non-Markov) do not solve the problem of flu state changing. The model uses unsupervised topic modeling that can capture the transition (changes) between consecutive messages of a user.

The study [56] has shown that the HFSTM model can learn meaningful word distribution. Each word in the list belongs to one of the three states (S, E, I). It can also learn the state transition as shown in Figure 2.3. The HFSTM model is able to classify the state of tweets and captures the transitions. It is also capable of predicting flu trends. The results of HFSTM model were compared to the Pan American Health Organization (PAHO) weekly records and the results of other two models: Google Flu Trend (GFT) and the baseline model that is based on word count and linear regression. GFT is a Flu trend prediction system that uses the volume of flu related search queries for the prediction process. Many studies have been conducted to evaluate and improve GFT [57, 58, 59, 60, 61, 62]. The study has shown that the HFSTM model is better than the baseline model and is comparable with GFT. In some cases, HFSTM outperforms GFT. Results have shown that GFT overestimates the number of flu cases.

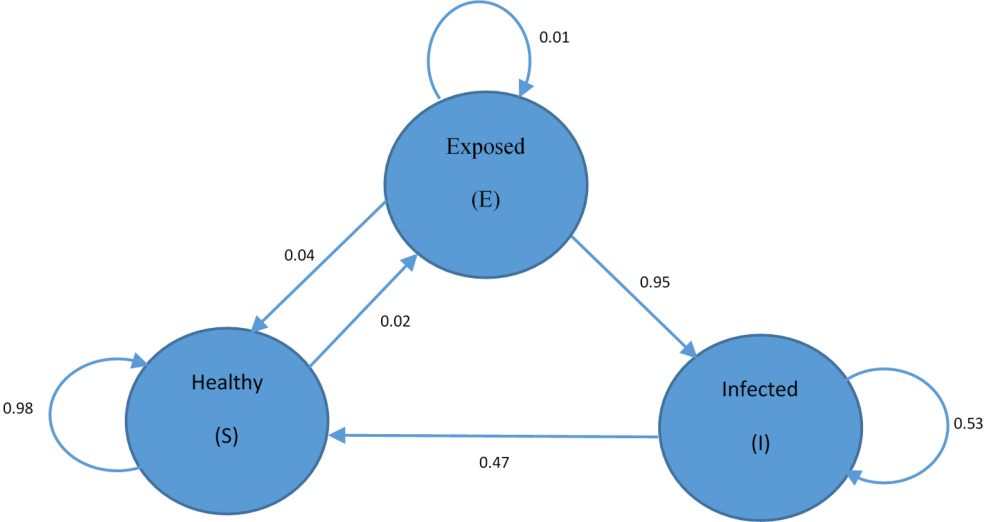


Figure 2.3: Health state transition diagram

2.3.4 Machine Learning Techniques

2.3.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method. Based on our survey, SVM is the most commonly used machine learning algorithm for the purpose of flu related posts classifications [63, 64, 65, 66, 67].

David A. Broniatowski, et al. [65] proposed a model that consists of three levels of classification using SVM for better distinction between the actual tweets about flu and the tweets that seem related but are not actually flu tweets (named "chatter" posts). The first classifiers is used to classify the collected posts to health-related/unrelated posts. The second one is used to extract the flu related posts, and the third one is used for infection classifications. The proposed algorithm was tested using a collection of tweets from Sep. 30, 2012 to May 31, 2013 (covering the season flu of 2012–2013) for the NYC location and the USA in general (local and national). To measure the performance, the results of the proposed algorithm was observed to have correlated with the CDC data ($r = 0.93$) and also with the data of the Department of Health and Mental Hygiene of New York City ($r = 0.88$).

It has been shown that the distinction between the infection and awareness tweets enhances the accuracy of the results. The goal of this distinction is to consider the infection posts only. Alex Lamb, et al. [68] proposed a machine learning based model that consists of two phases of classification to differentiate between the infection and awareness tweets. The accuracy of the model showed high correlation with CDC data using Pearson Correlation ($r = 0.9897$).

Eiji Aramaki, et al. [64] proposed a framework that consists of two parts. First, a crawler that works together with Twitter API to collect tweets was used, and then they were filtered for only flu-related ones. Second, an SVM-based classifier was used to extract only the actual Influenza tweets (positive tweets) and exclude the unrelated ones such as

news and questions (negative tweets). The initial dataset for this study was collected from Nov 2008 to June 2010. It included 300 million general tweets. Then, this dataset was filtered using "Influenza" keyword to get a set of only flu related tweets which contained 400,000 tweets. The flu-related dataset was divided into two parts: a training dataset, which contained 5,000 tweets (November 2008) and a test dataset, which contained all the remaining tweets from Dec 2008 to June 2010. The training dataset was assigned to a human annotator to label each tweet as either positive or negative. A tweet is labeled positive if it met two conditions. First, the flu tweet should concern the person who posted the tweet or about another person in a nearby area (maximum an area of the city). If the distance is unknown, the tweet is considered negative. Second, the flu tweet should be an affirmative sentence in the present tense or past tense with maximum period of 24 hours which can be checked using specific keywords such as "yesterday". The SVM classifier was implemented using the Bag-of-Words feature representation. The authors compared the accuracy of the SVM-based classifier with other six different machine learning methods and they found that the SVM was the most accurate method. For the purpose of evaluation, a Pearson Correlation was used to correlate between the results of this framework and the Japanese government data provided by the Infection Disease Surveillance Center (IDSC). The results of this framework showed high correlation ($r = 0.89$). The results also showed that news could impact the accuracy of the results. It has been shown that the swine flu related news in 2009 led to poor performance of this method and other methods.

José Carlos Santos, et al. [67] also applied SVM-based classifier to detect flu-like illness in Portugal using Twitter posts. For the purpose of training and testing, a dataset with 2,704 posts was manually annotated with 650 textual features. A subset of the annotated dataset was used to train the classifier. The classified tweets together with search queries were applied to a regression model as predictors. The results of the used model was evaluated and compared with the reports provided by Influenzanet: a system that mon-

itors Influenza Like Illness activities in Europe. The highest correlation ratio between the results of this method and Influenzanet data is 0.89 ($r = 0.89$). The classifier was implemented using the Bag-of-Words feature representation, and the feature selection process was based on a Mutual Information (MI) value that is used to pick the best set of features. Each feature is applied to a true class, and then MI value is assigned to the feature. The value of MI is based on how the feature is related to the true class. A feature with high MI value is more related to the true class.

Nanhai Yang, et al. [66] proposed a SVM-based method to predict flu trends from Chinese Social Networking Sites in Beijing. The authors claim that this is the first study to predict flu trend from Chinese Social Networking Sites. The collected data for this study included 3,505,110 posts from Sep. 2013 to Dec. 2013. Among those, 5,000 random posts were selected for manual annotation (sick and not sick labels) to be used for training and testing purposes-285 of sick posts and 285 of not sick posts were picked for training. For higher accuracy, word based features were used instead of character based features. Among the four types of word weighting techniques: Boolean weighting, term frequency weighting (TF), inverted document frequency weighting (IDF) and term frequency-inverted document frequency weighting (TFIDF), the TFIDF method was considered for classification purposes. Different classifiers were compared to decide the best one for the problem. The authors found that SVM was the best for big data problems. This method was able to predict the flu trend five days earlier than the China Nation Influenza Center (CNIC).

Mauricio Santillana, et al. [63] proposed a machine learning-based method that was capable of predicting flu related activities. In addition to CDC ILI reports that have been used as the ground truth, the method used data from different sources for better results. The sources included Google searches, Google Flu Trends, Twitter posts, hospital visits records collected from AthenaHealth, and a surveillance system called FluNearYou. This study has shown that the results of prediction methods using combined data sources outperform the

results when using a single data source. The method utilizes well-known machine learning algorithms including support vector machine, stacked linear regression and AdaBoost with decision trees regression. The study has also shown that the three algorithms work perfectly together in combining the information from different sources for real time analysis and then better forecasting. It has been shown that this method can predict one week faster than the Google Flu Trend (GFT) with accurate and comparable results.

2.3.4.2 Neural Network

Vasileios Lampos et al. [69] proposed a method to track flu in the population using Social Networking Sites. The method analyzed flu-related and flu-symptoms-related keywords in Twitter. The extracted information was converted to flu-score using machine learning techniques. Computing the flu score from Twitter includes several steps. First, a set of selected keywords M is identified to represent the search keywords to look for in Twitter posts: m_i ; where $i \in [1, k]$. Second, a set of daily tweets is identified as $\tau = t_j$ where $j \in [1, n]$. When the marker m_i appears in the tweet t_j : $m_i(t_j) = 1$, otherwise $m_i(t_j) = 0$. The number of markers appeared in t_j divided by the total number of markers is denoted as $s(t_j)$ and calculated using Equation 2.1.

$$S(t_j) = \frac{\sum_i m_i(t_j)}{k} \quad (2.1)$$

The flu-score of the daily tweet corpus $f(\tau, M)$ equals to the sum of all the flu-score of the tweets $s(t_j)$ of that day divided by the total number of the tweets n (Equation 2.2).

$$f(\tau, M) = \frac{\sum_j s(t_j)}{n} = \frac{\sum_j \sum_i m_i(t_j)}{k \times n} \quad (2.2)$$

An extension was made to the previous model in order to make a better prediction of Health Protection Agency (HPA) flu rate by adding weight w_i to each marker m_i (Equation 2.3). Therefore, the weighted flu-score for each tweet is:

$$S_w(t_j) = \frac{\sum_i w_i \times m_i(t_j)}{k} \quad (2.3)$$

Then, the weighted flu scores of all tweets of a day is summed up to get the weighted flu-score of the daily tweet corpus $f_w(\tau, M)$ (Equation 2.4):

$$f_w(\tau, M) = \frac{\sum_j S_w(t_j)}{n} = \frac{\sum_j \sum_i w_i \times m_i(t_j)}{k \times n} \quad (2.4)$$

The contribution of the marker m_i in the daily tweet flu-score f_w is considered as flu-subscore $f_{w_i}(\tau, m_i)$ (Equation 2.5):

$$f_{w_i}(\tau, m_i) = w_i \times \frac{\sum_j m_i(t_j)}{k \times n} \quad (2.5)$$

Using the flu-subscore $f_{w_i}(\tau, m_i)$, the daily tweet flu-score (Equation 2.6) could be represented as a vector of flu-subscore F_w of all the markers (keywords):

$$F_w = [f_{w_1}(\tau, m_1), \dots, f_{w_k}(\tau, m_k)]^T \quad (2.6)$$

The weights w_i of markers m_i can be learned by:

- (1) Initially, the unweighted flu-score vector F_w that is the sum of unweighted flu-subscore smoothed with 7-point moving average is found (Equation 2.7).

$$F = [f(\tau, m_1), \dots, f(\tau, m_k)]^T \quad (2.7)$$

- (2) The least square linear regression between F from the smoothed version, F from the expanded one, and smoothed HPA flu rate is performed.

To maximize the correlation with HPA flu rate, Vasileios Lamos et al. [69] also proposed a method to extract the markers (keywords) automatically. This method consisted of two steps. First, a list of candidates was created by extracting them from trusted web documents related to Influenza. Second, the most informative ones were picked using

the Least Absolute Shrinkage and Selection Operator (LASSO) method that discards the redundant features of the candidates. The use of LASSO method is explained in detail in [69].

Another machine learning technique that can be used in early trend prediction is neural network. Disease outbreaks can be predicted using Neural Network (NN) based approaches to analyze web data. Wei Xu et al. [70] proposed a model to detect Influenza outbreaks by analyzing web search queries using a neural network approach. Figures 2.4 and 2.5 show an overview of their proposed approach. This approach consists of several steps. The first step is to collect data from search engine queries and ILI data from the CDC. The second step is to select features automatically by reducing the dimension of the query and keeping only the most important features. The third step is to find the relationship between the Influenza Like Illness (ILI) and web data (query data) using different NN with different algorithms and architectures to measure the fitness values. The NN used with this model are: NN-GDX (Gradient descent with momentum and adaptive learning rate back propagation), NN-OSS (One-step secant back propagation), and NN-RP (Resilient back propagation). The 10-fold cross validation method is used to validate the different NN algorithms. The fourth step is to select the best NN as a detector using the cross validation method. The fifth step is to use the selected NN (detector) with the best features subset to predict flu activities. The accuracy (ACC) of the results of each NN is measured using Equation 2.8. If A_i are the actual values, D_i the detection values, and N the number of given pairs (A_i, D_i) , then

$$ACC = \frac{1}{N} \sum_{i=1}^N \frac{D_i}{A_i} \quad (2.8)$$

Results show that NN-RP was the best to be used for Influenza detection. NN-RP had the best average of ACC values.

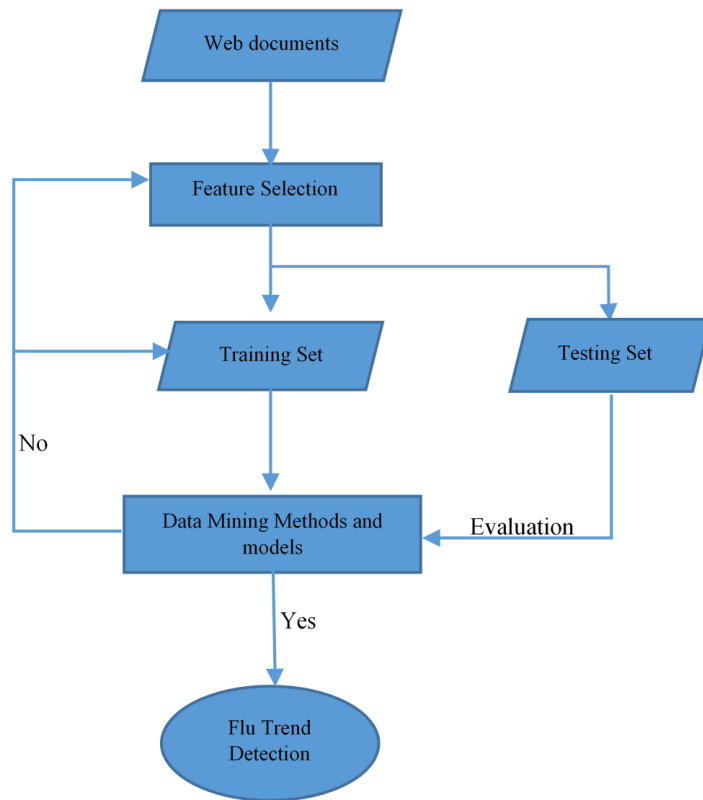


Figure 2.4: A framework for Influenza outbreak detection

2.3.4.3 Naïve Bayes

Kenny Byrd, et al. [71] proposed a framework based on Naïve Bayes classifier. The framework consisted of several steps. The first step was tweets collection with a location filter. The collected tweets were from Oct. 27 to Nov. 30 of 2015. The dataset included a total of 1,848,130 tweets. The used location filter was provided as latitudes and longitudes pairs (a comma separated list) to specify a bounding box of a required area. The Google Maps Developer tool was used to determine the bounding boxes of the required areas (cities). For this study, the used location was the area of Ottawa and its surrounding

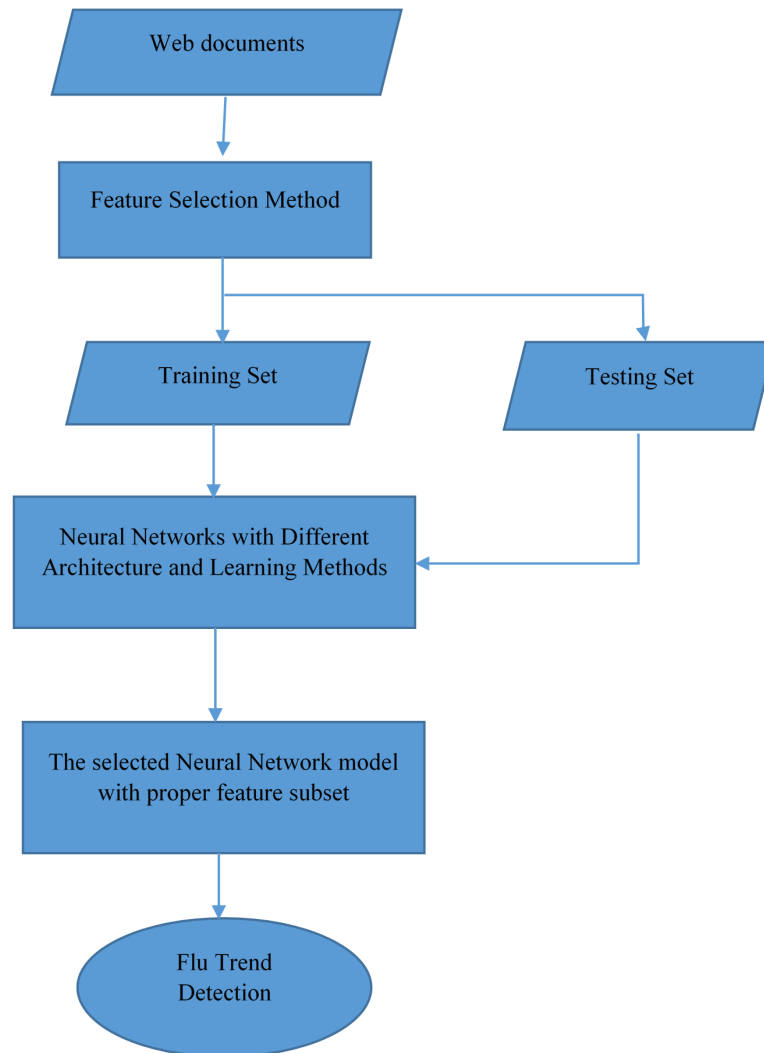


Figure 2.5: The process of Neural Networks based detection

areas. The second step was flu-related tweets filtration. The used keywords for the filtration process were "sick", "flu" and "cough". The total of filtered tweets were 4,696 posts. The third step was pre-processing which included: stop words elimination, URL's removing, words stemming, and retweets removing. The fourth step was sentiment analysis by applying machine learning techniques for classification (positive, negative, neutral). Three machine learning algorithms were evaluated, and this study found that the highest accuracy method was the Naïve Bayes classifier. The Naïve Bayes classifier was implemented us-

ing the Stanform core NLP (Natural Language Processing) and trained using the OpenNLP training dataset which includes 100 annotated tweets. The sentiment analysis is considered accurate when there is a matching between the predicted sentiment polarity with the manual assigned opinion of the sentiment. The authors found that Naïve Bayes was the most accurate one with 70% matching.

2.3.4.4 Prediction Market Using Support Vector Machine Regression Algorithm (SVR)

The prediction market is a mechanism that can be used for future prediction based on creating *shares* for an event. People can trade these shares with prices determined by the market. The prices can be used as probability of the event occurrence. This is considered as one of the optimal prediction solutions, and it is less expensive than other prediction methods. Disease outbreak can be predicted using the prediction market together with the Support Vector Machine regression algorithm (SVR) using share prices [72]. Joshua Ritterman et al. [72] have shown that the prediction of swine flu in 2009 was more accurate when adding some features extracted from Social Networking Sites to the SVR. The prediction market is modeled in two different ways: internal market and external market.

Internal Market The internal market is based on time series. It uses historical prices for today's price prediction. Technically, the prediction for a given day F_n is achieved by using the average price of the previous day $AvgP_{n-1}$ divided by the sum of the average prices for the previous 5 days (Equation 2.9).

$$F_n = \frac{AvgP_{n-1}}{\sum_{i=2}^6 AvgP_{n-i}} \quad (2.9)$$

The SVR is trained using extra features. The first feature is to use the Short-Term history feature $F(n) = AvgP(n - 1)$ that is the average price of the previous day. It gives a quick overview of the price movement. The second feature is the Mid-Term history feature

that is the moving average price, calculated using Equation 2.9. This determines a longer period than the first feature. The third extra feature is the Long-Term feature that is the sum of a vector of binary values M , as shown in Equation 2.10. The Long-Term feature is used to indicate the market direction for a long time.

$$F(n) = \sum_{i=0}^{n-1} M_i, M_i = \begin{cases} M_{i-1} + 1 & \text{if } Avg(P_i) \geq Avg(P_{i-1}) \\ M_{i-1} - 1 & \text{if } Avg(P_i) < Avg(P_{i-1}) \end{cases} \quad (2.10)$$

External Market This way of modeling considers the fundamental products of the company and the events occurring around the world. The SVR classifier is trained using social media data. By using the social media data, SVR is trained with unigram and bigram features and their frequencies using social media data (i.e. daily counts of unigrams and bigrams). No internal market is given for training. This gave lower performance compared to training with only a subset of data. For better performance, the system should be trained with only relevant data. This can be accomplished by training the SVR with unigrams and bigrams for a specific period of time based on historical context provided to the system. The length of the period is decided by the system using the historical context to determine the news cycle.

Joshua Ritterman et al. [72] have shown that combining the prediction market with features extracted from Social Networking Sites leads to better results. This demonstrates that social media data played an important role in the 2009 swine flu trend prediction.

2.3.5 Math/Statistical Based Models

2.3.5.1 Autocorrelation Function (ACF)

ACF finds the correlation of the values of the same variables at different times $(x_i, x_{(i+1)})$. Therefore, this method can be used for disease outbreak predictions. Disease outbreak trends in Social Networking Sites can be monitored by tracking a sudden high

frequency of disease–content posts using ACF. It compares the averaged disease–related posts per day with the actual number of the same disease posts of that day. Courtney D Corley et al. [73] proposed a method to track ILI in social media using ACF and to identify possible web and social media communities [73]. This method tracks a sudden high frequency of flu–content posts using ACF. The method defines a seven-day period as a period cycle for better accuracy and anomaly detection. The period starts on Sundays and ends on Saturdays.

The results of this methodology showed strong correlation with CDC reports. The Pearson Correlation coefficient is used for evaluation. The value of r was 0.767 with a confidence level of 95%.

Web Social Media (WSM) community identification and analysis was used as a part of their methodology for better results by using link analysis. Link analysis was also used to identify the first responder or influential user of a community. Only the links between flu posts are considered. The links between a flu–related post and non-flu-related post are not considered in the defined community. Closeness, Betweenness and Page Rank measures were used to rank flu communities to tell how a blog’s influence disseminates flu information. Blogs with high closeness and page rank can spread flu–information (response) more quickly.

Closeness It is used to find the average of the shortest paths between actor v and the other reachable actors. It is defined as shown in Equation 2.11 [74]. Let i and j be actors, $d(i, j)$ be the distance function that finds the number of geodesics between i and j , and $\sum_{j=1}^N d(i, j)$ be the total distance of i from all other actors. Closeness is defined as follows:

$$C_c(i) = \left[\sum_{j=1}^N d(i, j) \right]^{-1} \quad (2.11)$$

Betweenness It measures how a blog is central among other blogs. It is defined as shown in Equation 2.12 [74]. Let g_{jk} be the number of geodesics between j and k , and $g_{jk}(i)$ be the number of geodesics between j and k that contain actor i . Betweenness is defined by the following formula:

$$C_B(i) = \sum_{j < k} \frac{g_{jk}(i)}{g_{jk}} \quad (2.12)$$

Page Rank It is an eigenvector centrality that measures the importance of a node. It is defined as shown in Equation 2.13 [73]. Let $d = 0.85$ be a factor, where the pages are represented using the symbol P_n , the set of pages linked to P_n is represented using $M(p_n)$, and the out links on page P_j is represented using $L(p_j)$. Page Rank relationship is shown as follows:

$$R_{p_n} = \frac{1-d}{N} + d \sum_{p_j \in M(p_n)} \frac{PR(p_j)}{L(p_j)} \quad (2.13)$$

2.3.5.2 Auto Regression Moving Average (ARMA)/SNEFT

ARMA is a stochastic model that is composed of two forms: Auto Regression (AR) model and Moving Average (MA) model. The AR model is a prediction model. Its output depends linearly on the past values, a random value as an error, and a constant value. The MA model is used to represent the correlation between the past values and the white noise using linear regression.

Based on the ARMA model, Harshvardhan Achrekar et al. [40] proposed a framework called Social Network Enabled Flu Trends (SNEFT) that utilizes the ARMA model and the data obtained from CDC. Both are used in collaboration for better flu prediction trends. The architecture of the SNEFT framework is shown in Figure 2.6. The architecture consists of two main parts. The first part is used to predict Influenza and Influenza Like

Illness (ILI) using CDC data. The second part is used to provide flu warnings using Twitter data. The Auto regression Moving Average (ARMA) model is used to predict ILI incidence as a linear function of current and old Social Network data and historical ILI data (CDC data). The results indicated that Twitter data improved the output of the statistical models that were used for prediction. The SNEFT framework was tested with and without Twitter data together with CDC reports. The study has found that the Twitter data improved the accuracy of the prediction model. Based on the authors' findings, it is clear that Twitter could provide real time measurement of Influenza activity in the population.

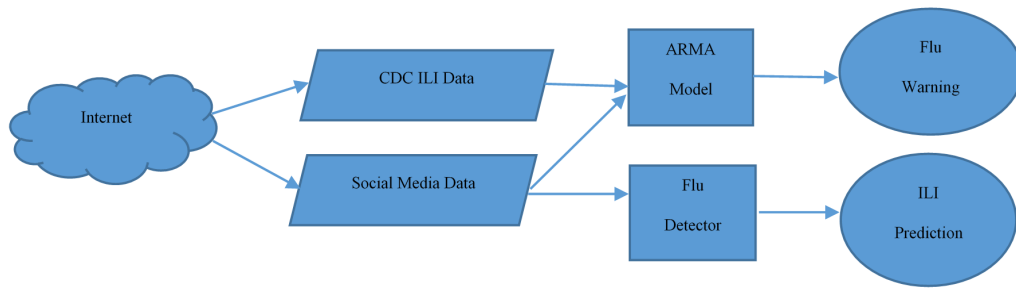


Figure 2.6: SNEFT architecture

2.3.5.3 Numerical-Based Analysis

Sangeeta Grover, et al. [75] proposed a framework to detect flu outbreak with respect to three stages of epidemics (beginning of epidemic, spread of epidemic, absence of epidemic) using the Bag-Of-Words (BOW) technique. The BOW is a technique that learns a vocabulary from all the documents, then models each document by counting the number of times each word appears. The implementation of this framework consists of the following steps:

- Collect tweets using Twitter API.
- Store the collected tweets in MangoDB.

- Build Bag–Of–Words (BOW) for each stage of epidemic (beginning of epidemic, spread of epidemic, absence of epidemic)
- Apply the Swine Epidemic Hint Algorithm (SEHA) on the tweets. The text of a tweet is tokenized for numerical analysis. The numerical analysis checks how relevant the tweet is to the epidemic stages.
- Classify the tweets into the 3 stages of the epidemics. The classification process is based on the numerical results from the previous step.
- Evaluate the results of this framework using 6 cross validation of Gaussian regression and prediction model. The results show that the framework was fairly accurate since the average value of the error rate was about 1.1.

2.3.6 Mechanistic disease models

Mechanistic disease models are used to provide a better understanding of any epidemic dynamics. Unlike statistical models, the mechanistic models consider different features to estimate key epidemic parameters such as intensity and severity that impact public health decision responses [76, 77]. Within the various mechanistic models, metapopulation models, compartmental models, and agent–based models provide information on population epidemic states and individual progress of an epidemic.

2.3.6.1 Metapopulation models

Metapopulation models, such as Global Epidemic and Mobility (GLEAM) model, are spatial, stochastic, and individual based models that can simulate the spread of epidemic diseases at worldwide scale. The model divides the world into smaller regions defining subpopulation networks and connections between the subpopulation that represent the individual fluxes because of the transportation and mobility infrastructure [78].

Qian Zhang, et al. [77] proposed a seasonal flu forecasting framework based on mechanistic disease model (GLEAM). The framework was validated and tested by comparing the results from the framework with the official government data in the U.S., Italy, and Spain in the 2014–2015 season and 2015–2016 season. The framework is a combination of the social media data, official surveillance data and mechanistic modeling approach. It consists of three stages. In the first stage, data from official surveillance systems and Twitter is used for model initialization. A set of English ILI-related tweets for a given region is used as an initial condition of relative flu incidences and as an input for the framework. The data from official surveillance systems is used to evaluate the coefficient of determination of the used ILI search keywords. The second stage consists of exploring important parameters: population, infectious period and the effective reproduction number (number of infected individuals in a region). The third stage is parameter selection and prediction. The study has shown that the framework provides reliable results for epidemic intensity and peak timing up to six weeks in advance. The accuracy of the framework showed high correlation with official surveillance data using Pearson Correlation (the highest r value is 0.98 for the flu prediction with one week in advance).

2.3.6.2 Compartmental models

Compartmental models define the rate at which individuals move between defined compartments and divide the population into subpopulation based on disease states. Examples include Susceptible- Infectious-Recovered (SIR) and Susceptible-Infections-Recovered-Susceptible (SIRS) [79].

Liangzhe Chen et al. [56] proposed a model called Hidden Flu-State from Tweet Model (HFSTM) based on the concept of epidemiological compartmental models. It analyzes a stream of a user's tweets and captures the disease states and the associated transitions.

Jeffrey Shaman, et al. [44] proposed a framework that predicts a seasonal flu using the compartmental model (SIRS) along with common used techniques in numerical weather predictions. Epidemic disease dynamics are non-linear that are similar to weather dynamics. The non-linearity of the epidemics makes the prediction systems sensitive to the initial and current conditions. Like any non-linear system, it is possible that the error rate of the system will grow with further uses that leads to inaccurate results. To overcome the growth of error rates with the non-linear systems, data assimilation techniques such as filtering are used to update and adjust the system using the latest available observations. The applied data assimilation method in the presented framework is the Ensemble Adjustment Kalman Filter (EAKF) method for the updating process using weekly observations obtained from Google Flu Trend (GFT). This method combines the weekly GFT observations with the Susceptible Infections Recovered Susceptible (SIRS) model. The EAKF is a recursive filtering technique to estimate the state of the model using a combination of the observations and the evolving ensemble of the model simulations. The framework was validated and then used to perform simulation of Influenza prediction in the New York City for the 2004–2005 and 2007–2008 flu seasons. The study has shown that the proposed framework is able to predict the peak timing up to seven weeks in advance.

2.3.6.3 Agent-Based models

Agent-based models define entities (agents) that interact with each other and the surrounding environment based on specific rules. These models provide better understanding of the change of individual behaviors during an epidemic which help in outbreak predictions [79].

Suruchi Deodhar, et al. [80] developed a large scale web application called FluCaster for flu epidemic forecasting using agent-based models. This model can distinguish FluCaster from other available systems. It produces fine-grained results that helps decision

makers in performing detailed analysis. For example, filtering the results of the flu forecast by a specific location for a specific age sub-population in a specific time can be provided by this model. FluCaster was implemented using CDC surveillance data and Google Flu Trend (GFT).

2.3.7 Detection Based on Filtered Keywords and Documents

Simple flu related keywords can be used to produce accurate results with a high correlation with CDC weekly reports. The method of selecting search keywords is very important. It impacts the accuracy of the results. Selecting keywords based on correlation with national statistics may cause inaccurate results. For example, the "flu shot" term has a high correlation, but it does not necessarily reflect the spread of flu. It could be just a general discussion about it or an advertisement. Therefore, a document classifier to remove spurious matches (such as advertisements) can be used to get more accurate results and reduce the error rates [38]. Aron Culotta [38] presented a method of correlating the keywords with ILI rates from CDC. Let P be the ILI symptoms reported by providers, $W = \{w_1, w_2, \dots, w_k\}$ be the set of keywords, D be a document collection, D_w be a set of documents that at least contain a keyword in W , B_1 and B_2 be coefficients, e be error terms, and $Q(w, D) = |D_w|/|D|$ be a query fraction, then

$$\log(P) = B_1(\log(Q(w, D))) + B_2 + e \quad (2.14)$$

Removing spurious keywords such as a keyword within government announcements and advertisements may also help produce better results and improve the correlation with ILI reports. Aron Culotta [38] also proposed a document classifier that can be used for document filtration. It labels the messages as ILI related or not. Then, the classifier calculates the probability of the ILI reporting messages. This classifier should be trained using logistic regression (Equation 2.15) with a parameter θ that can be computed using

the limited memory quasi-Newton method for large scale optimization (L-BFGS). Details of the L-BFGS method and its implementation are discussed in [81]. Let y_i be a binary random variable where (1) is a positive document and (0) otherwise, $x_i = \{x_{ij}\}$ be a vector of random values where x_{ij} is the number of times word j appears in document i , D be a document collection, θ can be computed using L-BFGS gradient descent [81]

$$P(y_i = 1|x_i; \theta) = \frac{1}{1 + e^{(-x_i \cdot \theta)}} \quad (2.15)$$

The filtration process was combined with regression in Equation 2.14 by considering two kinds of classifying methods: soft classification and hard classification. The soft classification finds $Q_s(W, D)$ of positive documents using Equation 2.16. This method assigns the probability as a weight to each matched document in D_w . The hard classification finds $Q_h(W, D)$ by considering and counting only the documents with probability of positive class > 0.5 using Equation 2.17. Afterwards, the value $Q(w, D)$ is substituted in Equation 2.14.

$$Q_s(W, D) = \frac{\sum_{d_i \in D_w} P(y_i = 1|x_i; \theta)}{|D|} \quad (2.16)$$

$$Q_h(W, D) = \frac{\sum_{d_i \in D_w} (P(y_i = 1|x_i; \theta) > 0.5)}{|D|} \quad (2.17)$$

The results show strong correlation for most of the picked keywords (e.g. flu, cough, sore throat, and headache). Comparing the results with another study's results by Lampose and Christianini (2010) [69] has shown that the results are competitive and yield less complexity. This concludes that flu trends could be predicted in a population by using simple methods.

2.4 Discussion

A summary of the used data sets in the reviewed studies is shown in Table 2.1. The performance of the discussed methods is shown in Table 2.2. Most studies use Pearson Correlation and Root Mean Squared Error (RMSE) for performance measurement. Therefore, in Table 2.2, the Person correlation measure is included for comparison.

Pearson Correlation is a metric that evaluates the correlation between two datasets using the symbol r . It ranges between (1) and (-1): the value of $r = 1$ when both datasets exactly match and the value of $r = 0$ when there is no correlation between the two datasets. Let y_i be the observed value of the ground truth (CDC ILINet data), x_i be the predicted value by a proposed model, and \bar{y} and \bar{x} be the average values of $\{y_i\}$ and $\{x_i\}$, respectively. Using these notations, Pearson Correlation value r is defined as shown in Equation 2.18 [63].

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (2.18)$$

Root Mean Squared Error (*RMSE*) is an evaluation metric that provides an indicator of comparison between predicted and real values. Lower value of *RMSE* indicates more accurate results of the used model and less errors. Using the same notations for Pearson Correlation, the *RMSE* value is defined as shown in Equation 2.19 [63].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2} \quad (2.19)$$

As shown in Table 2.2, the SNEFT yields a very high correlation coefficient with the used ground truth (0.9846). The study [40] has shown that the best results is obtained when the dataset is filtered to not include redundant posts (retweet) as well as posts from the same user within one week. In addition, the authors use Root Mean Squared Error (RMSE) to evaluate the accuracy of SNEFT. It has been found that the value of RMSE of the same filtered dataset is 0.318. Further enhancement of the accuracy can be achieved

by considering only the tweets about infection as shown in [68]. The distinction between the infection and awareness tweets shows high correlation with CDC data using Pearson Correlation ($r = 0.9897$). The other methods were evaluated using different measures. The neural network approach was evaluated by comparing the accuracy of different neural network algorithms using the *ACC* measure which is calculated using Equation 2.8. The study [70] has shown that the best average value of *ACC* is 0.9532. The HFSTM model was evaluated by comparing it with the Google Flu Trend (GFT). The study [56] has shown that the HFSTM model outperforms the GFT even with no optimization. The evaluation of the prediction market was conducted using Mean Square Error (MSE) measure. The study [72] has shown that the MSE was lowered dramatically when using historical context with the bigram model. The best value of MSE is 40.67. For the Historical pattern method, the study [53] has shown that the precision for 1–day prediction is 0.8 (with mean of 0.52) and 0.6 (with mean of 0.46) for 7–days prediction. The Journal/conference backgrounds of the reviewed studies are listed in Table 2.3

Table 2.1: Summary of the used data sets in the reviewed studies

Method Category	Method Name	Reference	SNS	Language	Timeframe	Location
Graph Data Mining	Graph Data Mining	[37]	Twitter	English	Oct 2008 – March 2009	US
Text Mining	Historical Patterns Co-occurrences	[53]	Twitter	English	March 2009 – Jul 2009	US
		[52]				
Topic Models	ATAM	[54]	Twitter	English	May 2009 – Oct 2010	US
	ATAM+	[55]	Twitter	English	May 2009 – Oct 2010	US
	HFSTM	[56]	Twitter	English	Dec 2012 – Jan 2014	South America
Machine Learning	Neural Network	[69]	Twitter	English	Jun 2009 – Dec 2009	UK
		[65]	Twitter	English	Sep 2012 – May 2013	US
		[64]	Twitter	English	Nov 2008 – Jun 2010	Japan
		[67]	Twitter	Portuguese	March 2010 – Feb 2012	Portugal
		[66]	Chinese Sina	Chinese	Sep 2013 – Dec 2013	China
		[68]	Twitter	English	May 2009 – Oct 2010	US
[63]	Twitter	English	Nov 2011 – Feb 2015	US		
Math/Statistical Models	Prediction Market using SVR Naïve Bayes	[72]	Twitter	English	April 2009 – Jun 2009	Ottawa
		[71]	Twitter	English	Oct 2015 – Nov 2015	
		[40]	Twitter	English	Oct 2009 – Oct 2010	
Mechanistic Disease Models	Numerical-Based Analysis (SEHA using BOW)	[73]	Twitter	English	Aug 2008 – Sep 2008	US, Spain, Italy South America
		[75]	Twitter	English	Oct 2009 – Oct 2010	
		[77]	Twitter	English	2014–2015, 2015–2016	
Keys/Documents Filtration	Metpopulation Model Compartmental Model Agent-Based Model	[44]	Twitter	English	Dec 2012 – Jan 2014	US, Spain, Italy South America
		[80]	Twitter	English	2014–2015, 2015–2016	
		[38]	Twitter	English	Sep 2009 – May 2010	

Table 2.2: Summary of the reviewed methods and techniques

Method Category	Method Name	Study Reference	Performance Metric	Metric Value
Graph Data Mining	Graph Data Mining	[37]	Pearson Correlation	r=0.545
Text Mining	Historical Patterns	[53]	The precision for 1-day prediction is 0.8 (with mean of 0.52) and 0.6 (with mean of 0.46) for 7-days prediction.	
	Co-occurrences	[52]		
Topic Models	ATAM	[54]	Pearson Correlation	r=0.934
	ATAM+	[55]	Pearson Correlation	r=0.958
	HFSTM	[56]	Mean Square Error (MSE)	MSE = 40.67
Machine Learning	Neural Network	[69]	ACC (Equation 2.8)	ACC=0.9532
		[65]	Pearson Correlation	r=0.93
	SVM	[64]	Pearson Correlation	r=0.89
		[67]	Pearson Correlation	r=0.89
		[66]		
		[68]	Pearson Correlation	r=0.9897
[63]				
Prediction Market using SVR	Naïve Bayes	[72]	Sentiment polarity is used to determine the accuracy of the used method (Naïve Bayes polarity is 70%)	
		[71]		
Math/Statistical Based Models	SNEFT	[40]	Pearson Correlation	r=0.9846
	ACF	[73]	Pearson Correlation	r=0.767
	Numerical-Based Analysis (SEHA using BOW)	[75]	RMSE	Avg (RMSE)=1.1
Mechanistic Disease Models	Metpopulation Model	[77]	Pearson Correlation	r=0.98
	Compartmental Model	[44]		
	Agent-Based Model	[80]		
Keys/Documents Filtration	Keys/Documents Filtration	[38]		

Table 2.3: Journal/Conference backgrounds of the reviewed studies

Method Category	Method Name	Study Reference	Journal/Conference Background
Graph Data Mining	Graph Data Mining	[37]	Environment and public health
Text Mining	Historical Patterns Co-occurrences	[53] [52]	Web intelligence Infection control
Topic Models	ATAM ATAM+ HFSTM	[54] [55] [56]	Health Social media Data mining
Machine Learning	Neural Network SVM	[69] [65] [64] [67] [66] [68] [63]	Cognitive information processing Multiple scientific disciplines Natural language processing Computational linguistic Biology and medicine developments Living system Computational biology
Math/Statistical Models	Pred. Market using SVR Naïve Bayes SNEFT ACF Numerical-Based Analysis (SEHA using BOW)	[72] [71] [40] [73] [75]	Social media mining Health care Networking systems Bioinformatics Sustainable global development
Mechanistic Disease Models	Metpopulation Model Compartmental Model Agent-Based Model	[77] [56] [44] [80]	World wide web Data mining Multiple scientific disciplines Healthcare informatics
Keys/Documents Filtration	Keys/Documents Filtration	[38]	Repository of pre-prints

2.5 Challenges

Using social media data for disease outbreak detections calls for certain challenges to be addressed [82, 83, 84, 85, 86].

2.5.1 Data Collection

The first challenge is the restriction on data collection. Social media providers use unknown and undocumented sampling filtration algorithms that allow for collecting only a sample of the overall data. In addition, there are restrictions on some private data that may be needed for the detection process. Also, users may not include some other important information. This may lead to inaccurate results produced by the tools of disease trend detection.

2.5.2 Data Size

The size of social media data is another challenge. Today, Social Networking Sites have become very popular and have millions of users. This challenge would make it difficult to process such size of data by certain techniques.

2.5.3 Language

The used language in Social Networking Sites is usually informal and sometimes with spelling errors. Users may spell one word in different ways.

2.5.4 Heterogeneity

Social Networking Sites are heterogeneous. They have different kinds of users with different capabilities, activities, ages, and languages. This leads to the need for awareness of what to analyze using the data of Social Networking Sites.

2.5.5 Sampling bias

One of the serious challenges is the bias of data samples. The user population of Social Networking Sites may not represent a sample of a society [84, 85, 86]. Alan Mislove et al. [84] analyzed the data of a very large number of Twitter users from United States to compare the Twitter population to the actual one. The study has shown that the Twitter users are not a random sample of the whole population and misrepresent the real distribution of race or ethnicity. Understanding this challenge will help in correcting the prediction process using social networking data if there is any bias. The correction process includes using different methods of bias quantification for further analysis and adjustment [85].

2.5.6 Dataset Consistency

Social media providers such as Twitter do not allow sharing collected datasets. This is a limitation when it comes to comparing a new proposed method and the existing ones. Consistent datasets are required for fair comparisons.

2.5.7 User Location

There is a lack of accurate user locations in SNS. A user may not share location information. In addition, the users who release this information may not update it when moving or visiting a different place.

2.5.8 Proxy Population

There are difficulties of defining a target population for the purpose of analysis. Populations are not self-labeled. Therefore, researchers tend to use proxy populations such as all users who use pain relievers to study the impact of pain. Using proxy population is biased and may lead to incorrect results [85].

2.5.9 Spams

There are many spam accounts that appear as normal and are frequently used to post different topics. Researchers should be aware of these accounts and find a way to exclude them when analyzing SNS data.

2.5.10 Evaluation

Evaluation is a challenging process. CDC ILINet data can be used as a ground truth for the Influenza trend detections but there is lack of ground truth for some other diseases.

2.6 Concluding Remarks

Social Networking Sites have become part of people's lives. This has provided researchers with the opportunity to conduct different studies and researches to enhance event detection and prediction process from the data of Social Networking Sites. In the public health area, the data of Social Networking Sites can be used to provide early warnings of disease outbreaks such as seasonal Influenza. The survey shows that the researchers have developed various methods and frameworks of flu trend detection from Social Networking Sites. From the survey, we conclude that the research in this area is still active. More methods and frameworks may be developed to improve the efficiency of the detection processes, and the accuracy of the results that can potentially be used for new disease outbreaks for better public health.

CHAPTER 3: RESEARCH METHODOLOGY

The proposed framework consists of three main modules: text classification, mapping, and a linear regression-based estimator for weekly flu rate predictions. Figure 3.1 shows a general overview of the proposed framework. The classification module, which is used to classify flu-related tweets, is implemented using the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a well-known standard for implementing data mining frameworks. This standard includes six steps [87]:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

Based on the CRISP-DM standard, the methodology for this study is presented in Figure 3.2.

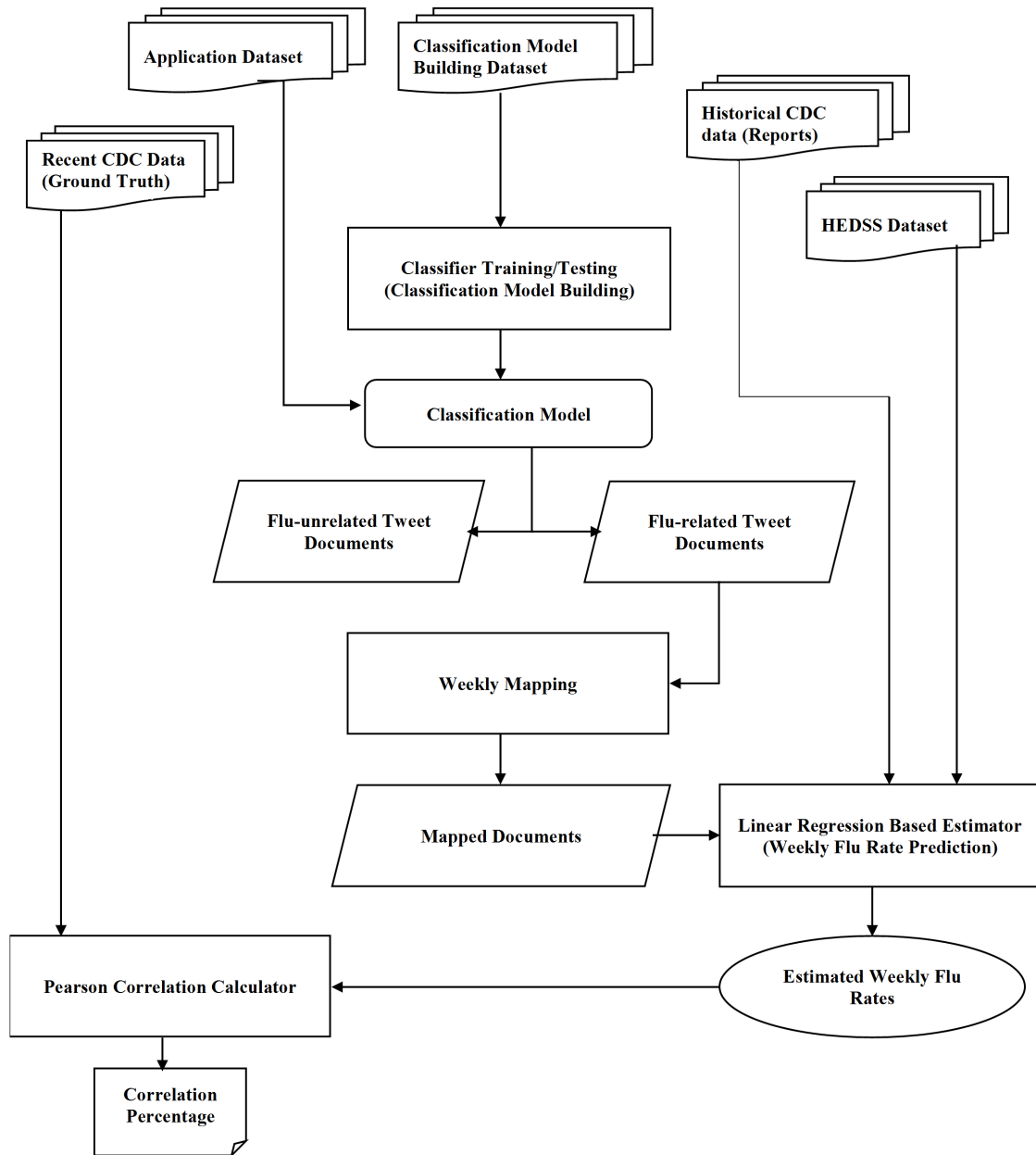


Figure 3.1: Proposed framework overview

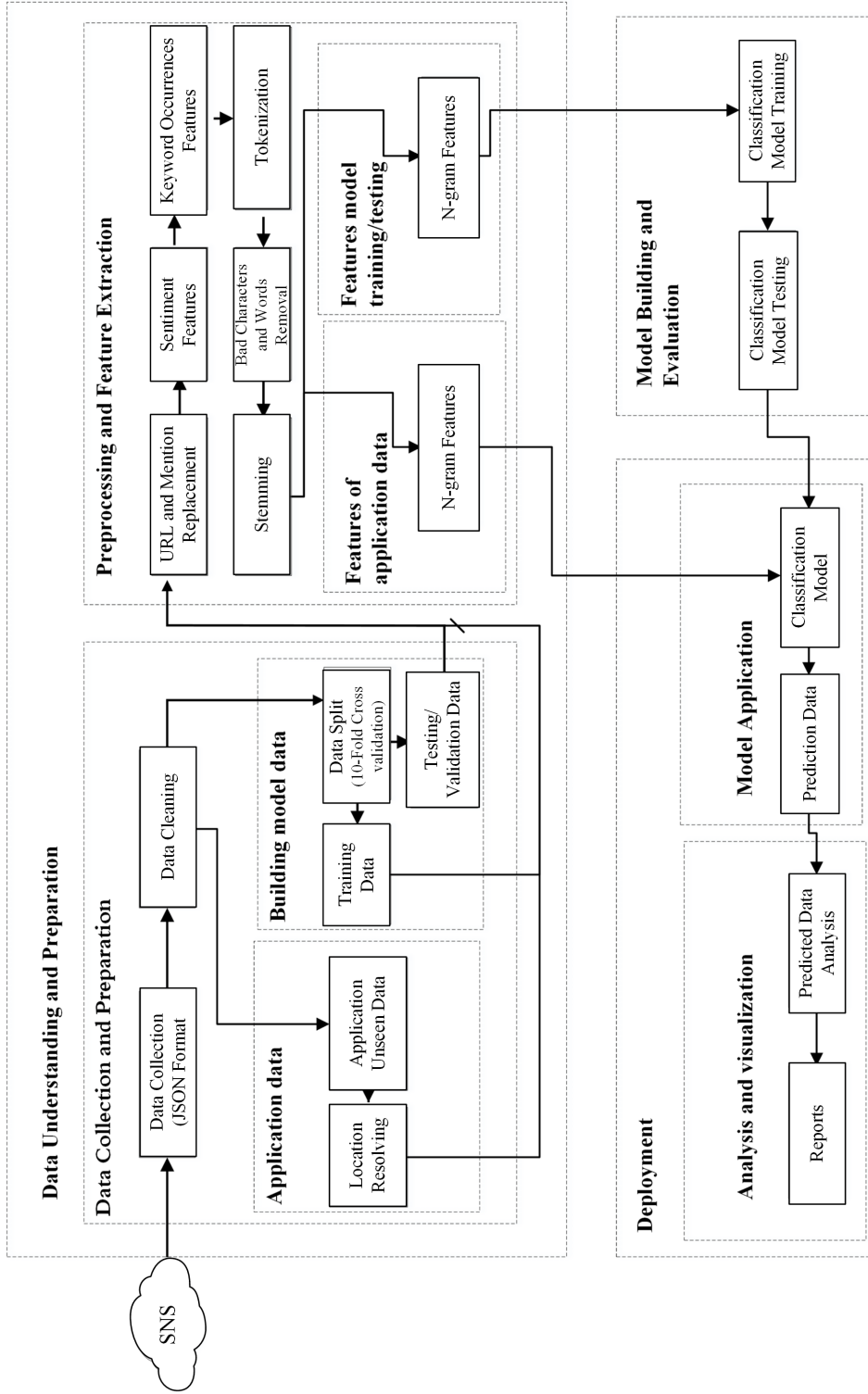


Figure 3.2: Methodology for text classification of flu tweets

3.1 Data Collection and Preparation

3.1.1 Classification Model Data

For classification model training and testing, we prepared a labeled dataset that is a combination of multiple manually labeled datasets obtained from [68, 88]. This makes the total instances of the merged dataset 10,592 tweets (5,249 flu-related and 5,343 flu-unrelated posts). Because of Twitter guidelines, the tweets in the obtained datasets were released with tweet IDs instead of the text of the tweets. Therefore, we developed a script that works together with the Twitter API to retrieve the corresponding tweet texts using the given IDs. The collected tweets were cleaned to include only the texts for training and testing purposes. Then, we divided the merged dataset into two parts: training set and testing set.

3.1.1.1 Twitter Influenza surveillance dataset

The labeled dataset obtained from [68] was initially filtered to contain any posts that have flu-related keywords. Then, every post in the dataset was labeled manually. The dataset was prepared to train and test three flu-related classifiers that were used as a part of an algorithm for seasonal flu predictions. The dataset was divided into three sets, one for each classifier. The first set consisted of tweets that were labeled as either flu-related tweets or unrelated. The second one had tweets with labels of flu infections or flu awareness. The tweets in the last set were labeled as either the flu tweet being about the author or about someone else. For the training dataset, we consider the tweets in the second and third datasets as flu-related tweets and combine all of them with only two labels: flu-related or unrelated.

3.1.1.2 Sanders dataset

The labeled dataset obtained from [88] was prepared manually to train and test sentiment analysis algorithms. Each record in the dataset is annotated with a sentiment label, indicating a feeling toward either Google, Twitter, Microsoft, or Apple. The labels are: positive, neutral, negative, and irrelevant. Since this dataset was prepared for sentiment analysis of topics that are not related to flu, we used all the tweets in this dataset except the ones with irrelevant labels as flu-unrelated tweets.

3.1.2 Application Dataset

For validation purposes, we prepared an application dataset by collecting a set of Twitter posts for the first 20 weeks of the year 2018 within the boundary box of the state of Connecticut as a location filter using its associated longitude and latitude. The data was collected from Twitter SNS using a crawler that works with the Twitter API to stream tweets. The crawler is designed to filter the tweets based on keywords that are directly related to flu and verified by healthcare professionals. The list contains 11 flu-related keywords: fever, headache, sick, respiratory virus, ache, stuffy nose, dehydration, flu, Influenza, contagious, and cough. Because of some technical problems, we were able to collect few Twitter documents for the 10th week. Therefore, we did not include the period of the 10th week in our experiments. The total number of tweets over the 19 weeks is 8,440,670.

3.1.3 CDC ILINet Data

ILI weekly rate produced by the CDC ILINet is used as a gold standard for comparison. The official ILI rates consider outpatients with symptoms of Influenza who have visited any location of ILINet-participated healthcare providers around the United States. The data is obtained from the official CDC website: (<https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>).

3.1.4 Data of Hospital Emergency Department Syndromic Surveillance (HEDSS) System

This data consists of the number of patients who have visited any location of the Emergency Departments (ED) of the hospitals in Connecticut. HEDSS generates daily reports about the daily patient visits based on the information received from the Emergency departments. The generated reports include a percentage of patient visits for Influenza [89]. This data is used to train the linear regression model for the final flu rate prediction for the state of Connecticut.

3.2 Preprocessing

During data preprocessing, stop-words, punctuations and symbols were removed before the training and testing processes using the Natural Language Processing Toolkit (NLTK) [90]. Stop words such as “the” or “are” are very frequent and may lead to inaccurate classification results if used as features. The preprocessing also includes stemming that is used to reduce words to their roots. There are many stemming algorithms available for use. For this study, the stemming algorithm employed is Porter Stemming. It is one of the most commonly used stemming algorithms. It is a rule-based algorithm with five steps that is designed based on the idea that English suffixes are made of smaller and simpler ones. A suffix is removed if a rule in the five steps passes the conditions and is then accepted [91]. Figure 3.3 shows the overall preprocessing steps that are used for this study.

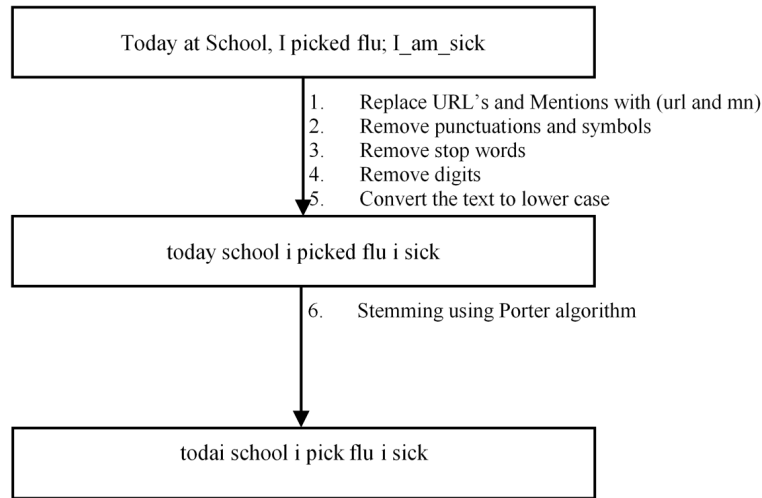


Figure 3.3: Text preprocessing

URL's, Hashtags, and Mentions in the tweets were kept in the corpus. They can be used as features for classification. URL's were replaced with the keyword (url), and Mentions were replaced with the keyword (mn) to be used as one feature for classification.

3.3 Feature Extraction

A maximum classification accuracy can be achieved by selecting the best set of features. Therefore, feature selection is a crucial process in any classification problem. In text classification, the set of features is a subset of words (n -gram) that can be used to distinguish different classes [92]. The selected words should provide useful information to be used for classification purposes. Thus, it is important to consider different techniques to convert the text in a way that can be processed to gain the required information. In this work, we consider additional features to enhance the classification accuracy. The additional features are sentiment based features, stylometric features, and flu-related keyword features (Algorithm 1).

3.3.1 Textual Features

The default features in text classification are the terms and words that make up the document/text. Text classifiers are trained and tested using n -gram features, as basic features, by breaking down the documents/texts into single words (uni-grams), terms composed of two words (bi-grams), and terms composed of three words (tri-grams) and/or more. A basic technique in text classification is to count n -gram features including the un-informative ones that may yield inaccurate results. Therefore, using smarter techniques is important. One of these techniques is the word/term weighting technique, which weighs the count for every word/term in the text. There are different techniques of word weighting that include Boolean weighting, Term Frequency weighting (TF), Inverse Document Frequency weighting (IDF) and Term Frequency–Inverse Document Frequency weighting (TF–IDF). Among the four types of word weighting techniques, only the IDF and TF–IDF techniques consider the importance of a word/term in the entire corpus instead of the importance of the word/term in only a document. The study [66] has shown that TF–IDF is more accurate than IDF. Therefore, TF–IDF has been used to weigh the n -gram features for the conventional machine learning classifiers.

TF–IDF value is obtained by multiplying the value of the Term–Frequency value by the value of Inverse Document–Frequency (Equation 3.1). TF is the ratio between the term t with frequency n_t in a given document d and the total numbers of terms n in the document d (Equation 3.2). IDF is the inverse of the number of documents that has the term t at least once. IDF is calculated using Equation 3.3, which is the ratio between the frequency N_d of the documents d that have term t , and the total number N of documents d in the analyzed corpus.

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (3.1)$$

$$TF(t, d) = \frac{n_t}{n} \quad (3.2)$$

$$IDF(t) = \frac{N_d}{N} \quad (3.3)$$

For the FastText classifier, the representations of textual features of a document are averaged and weighted to be fed to the classifier. For word ordering, FastText utilizes only partial information about the order by using bag of n -grams instead of bag-of-words with the full information of the word ordering [93].

3.3.2 Stylometric features

Stylometric features of Twitter posts include Retweets (RT), Mentions, and URL links. These features were kept in the corpus to be used for classification. URL links and Mentions to others were preprocessed by replacing them to url and mn keywords.

3.3.3 Topic-related keywords based features

It is common to use seed words in text classification. For example, in sentiment analysis, a list of words, including nice and good, is used for positive sentiment and another list of words, including bad and poor, can be used for negative sentiment. In this study, a set of flu-related keywords/terms was used as a set of features for flu-related tweets. The list includes some important Influenza-related keywords, symptoms, and treatments. The list of the keywords is kept in an array and then each tweet is compared against these keywords to keep track of their occurrences.

3.3.4 Sentiment based features

Sentiment analysis is the process of extracting the sentiment of a text using contextual polarity. Sentiment analysis is commonly used in classifying reviews of different

products on the Internet such as the sentiment of movies. In this study, we used TextBlob library to assign a sentiment to each tweet [94]. TextBlob is a Python library that is used to analyze textual data. Based on the polarity score of a tweet, a sentiment value is assigned to the text: positive or negative.

Algorithm 1 Additional Feature Extraction

```

twt_txt ← tweet_document.text
txtlen ← length of twt_txt
Feature Set1 :                                ▷ %Preprocessing and stylometric Feature Extraction
if URL in twt_txt then
    twt_txt ← replace URL with a keyword url
if Mention in twt_txt then
    twt_txt ← replace Mention with a keyword mn
token ← tokenize(twt_txt)
for (i = 0; i < txtlen ; i = i + 1) do
    if (token(i) is ineffective char) OR (token(i) in Stop_Word_lst) then
        remove token(i)

        stem ( lower (token (i))
twt_txt ← token
Feature Set2 :                                ▷ %Sentiment Feature Extraction
sent_ft ← 0
Polarity_score ← find_polarity(twt_txt)
if Polarity_score > 0 then
    sent_ft = 1
else
    sent_ft = 0
Feature Set3 :                                ▷ %Keyword Occurrences Feature Extraction
hsKwrд_ft ← 0
kwrд_lst_len ← length of keyword_lst
for (i = 0; i < kwrд_lst_len ; i = i + 1) do
    if keyword_lst(i) in twt_txt then
        hsKwrд_ft = 1
twt_txt_w_features ← concatenate(twt_txt, _sent_(sent_ft), _hsKwrд_(hsKwrд_ft) )

```

3.4 Classification Model Building - Training and Testing

For the sake of accuracy and efficiency, various classifiers are evaluated, including FastText and six conventional machine learning algorithms [95, 96]:

3.4.1 FastText

FastText (FT) was proposed by Facebook for word embeddings and text classification. In this study, FastText is used for text classification. FT produces accurate classification results that are comparable with the results produced by deep neural network classifiers. In addition, the processes of FT training and classification are very fast using a standard computer with a multicore processor. A FastText model can be trained using a billion of words in just a few minutes and can classify about five hundred thousand sentences in less than a minute [93].

FastText utilizes several techniques to enhance the efficiency. It is a linear-based model, scaled to very large data and large output space using a rank constraint and a fast loss approximation. It uses a hierarchal softmax function for a faster search. In addition, only partial information about the word order is utilized for prediction. Furthermore, FT utilizes the technique of hashing for textual feature mapping [93].

3.4.2 Conventional Machine Learning Classifiers

For training and testing, several supervised classification methods were evaluated to determine the one with better classification accuracy [95]. The evaluated conventional classifiers include Random Forest, Naïve Bayes, SVM, C4.5 Decision Tree, K-nearest neighbors classifier (KNN) using the Instance Based learning algorithm (IBK), and AdaBoost. The preprocessed labeled dataset was used to train and test the model of different classifiers using 10-fold cross validation as the experimental setting. The 10-fold cross validation is

a method to validate the studied/built model by iterating through the labeled data 10 times with different subsets of training and testing for each iteration.

3.4.2.1 Support Vector Machines (SVM)

SVM was proposed in 1998 by Vapnik [97]. It uses the features of the provided training dataset to decide the classification boundary (hyperplane) that divides the space into regions, one for each class. SVM only chooses part of training samples that are close to the boundary to form the support vector instead of using the whole feature space in order to distinguish between the classes.

3.4.2.2 K-Nearest Neighbor (KNN)

KNN is a simple classifier. It uses the provided training set as an input vector to form different regions for different classes. Each sample in the training dataset is mapped to a point in the feature space. When a new unlabeled sample requires classification, KNN identifies the approximate distances between its associated point and k -neighbors in the space and then assigns the point to the class of the majority of the k -nearest neighbors. The K value plays an important role in the performance of KNN classifiers [98]. A large value may increase the classification accuracy, but it requires more computation time. In this study, we use an extended version of KNN that is implemented based on Instance-Based learning algorithm (IBK) [99].

3.4.2.3 Random Forest (RF)

RF consists of several decision trees (n). Each tree is trained separately using a random feature subset of the training dataset T . The decision trees altogether make an accurate classifier. An unlabeled instance is classified by all the trees separately. Then, the final decision about the class of the unlabeled instance is decided by the majority voting

technique [100].

3.4.2.4 C4.5 Decision Tree

For this study, J48 that has been used is an implementation of the C4.5 decision tree [101]. It consists of decision and leaf nodes. A decision node (internal node) is the node that has a child and a leaf (terminal) node is the node that specifies the class value. The decision nodes are used to represent different attributes. When a decision node represents a discrete attribute, its child nodes indicate different possible values of the discrete attribute. A decision node for a continuous attribute has two child nodes. Each child indicates a certain range of the continuous attribute that is determined by using a threshold value. Lastly, the terminal nodes (leaves) represent the final value of the class labels. A decision tree is constructed and tested using a training dataset. Then, a new instance is classified by the tree based on the values of the attribute (features) [101, 102].

3.4.2.5 Naïve Bayes

Naïve Bayes is a supervised classifier that uses the conditional probability formula for classification. It is constructed using a training dataset with a prior knowledge about the relationships between the attributes (features), and a directed acyclic graph (DAG) that is used to represent conditionally independent features for a given class and their relationships. Each feature is represented by a node, and each relationship is represented by a link in the graph. The links indicate the influences between different variables [103].

3.4.2.6 Ensemble classifier

An ensemble classifier is a combination of multiple classifiers that work together to enhance the accuracy of classification. Each classifier can be trained individually with different subset of a training dataset to improve the performance of classification [104, 105,

106]. In this study, we evaluate the AdaBoost ensemble classifier.

AdaBoost was introduced by Freund et al. [106]. Boosting is an enhanced version of bagging. It consists of multiple base learners that are trained sequentially. The first learner is trained using a subset (bag) of random selected instances n from a training dataset T . The trained model is then tested using the training dataset T . During the testing process, weights are assigned to the examined instances with high weight values for the misclassified instances. Based on the weight values, the instances are picked for the next bag. Thus, instances with higher weights should be picked to train the next base learner in the sequence.

3.5 Mapping

The flu-related classified documents must be summarized on a weekly basis by counting all the documents that belong to the same week to be passed to the estimator to find the weekly flu rate. The mapping method takes an input as a pair (week number and post), groups all the posts associated with the same week number, and then merges all the pairs with the same week number by counting the associated posts.

Since Social Networking Sites have enormous data, Hadoop systems could be utilized for the mapping process. These tools and techniques can be used to parallelize the MapReduce Programming approach that allows programmers to utilize the resources of large distributed systems [107]. A MapReduce (MR) approach can be used to process the large dataset of tweets. MR consists of two main functions: Map and Reduce. The Map function takes an input as a pair (week number and post), groups all the posts associated with the same week number, and generates intermediate pairs to be passed to the Reduce function. The Reduce function merges all the pairs with the same week number after processing the associated values such as counting or summing them up [108]. Figure 3.4 shows a general overview of the flow of Hadoop MapReduce programming approach.

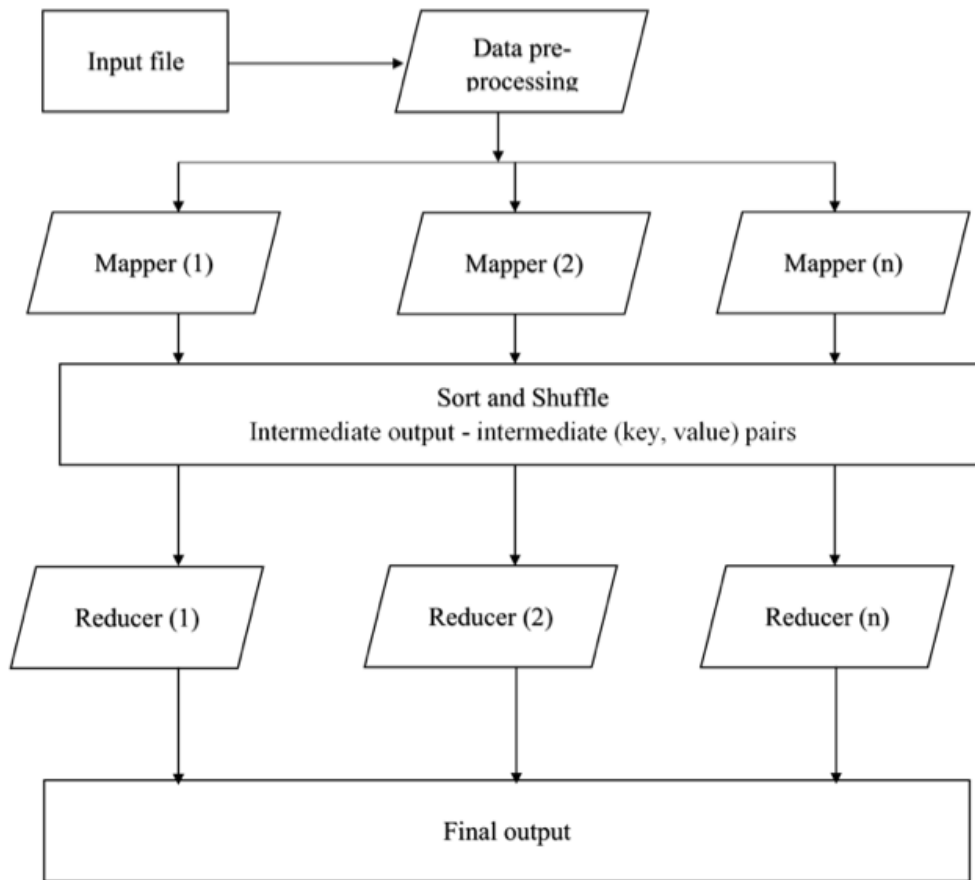


Figure 3.4: General flow of Hadoop MapReduce programming approach

The mapping process can also be achieved by employing features of other big data tools such as Hadoop Eco Systems, and Apache Spark:

Hadoop Eco Systems Hadoop Eco Systems such as Hive are data analytic tools to manage and query large datasets. They are built on top of Hadoop to provide an easy way to query and manage data. Hive allows users to query large datasets using a SQL-Like script (HiveQL) instead of MapReduce programming. The performance of queries written in HiveQL are similar to the ones written in MapReduce framework [109].

Apache Spark Spark was developed in 2009. It supports real time streaming data and fast queries. Spark runs on top of Hadoop to replace the data batch process of the traditional MapReduce model in order to support real time streaming data processes. Spark performs tasks based on two concepts. The first concept is the Resilient Distributed Dataset (RDD), which is an abstract collection of an element that can be processed in parallel. It is a read-only collection of objects partitioned across a set of nodes. RDD supports two kinds of operations: Transformations and Actions. Transformation operations take RDDs and only return new RDDs and nothing evaluated. Transformation functions include map, filter and reduceByKey. Action operations are also applied on RDDs that include evaluation and returning new values. Action functions include reduce, collect and take. The second concept concerns the Directed Acyclic Graph (DAG), which is an engine that supports cyclic data flow. Spark creates a DAG for each job that consists of task stages (map and reduce) to be performed on a cluster [110].

3.6 Weekly Flu Rate Estimation

To predict the Influenza rate for a certain week, a proposed estimator based on a regression model is used as a component of the framework. The predictors for the regression model include a combination of the rate of flu tweets and the average ILI rate of the same week number of past years (from 1998 to 2016). The proposed flu rate estimator has been evaluated using different regression models to determine the one with better estimation accuracy.

A regression model is trained (fitted) using available data of flu rates, such as the data obtained from FluNearYou [111]—a web application that uses weekly surveys to collect the health status of individuals—or the data of flu emergency visits obtained from HEDSS. For this study, we used the data of HEDSS for regression models training, where the average ILI rates of previous years and rates of flu-related tweets obtained from the classification

results are passed to the regression models as predictors. The regression models are then tested and validated using CDC ILINet data.

3.6.1 Linear Regression Model

Linear Regression is used when the dependent variable (response) is continuous and the independent variables (predictors) are either continuous or discrete, and the relationship between the dependent and independent variable(s) is linear. The linear regression indicates that the rate in the change of the mean of the response value is constant with respect to the value of the predictor(s). Therefore, the relationship is represented by an equation of a line [112].

Using the proposed combination of predictors for the weekly rate estimator, our proposed linear regression model has the following form:

$$F_w = \beta + \alpha_1 \frac{1}{2016 - 1998} \sum_{y=1998}^{2016} F_w^y + \alpha_2 T_w \quad (3.4)$$

where F_w indicates the flu rate at week w , β is the intercept which is the mean value of F_w when all predictors are 0, α_i values represent the regression coefficients, F_w^y is the actual rate of flu incidents in week w of year y , and T_w is the rate of flu tweets in week w .

3.6.2 Other Regression Models

In addition to our proposed linear regression model, three different regression techniques were evaluated to determine the technique with better estimation accuracy. The evaluated techniques are Polynomial Regression, Logistic Regression, and Support Vector Regression. The measure of Pearson Correlation, which is discussed in Chapter 4, is used to find the most accurate model to be used for the final weekly flu rate estimation.

Logistic Regression Logistic regression is commonly used for binary classification problems. It is used with a binomial distribution of dependent variables. Thus, it includes a function of Logit transformation that is suitable for the distribution. The Logit function is applied to handle different types of relationships between the dependent and independent variables.

The logistic regression finds the probability of the categorical values of the dependent variables [112]. In this study, the probability values, which ranges between 0 and 1, have been used to indicate the weekly flu rates.

Polynomial Regression Polynomial Regression is used when the dependent variable (response) is continuous and the independent variables (predictors) are either continuous or discrete, and the relationship between the dependent and independent variable(s) is not linear. The polynomial regression indicates that the rate in the change of the mean of the response value is not constant when the value of predictors increase or decrease [112].

Support Vector Regression Support Vector Regression (SVR) is a technique, which uses the same concepts of the classification method (SVM), for regression. It predicts the continuous values of dependent variables with respect to the values of independent variables (predictors). SVR can be used for both linear and non-linear problems [113].

CHAPTER 4: IMPLEMENTATION AND TESTING

The proposed framework consists of data preprocessing, which includes stemming and removal of stop words and ineffective characters. The preprocessing phase is implemented using the Python Natural Language Processing Toolkit (NLTK). The framework also consists of text classification module that utilizes the features of sentiment analysis and predefined keyword occurrences. This module is evaluated by using various classifiers to identify the most efficient and accurate one. The framework has been trained and tested using a pre-labeled dataset of flu-related and unrelated Twitter postings. The classification results demonstrate that FastText improves the accuracy and the efficiency of flu disease surveillance systems using SNS data. The trained classification model is then used to classify over 8,400,000 tweet documents that are collected using a developed crawler. The crawler works together with the Twitter API to stream tweets for the first 20 weeks of the year 2018 within the boundary box of the state of Connecticut. The flu-related documents are then mapped on a weekly basis using a mapping module that groups all the posts associated with the same week number and then merges all the ones with the same week number by counting them. The results are passed together with historical CDC data, as a combination of predictors, to an estimator module for weekly flu rate predictions. Finally, the weekly prediction results are compared to the available recent data from CDC.

4.1 Flu Post Classification

To build a classification model with better accuracy and efficiency, FastText and several supervised classification methods using the proposed additional features were evaluated. In addition to FastText, the evaluated classifiers are Random Forest, Naïve Bayes, SVM, C4.5 decision tree, K-nearest neighbors classifier using the Instance Based learning algorithm (IBK), and AdaBoost. The preprocessed labeled dataset was used to train and test models of the different classifiers with the TF-IDF based n -gram features and the proposed additional ones that are presented in the Feature Extraction Section (3.3).

For a better FastText model, we evaluated 28 different feature settings using FastText with the parameter values of learning rate = 0.8 and epoch = 8, to determine the best feature set. Initially, the model was trained and tested using one setting of n -gram features ($n=1$ to 6), which are tokens of (n) words including the stylometric features. Then, different settings of the additional features are combined with the tweet text for training and testing using n -grams ($n=1$ to 6). The settings include a combination of text and sentiment features, a combination of the text and keyword occurrence features, and a combination of all additional features (text + sentiment + hasKeyword):

label<related/unrelated> TEXT _sent_<neg/pos> _hasKeywrд_<yes/no>

With a standard computer (2.6 GHz Intel Core i7 processor, and 16 GB RAM), the preprocessed labeled dataset was used to train and test the models using 10-fold cross validation as well. The 10-fold cross validation is a method to validate the studied/built model by iterating through the labeled data ten times with different subsets of training and testing for each iteration.

Baseline Classifier The literature shows that most of the existing models are based on Support Vector Machine (SVM) for the text classification to distinguish between related

and unrelated flu tweets. Therefore, SVM with basic textual features have been considered as a baseline classifier for comparison purposes.

4.2 Performance Metrics

4.2.1 Text Classification

This section presents the used performance metrics. The performance of the classifiers are evaluated using different metrics: accuracy (Equation 4.1), precision (Equation 4.2), recall (Equation 4.3), and F-measure (Equation 4.4). These metrics are used to provide a better overview of the model performance. The accuracy measure by itself is not a perfect measure if the dataset is not balanced. Precision and recall are better measures in the case of imbalanced datasets. The selected metrics can be computed using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) measures, where TP refers to the rate of correctly classified instances as positive, TN refers to the rate of correctly classified instances as negative, FP refers to the rate of incorrectly classified instances as positive, and FN refers to the rate of incorrectly classified instances as negative. In this work, we mainly use F -measure as a performance metric for evaluation and comparison. F -measure is a weighted average of two different performance metrics: precision and recall. Its value ranges between 0 (worst) and 1 (best).

Accuracy Accuracy is a measure to evaluate the performance of a prediction model. It is the rate of the correctly classified labels. It is calculated by using Equation 4.1 :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Precision Precision measures the true positive predictions. The precision of a model is calculated by using Equation 4.2:

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

Recall Recall is a sensitivity measure. It is used to evaluate a model's performance in predicting positive labels. It is calculated by using Equation 4.3:

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

F-Measure F-Measure takes into account both measures: recall and precision. It can be considered as a weighted average of precision and recall measures with a value ranging between 0 (worst) and 1 (best). F-measure is calculated using Equation 4.4:

$$F-Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

4.2.2 Flu Rate Estimation

The performance of flu rate estimation is evaluated using Pearson Correlation. This measure is used to evaluate the performance of the flu rate estimator using different regression models. CDC weekly reports are used as the ground truth to be correlated with the output of the proposed estimator.

Pearson Correlation Pearson Correlation is a metric that evaluates the correlation between two datasets. Let y_i be the observed value of the ground truth (CDC ILINet data), x_i be the predicted value (estimated weekly flu rate), and \bar{y} and \bar{x} be the average values of $\{y_i\}$ and $\{x_i\}$, respectively. Using these notations, the Pearson Correlation value r is defined as shown in Equation 4.5 [63].

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (4.5)$$

CHAPTER 5: RESULTS

5.1 Classification Results

The results show that the proposed model improves the performance of flu post classifications using a combination of the additional features. The performance results of the evaluated classifiers are shown in Table 5.1 using the discussed metrics in the previous chapter. The Random Forest method achieved the highest accuracy results, with an F -measure of 90.1%. In addition, we used the Receiver Operating Characteristic (ROC) metric to evaluate the utilized classifiers. ROC is a curve with points that represent the pair of true positive rate (Sensitivity) and false positive rate (Specificity). A perfect curve is the one that passes through the upper left corner representing 100% sensitivity and 100% specificity. Thus, the closer the curve is to that corner, the better the accuracy is [114]. As shown in Figure 5.1, Random Forest appears to be the best classifier. The high accuracy results demonstrate the efficiency and effectiveness of the extracted features.

Moreover, the performance results of FastText with different sets of features is presented in Figure 5.2. The overall accuracy using the F -measure metric ranges between 86.47% and 89.9%. The results demonstrate the efficiency of the FastText classifier. The highest classification accuracy is achieved by using the 5-gram features together with all the proposed additional features (F -measure = 89.9%) in only 21.53 seconds for training and testing using 10-fold cross validation on a standard computer (2.6 GHz Intel Core i7 processor, and 16 GB RAM). It has been shown that FastText can produce, in a short time,

accurate results that are comparable to the results produced by the state-of-the-art deep neural network classifiers [93]. The high accuracy together with the efficiency of FastText make it an optimal classifier for flu disease surveillance models/systems with very large data. Therefore, FastText will be used for our further analysis.

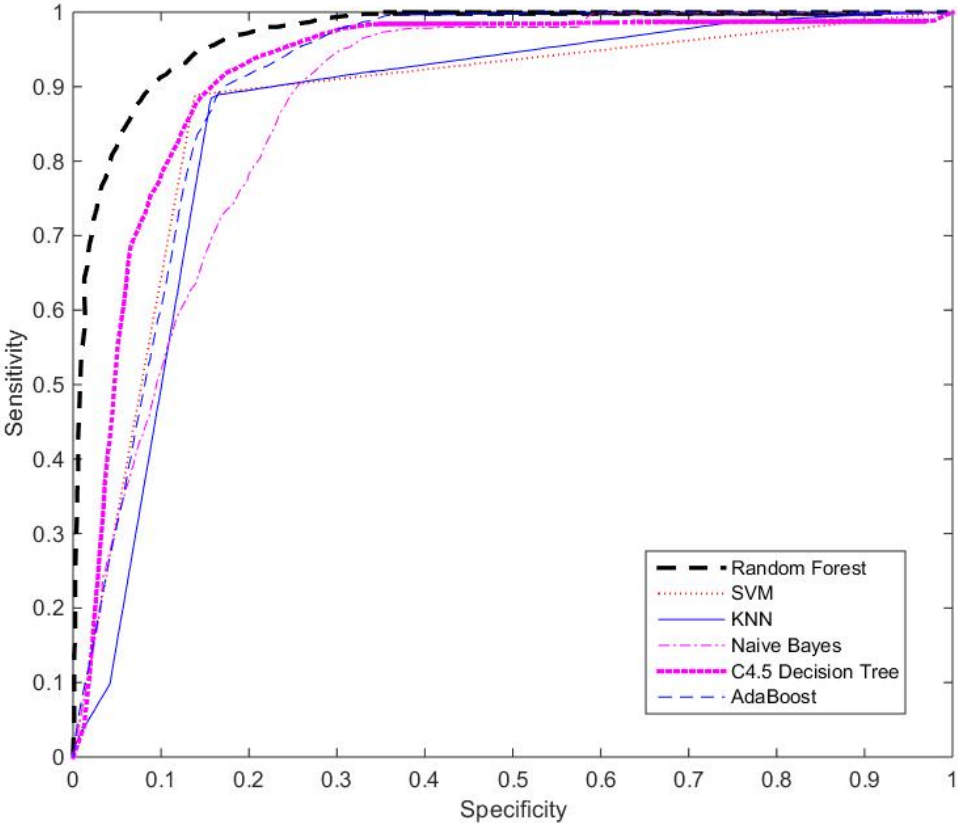


Figure 5.1: Performance comparison using ROC

Table 5.1: Performance of classifiers

Classifier name	Precision	Recall	F-measure
C4.5 Decision Tree	0.876	0.85	0.873
Random Forest	0.905	0.902	0.901
SVM	0.883	0.883	0.883
Naïve Bayes	0.846	0.826	0.824
AdaBoost	0.867	0.864	0.864
KNN	0.874	0.872	0.872
FastText	0.899	0.899	0.899

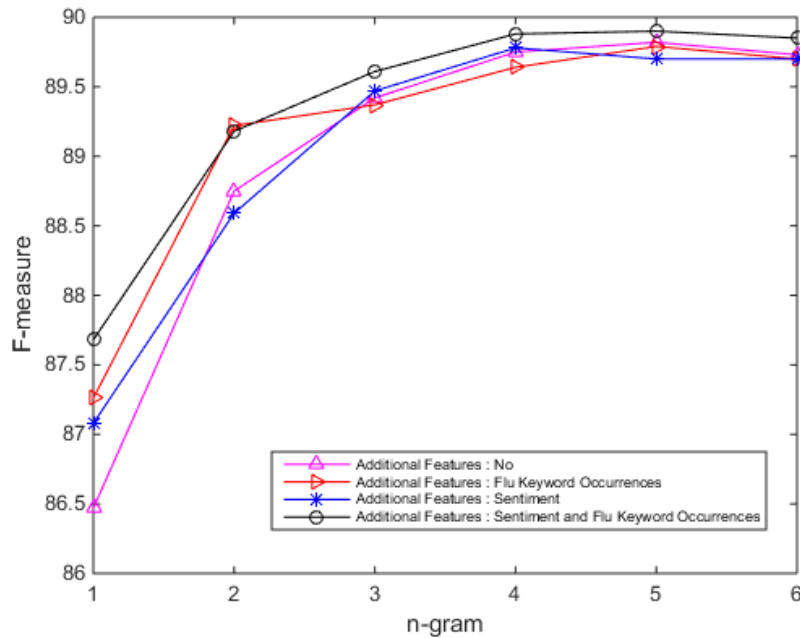


Figure 5.2: FastText performance using different sets of features

Many studies have utilized the available data from Twitter to build faster Influenza surveillance systems [39]. All the reviewed studies use conventional machine learning

methods to distinguish between flu-relevant and irrelevant posts for further analysis. A summary of the performance results of previous works, which include tweet classification for Twitter-based flu surveillance systems, and a comparison between the performance results of the proposed framework and the baseline are shown in Table 5.2. The metrics are reported as percentages. The evaluation of flu tweet classification using the F -measure shows that the proposed framework using FastText together with the extracted features, has achieved high accuracy with F -measure value of 89.9% in only 21.53 seconds for training and testing using 10-fold cross validation. This is while the F -measure value of the baseline is 86.6 and requires more than 30 minutes for training and testing using the same settings.

Table 5.2: Summary of the reviewed flu posts classifiers (Flu-Relevant / Flu-Irrelevant)

Reference	Classifier name	Precision	Recall	F-measure	Note
[65]	SVM and Logistic Regression	67	87	75.62	Multi-level classification
[68]					
[67]	Naïve Bayes and SVM	N/A	N/A	83	
[64]	SVM	N/A	N/A	75.6	
[66]	SVM	87.49	92.28	89.68	
[71]	Naïve Bayes	N/A	N/A	N/A	Accuracy= 70
Baseline Classifier	SVM	86.6	86.6	86.6	
Proposed Framework	Random Forest	90.5	90.2	90.1	
Proposed Framework	FastText	89.9	89.9	89.9	

5.2 Weekly Flu Rate Estimation Results

The framework was evaluated by applying the trained FastText model on the application data, which includes over 8,400,000 tweets, for classification. Then, the classification results together with the historical CDC data were passed on to the proposed regression-based estimator as predictors to obtain weekly flu-rates. The results of the flu estimator show a highly correlated output to the gold standard data (CDC). The estimator was evaluated using several regression models. Every model was fitted using the data of flu emergency visits obtained from HEDSS. Then, it was tested on CDC ILINet data from January 1, 2018 to May 19, 2018.

The performance results of the proposed flu rate estimator based on different regression models are shown in Table 5.3. The table demonstrates the accuracy results using the Pearson Correlation measure that has been discussed in the previous chapter. The Linear Regression based estimator achieved the highest accuracy results, with a Pearson Correlation of 96.2%. Figure 5.3 also shows that Linear Regression is the most correlated model with the ground truth (CDC). In addition to the efficiency of linear regression, the experimental results demonstrate the model accuracy and confirm the linear relationship between the rates of weekly flu (dependent variable) and flu-related tweets (independent variable). Therefore, the linear regression model is used for the module of weekly flu rate estimation.

Table 5.3: Performance of Flu rate estimator using different regression models

Regression Model	r Value
Polynomial Regression	$r=0.895$
Logistic Regression	$r=0.917$
Support Vector Regression	$r=0.930$
Linear Regression	$r=0.962$

Figure 5.4 shows the normalized rate of ILI patients obtained from CDC and the normalized rate of ILI Twitter posts obtained from the output of our proposed solution during the period of January through May of 2018 for the state of Connecticut. The rate values of the proposed framework and ILINet are normalized to a common scale for comparison.

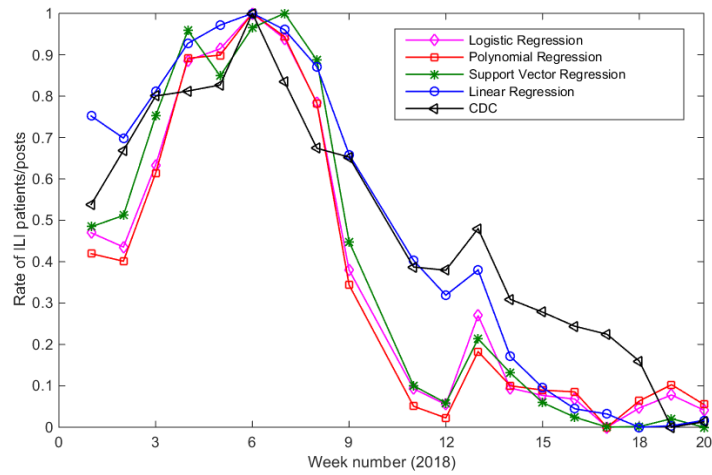


Figure 5.3: Correlation between the proposed framework and CDC ILI rate using different regression models

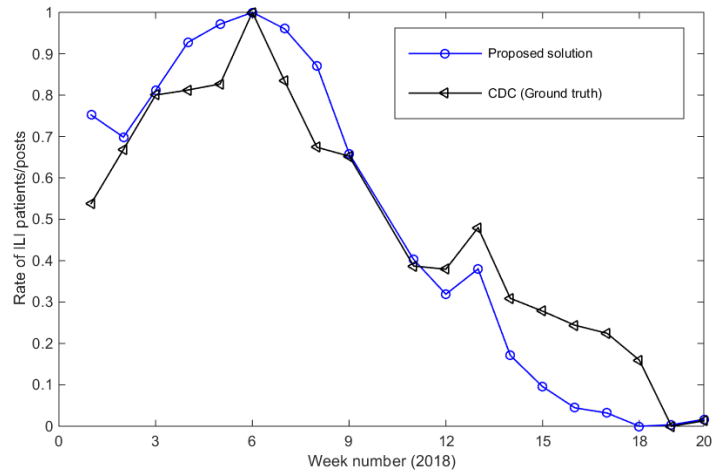


Figure 5.4: Correlation between the proposed framework and CDC ILI rate

CHAPTER 6: DISCUSSION AND VALIDATION

The performance of weekly flu rate estimation is evaluated using Pearson Correlation. It measures the correlation between two datasets using the symbol r that ranges between (1) and (-1): the value of $r = 1$ when both datasets exactly match and the value of $r = 0$ when there is no correlation between the two datasets. An available ground truth is usually used to evaluate the quality of the results of the proposed methods and frameworks. This study used recent CDC weekly reports as the ground truth to be compared with the proposed solution. The Pearson Correlation value r is defined as shown in Equation 4.5 [63].

As shown in Table 6.1 and depicted in Figure 5.4, the results show a strong correlation (96.2% Pearson Correlation) between the output of the proposed framework and the CDC reports. This correlation percentage shows that our proposed solution provides accurate results on a par with the best results in our survey, while being more efficient (faster). In addition, we believe that this study is the first work that utilizes Twitter postings for flu trend predictions in the state of Connecticut with strong correlated results. To the best of our knowledge, this is also the first work that shows a Twitter-based solution for flu prediction using recent data that is collected in the year of 2018.

6.1 Computational Complexity

The experiments show that FastText produces accurate classification results in only 21.53 seconds for training and testing using 10-fold cross validation on a standard com-

Table 6.1: Summary of the reviewed studies with reported Pearson Correlation

Study Reference	TimeFrame	Location	<i>r</i> Value
[65]	Sep 2012-May 2013	US	r=0.93
[64]	Nov 2008-Jun 2010	Japan	r=0.89
[67]	Mar 2010-Feb 2012	Portugal	r=0.89
[68]	May 2009-Oct 2010	US	r=0.9897
[66]	Sep 2013-Dec 2013	China	N/A
[71]	Oct 2015-Nov 2015	Ottawa	N/A
Proposed Framework	Jan 2018 -May 2018	CT, US	r=0.962

puter (2.6 GHz Intel Core i7 processor, and 16 GB RAM). FastText is an efficient linear based model. It uses a hierarchal softmax function that reduces the computational complexity to become logarithmic $O(\log n)$, leading to faster classification training and testing [93]. For word ordering, only partial information about the order is utilized by using a bag of n -grams instead of a bag-of-words with the full information of the word ordering. For more efficiency, the bag of n -grams are mapped using hashing techniques [93]. On the other hand, the experiments show that Random Forest, which is the most accurate conventional classifier in our experiment with F -measure value of 90.1, requires a longer time (39 minutes and 26 seconds) for training and testing using the experimental settings. The worst time complexity of Random Forest is quadratic for training $O(n^2 \log n)$ and linear for prediction $O(n)$ [115]. This together with the experimental results demonstrate the efficiency and the accuracy of FastText classifier. FT is an optimal classifier to detect new outbreaks with new signs and symptoms published in posts of Social Networking Sites. Therefore, FastText has been used for our further analysis.

6.2 Statistical Power Analysis

The power analysis has been performed to justify and ensure the appropriateness of the number of instances that are used for this study. Experimental results show that the

accuracy of flu tweet classification component using FastText with the proposed additional features outperform FastText with only textual features. Therefore, the power analysis is also used to prove this hypothesis that is stated as an alternative hypothesis H_a , whereas the null hypothesis H_0 is the hypothesis where there is no change in the accuracy using proposed features with respect to only textual features. With the power analysis, a statistical test rejects the null hypothesis when it is false. With this, we can conclude that there is a difference between the accuracies (better accuracy) using additional features and can confirm our alternative hypothesis H_a . If the null hypothesis is not rejected, then the alternative hypothesis should be rejected. The opposing hypotheses for our work can be stated as follows:

$$H_0 : \mu_{proposed} = \mu_{textual} \quad (6.1)$$

$$H_a : \mu_{proposed} > \mu_{textual} \quad (6.2)$$

where $\mu_{proposed}$ is the accuracy average of FastText using the proposed additional features and $\mu_{textual}$ is the accuracy average of FastText using only textual features for flu tweet classification.

To determine the required sample size n , four parameters/factors must be known or estimated:

- α : Significance level (1% or 5%)
- p : Desired power of the test (80%)
- σ : Population standard deviation.
- d : Effect size (the difference between the two groups)

The values of the first two parameters are generally fixed. The parameter of significance level α is usually set to either 0.05 or 0.01 and is the probability of rejecting the null

hypothesis when it is true. The power parameter p is the probability that the effect will be detected and is usually set to either 0.8 or 0.9. On the other hand, the last two parameters are problem dependent. For our analysis, the last two parameters are estimated based on our previous experiments. Thus, the values of all the four required parameters are stated below:

- $\alpha = 5\%$
- $p = 80\%$
- $\sigma = 0.27$
- $d = 0.012$

Using these parameters together with the z-test model to obtain z-scores, the sample size n can be computed by using Equation 6.3.

$$\text{Sample size}(n) = 2 \times \left(\sigma \times \frac{z_{1-\frac{\alpha}{2}} + z_p}{d} \right)^2 \quad (6.3)$$

Given the estimated values of the required parameters, we will have:

$$\begin{aligned} \text{Sample size}(n) &= 2 \times \left(0.27 \times \frac{z_{1-\frac{0.05}{2}} + z_{0.8}}{0.012} \right)^2 \\ \text{Sample size}(n) &= 2 \times \left(0.27 \times \frac{1.959 + 0.8416}{0.012} \right)^2 \\ \text{Sample size}(n) &= 7941 \end{aligned}$$

Using the obtained sample size n and the significance level α , the below parameters can be computed in order to apply the z-test and then make a decision on accepting or rejecting our alternative hypothesis:

$$\text{Mean}(\hat{x}) = \frac{\sum x}{n} = \frac{7292}{7941} = 0.918 \quad (6.4)$$

$$\text{Variance}(\sigma^2) = \frac{\sum(x - \hat{x})^2}{n} = 0.075 \quad (6.5)$$

$$\text{Standard Deviation}(\sigma) = \sqrt{\sigma^2} = 0.27 \quad (6.6)$$

$$\text{Critical } z = z_{1-\frac{\alpha}{2}} = 1.96 \quad (6.7)$$

$$\text{Standard Error}(S_x) = \frac{\sigma}{\sqrt{n}} = 0.003 \quad (6.8)$$

$$\text{Lower limit} = \hat{x} - \text{Critical } z \times S_x = 0.912 \quad (6.9)$$

$$\text{Upper limit} = \hat{x} + \text{Critical } z \times S_x = 0.923 \quad (6.10)$$

$$\text{Null Hypothesis } (H_0) : \mu_{\text{proposed}} = \mu_{\text{textual}} = 0.864 \quad (6.11)$$

$$Z_{\text{test}} (Z) = \frac{\hat{x} - \mu_{\text{textual}}}{S_x} = 18 \quad (6.12)$$

Since the obtained value of the z-test (18) is higher than the critical value ($18 > 1.96$), the observed difference is significant and shows that the additional features enhance the accuracy of FastText to classify flu tweets. In other words, results of the z-test show that the null hypothesis (H_0) should be rejected, and the sample set of 7,941 tweets is sufficient to prove that FastText with the proposed additional features is more accurate than FastText with only textual features for flu tweet classification. Our experimental results included over 10,000 tweets which is more sufficient to prove the hypothesis claims.

CONCLUSION

For disease surveillance models, gathering related information about diseases and then issuing necessary reports at an early stage is crucial for outbreak prevention. Data of microblogging sites, such as Twitter, have become popular to be used as triggers for different event prediction such as disease outbreaks. Recently, many studies have utilized this data to build faster epidemic prediction models such as flu outbreak prediction. The literature indicates that most of the models utilize conventional machine learning methods to filter and distinguish between the flu-relevant and irrelevant posts for further analysis. In our study, we introduced a framework based on FastText, a state-of-the-art text classifier, that utilizes the features of sentiment analysis and flu keyword occurrences for classification. Then, a combination of the classified Twitter documents and historical CDC data is passed to a linear regression-based module for weekly flu rate predictions. The results demonstrate the efficiency and the accuracy of the proposed framework. The final predicted flu trend using Twitter documents show a strong Pearson Correlation of 96.2% with the ground truth data of CDC for the first few months of 2018.

For future directions, more application data can be collected to cover a longer period of time to validate the regression model for flu rate estimation. In addition, other regression predictors, such as the data of FluNearYou, can also be investigated for further enhancement of the weekly flu rate estimation. Moreover, our classification model can be trained with a larger training dataset that includes more posts of Social Networking Sites for better classification accuracy. Furthermore, additional features and classifiers can be examined

for further accuracy and efficiency enhancement of the proposed framework. Lastly, our proposed framework can be fine-tuned to predict different outbreaks and events using the data of Social Networking Sites.

REFERENCES

- [1] L. W. Zhang and D. H. Zhu, "Research of technical development trend and hot points based on text mining," in *Proceedings of the 2nd IEEE International Conference on Information Engineering and Computer Science (ICIECS)*, pp. 1–5, Dec 2010.
- [2] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using internet searches for influenza surveillance," *Clinical infectious diseases*, vol. 47, no. 11, pp. 1443–1448, 2008.
- [3] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National academy of sciences (PNAS)*, vol. 107, no. 41, pp. 17486–17490, 2010.
- [4] M. Scharnow and J. Vogelgesang, "Measuring the public agenda using search engine queries," *International Journal of Public Opinion Research*, vol. 23, no. 1, pp. 104–113, 2011.
- [5] A. F. Dugas, Y.-H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman, "Google flu trends: correlation with emergency department influenza rates and crowding metrics," *Clinical infectious diseases*, vol. 54, no. 4, pp. 463–469, 2012.

- [6] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, “Generank: using search engine technology for the analysis of microarray experiments,” *BMC bioinformatics*, vol. 6, no. 1, p. 1, 2005.
- [7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.
- [8] K. Lee, A. Agrawal, and A. Choudhary, “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*, pp. 1474–1477, 2013.
- [9] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter improves influenza forecasting,” *PLOS Currents Outbreaks*, 2014.
- [10] K. Talvis, K. Chorianopoulos, and K. L. Kermanidis, “Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages,” in *Proceedings of the 9th IEEE International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 83–87, 2014.
- [11] C. Chew and G. Eysenbach, “Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak,” *PloS one*, vol. 5, no. 11, p. e14118, 2010.
- [12] J. Mowery, “Twitter influenza surveillance: Quantifying seasonal misdiagnosis patterns,” *Online Journal of Public Health Informatics*, vol. 8, no. 3, 2016.
- [13] M.-H. Hwang, S. Wang, G. Cao, A. Padmanabhan, and Z. Zhang, “Spatiotemporal transformation of social media geostreams: a case study of twitter for flu risk analysis,” in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, pp. 12–21, 2013.

- [14] P. Kostkova, M. Szomszor, and C. St Louis, “#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 (TMIS)*, vol. 5, no. 2, p. 8, 2014.
- [15] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in *Proceedings of the second IEEE international conference on Social computing (socialcom)*, pp. 177–184, 2010.
- [16] S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving, “A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication,” *Journal of medical Internet research*, vol. 15, no. 4, p. e85, 2013.
- [17] A. Nurwidyantoro and E. Winarko, “Event detection in social media: A survey,” in *Proceedings of the IEEE International Conference on ICT For Smart Society (ICISS)*, pp. 1–5, 2013.
- [18] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, “Visual fusion of mega-city big data: an application to traffic and tweets data analysis of metro passengers,” in *Proceedings of the IEEE International Conference on Big Data*, pp. 431–440, 2014.
- [19] X. Wang, K. Zeng, X.-L. Zhao, and F.-Y. Wang, “Using web data to enhance traffic situation awareness,” in *Proceedings of the 17th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 195–199, 2014.
- [20] S. Zhang, “Using twitter to enhance traffic incident awareness,” in *Proceedings of the 18th IEEE International Conference on Intelligent Transportation Systems*, pp. 2941–2946, 2015.

- [21] R. Kosala, E. Adi, et al., “Harvesting real time traffic information from twitter,” *Procedia Engineering*, vol. 50, pp. 1–11, 2012.
- [22] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, “Twitcident: fighting fire with information from social web streams,” in *Proceedings of the 21st ACM International Conference on World Wide Web*, pp. 305–308, 2012.
- [23] T. Terpstra, A. de Vries, R. Stronkman, and G. Paradies, *Towards a realtime Twitter analysis during crises for operational crisis management*. Simon Fraser University, 2012.
- [24] N. Adam, J. Eledath, S. Mehrotra, and N. Venkatasubramanian, “Social media alert and response to threats to citizens (smart-c),” in *Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pp. 181–189, 2012.
- [25] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, “Semantics+ filtering+ search= twitcident. exploring information in social web streams,” in *Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 285–294, 2012.
- [26] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in *Proceedings of the 19th ACM international conference on World wide web*, pp. 851–860, 2010.
- [27] A. Qusef and K. Ismail, “Social media in project communications management,” in *Proceedings of the 7th International Conference on Computer Science and Information Technology (CSIT)*, pp. 1–5, July 2016.
- [28] J. Treboux, F. Cretton, F. Evéquo, A. L. Calvé, and D. Genoud, “Mining and visualizing social data to inform marketing decisions,” in *Proceedings of the 30th IEEE*

International Conference on Advanced Information Networking and Applications (AINA), pp. 66–73, March 2016.

- [29] S. Wan, C. Paris, and D. Georgakopoulos, “Social media data aggregation and mining for internet-scale customer relationship management,” in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 39–48, Aug 2015.
- [30] J. Burgess and A. Bruns, “Twitter archives and the challenges of ‘big social data’ for media and communication research,” *M/C Journal*, vol. 15, no. 5, 2012.
- [31] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, “Estimating mobile traffic demand using twitter,” *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 380–383, 2016.
- [32] A. Jackoway, H. Samet, and J. Sankaranarayanan, “Identification of live news events using twitter,” in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pp. 25–32, 2011.
- [33] S. Ishikawa, Y. Arakawa, S. Tagashira, and A. Fukuda, “Hot topic detection in local areas using twitter and wikipedia,” in *Proceedings of the IEEE ARCS Workshops*, pp. 1–5, 2012.
- [34] S. Petrovic, M. Osborne, and V. Lavrenko, “The edinburgh twitter corpus,” in *Proceedings of the NAACL HLT Workshop on Computational Linguistics in a World of Social Media*, pp. 25–26, 2010.
- [35] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis, “Bieber no more: First story detection using twitter and wikipedia,” in *Proceedings of the SIGIR Workshop on Time-aware Information Access*, 2012.

- [36] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, “Bad news travel fast: A content-based analysis of interestingness on twitter,” in *Proceedings of the 3rd ACM International Web Science Conference (WebSci)*, pp. 1–7, 2011.
- [37] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh, “Text and structural data mining of influenza mentions in web and social media,” *International journal of environmental research and public health*, vol. 7, no. 2, pp. 596–615, 2010.
- [38] A. Culotta, “Detecting influenza outbreaks by analyzing twitter messages,” *arXiv preprint arXiv:1007.4748*, 2010.
- [39] A. Alessa and M. Faezipour, “A review of influenza detection and prediction through social networking sites,” *Theoretical Biology and Medical Modelling*, vol. 15, pp. 1–27, Feb. 2018.
- [40] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, “Predicting flu trends using twitter data,” in *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 702–707, 2011.
- [41] R. Chunara, S. Aman, M. Smolinski, and J. S. Brownstein, “Flu near you: an online self-reported influenza surveillance system in the usa,” *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.
- [42] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, et al., “Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience,” *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 17–21, 2014.
- [43] Q. Zhang, C. Gioannini, D. Paolotti, N. Perra, D. Perrotta, M. Quaggiotto, M. Tizzoni, and A. Vespignani, “Social data mining and seasonal influenza forecasts: the

- fluoutlook platform,” in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 237–240, Springer, 2015.
- [44] J. Shaman and A. Karspeck, “Forecasting seasonal outbreaks of influenza,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 109, no. 50, pp. 20425–20430, 2012.
- [45] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, “Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports,” *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.
- [46] W. Shianghau and G. Jiannjong, “The trend of green supply chain management research a text mining analysis,” in *Proceedings of the 8th International Conference on Supply Chain Management and Information Systems (SCMIS)*, pp. 1–6, Oct 2010.
- [47] P. Meesad and J. Li, “Stock trend prediction relying on text mining and sentiment analysis with tweets,” in *Proceedings of the Fourth World Congress on Information and Communication Technologies (WICT)*, pp. 257–262, Dec 2014.
- [48] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.
- [49] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, “Knowledge management and data mining for marketing,” *Decision support systems*, vol. 31, no. 1, pp. 127–137, 2001.
- [50] M.-A. Mittermayer, “Forecasting intraday stock price trends with text mining techniques,” in *Proceedings of the 37th IEEE Annual Hawaii International Conference on System Sciences*, pp. 10–pp, 2004.

- [51] O. R. Zaiane, M. Xin, and J. Han, "Discovering web access patterns and trends by applying olap and data mining technology on web logs," in *Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries*, pp. 19–29, 1998.
- [52] D. Scanfled, V. Scanfled, and E. L. Larson, "Dissemination of health information through social networks: Twitter and antibiotics," *American journal of infection control*, vol. 38, no. 3, pp. 182–188, 2010.
- [53] K. Radinsky, S. Davidovich, and S. Markovitch, "Predicting the news of tomorrow using patterns in web search queries," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 01, pp. 363–367, 2008.
- [54] M. J. Paul and M. Dredze, "A model for mining public health topics from twitter," *Health*, vol. 11, pp. 16–6, 2012.
- [55] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 20, pp. 265–272, 2011.
- [56] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 755–760, 2014.
- [57] M. Santillana, D. W. Zhang, B. M. Althouse, and J. W. Ayers, "What can digital disease detection learn from (an external revision to) google flu trends?," *American journal of preventive medicine*, vol. 47, no. 3, pp. 341–347, 2014.

- [58] M. W. Davidson, D. A. Haim, and J. M. Radin, “Using networks to combine ‘big data’ and traditional surveillance to improve influenza predictions,” *Scientific reports*, vol. 5, p. 8154, 2015.
- [59] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen, “Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales,” *PLoS Comput Biol*, vol. 9, no. 10, p. e1003256, 2013.
- [60] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The parable of google flu: traps in big data analysis,” *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.
- [61] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, “Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic,” *PloS one*, vol. 6, no. 8, p. e23610, 2011.
- [62] S. Yang, M. Santillana, and S. Kou, “Accurate estimation of influenza epidemics using google search data via argo,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.
- [63] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, “Combining search, social media, and traditional data sources to improve influenza surveillance,” *PLoS Comput Biol*, vol. 11, no. 10, p. e1004513, 2015.
- [64] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using twitter,” in *Proceedings of the conference on empirical methods in natural language processing*, pp. 1568–1576, 2011.
- [65] D. A. Broniatowski, M. J. Paul, and M. Dredze, “National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic,” *PloS one*, vol. 8, no. 12, p. e83672, 2013.

- [66] X. Cui, N. Yang, Z. Wang, C. Hu, W. Zhu, H. Li, Y. Ji, and C. Liu, “Chinese social media analysis for disease surveillance,” *Personal and Ubiquitous Computing*, vol. 19, no. 7, pp. 1125–1132, 2015.
- [67] J. C. Santos and S. Matos, “Analysing twitter and web queries for flu trend prediction,” *Theoretical Biology and Medical Modelling*, vol. 11, no. 1, p. S6, 2014.
- [68] A. Lamb, M. J. Paul, and M. Dredze, “Separating fact from fear: Tracking flu infections on twitter,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 789–795, 2013.
- [69] V. Lampos and N. Cristianini, “Tracking the flu pandemic by monitoring the social web,” in *Proceedings of the 2nd IEEE International Workshop on Cognitive Information Processing*, pp. 411–416, 2010.
- [70] W. Xu, Z.-W. Han, and J. Ma, “A neural network based approach to detect influenza epidemics using search engine query data,” in *Proceedings of the IEEE International Conference on Machine Learning and Cybernetics*, vol. 3, pp. 1408–1412, 2010.
- [71] K. Byrd, A. Mansurov, and O. Baysal, “Mining twitter data for influenza detection and surveillance,” in *Proceedings of the IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS)*, pp. 43–49, 2016.
- [72] J. Ritterman, M. Osborne, and E. Klein, “Using prediction markets and twitter to predict a swine flu pandemic,” in *Proceedings of the 1st international workshop on mining social media*, vol. 9, pp. 9–17, 2009.
- [73] C. Corley, A. R. Mikler, K. P. Singh, and D. J. Cook, “Monitoring influenza trends through mining social media,” in *Proceedings of the International Conference on Bioinformatics and Computational Biology (BIOCOMP)*, pp. 340–346, 2009.

- [74] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994.
- [75] S. Grover and G. S. Aujla, “Twitter data based prediction model for influenza epidemic,” in *Proceedings of the 2nd IEEE International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 873–879, 2015.
- [76] J. Lessler and D. A. Cummings, “Mechanistic models of infectious disease and their impact on public health,” *American journal of epidemiology*, vol. 183, no. 5, pp. 415–422, 2016.
- [77] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, and A. Vespignani, “Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model,” in *Proceedings of the 26th International Conference on World Wide Web*, pp. 311–319, 2017.
- [78] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, “Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model,” *Journal of computational science*, vol. 1, no. 3, pp. 132–145, 2010.
- [79] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe, “A systematic review of studies on forecasting the dynamics of influenza outbreaks,” *Influenza and other respiratory viruses*, vol. 8, no. 3, pp. 309–316, 2014.
- [80] S. Deodhar, J. Chen, M. Wilson, M. Soundarapandian, K. Bisset, B. Lewis, C. Barrett, and M. Marathe, “Flu caster: A pervasive web application for high resolution situation assessment and forecasting of flu outbreaks,” in *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 105–114, 2015.

- [81] D. C. Liu and J. Nocedal, “On the limited memory bfgs method for large scale optimization,” *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
- [82] L. Tang, *Learning with large-scale social media networks*. PhD thesis, Arizona State University, 2010.
- [83] S. Volkova, *Predicting Demographics and Affect in Social Networks*. PhD thesis, Johns Hopkins University, 2015.
- [84] A. Mislove, S. Lehmann, Y. Ahn, J. Onnela, and J. N. Rosenquist, “Understanding the demographics of twitter users,” in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 11, p. 5th, 2011.
- [85] D. Ruths and J. Pfeffer, “Social media for large studies of behavior,” *Science*, vol. 346, no. 6213, pp. 1063–1064, 2014.
- [86] M. M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, “Population bias in geotagged tweets,” *People*, vol. 1, no. 3,759.710, pp. 3–759, 2015.
- [87] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29–39, 2000.
- [88] N. J. Sanders, “Sanders-Twitter Sentiment Corpus.” <http://www.sananalytics.com/lab/twitter-sentiment/>, 2011. [Online; accessed 20-October-2017].
- [89] Z. F. Dembek, K. Carley, and J. Hadler, “Guidelines for constructing a statewide hospital syndromic surveillance network,” *Morbidity and Mortality Weekly Report*, vol. 54, no. s21-24, 2005.
- [90] S. Bird, “Nltk: the natural language toolkit,” in *Proceedings of the Association for Computational Linguistics on Interactive presentation sessions*, pp. 69–72, 2006.

- [91] J. Singh and V. Gupta, “A systematic review of text stemming techniques,” *Artificial Intelligence Review*, vol. 48, no. 2, pp. 157–217, 2017.
- [92] T. Joachims, “A probabilistic analysis of the rocchio algorithm with tfidf for text categorization,” tech. rep., Carnegie-Mellon university, 1996.
- [93] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 427–431, April 2017.
- [94] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey, and others, “Textblob: simplified text processing.” <https://textblob.readthedocs.io/en/dev/index.html>, 2014. [Online; accessed 20-April-2018].
- [95] A. Al Essa and M. Faezipour, “Tweet classification using sentiment analysis features and TF-IDF weighting for improved flu trend detection,” in *Proceedings of the 14th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), Part I*, pp. 174–186, 2018.
- [96] A. Alessa, M. Faezipour, and Z. Alhassan, “Text classification of flu-related tweets using fasttext with sentiment and keyword features,” in *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 366–367, June 2018.
- [97] V. Vapnik, *Statistical learning theory*. Wiley, 1998.
- [98] S. Manocha and M. A. Girolami, “An empirical analysis of the probabilistic k-nearest neighbour classifier,” *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1818–1824, 2007.
- [99] D. Aha and D. Kibler, “Instance-based learning algorithms,” *Machine Learning*, vol. 6, pp. 37–66, 1991.

- [100] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.
- [101] R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
- [102] S. Ruggieri, “Efficient c4. 5 [classification algorithm],” *IEEE transactions on knowledge and data engineering*, vol. 14, no. 2, pp. 438–444, 2002.
- [103] G. H. John and P. Langley, “Estimating continuous distributions in bayesian classifiers,” in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, 1995.
- [104] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, “On combining classifiers,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [105] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, pp. 123–140, Aug 1996.
- [106] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [107] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [108] A. Al Essa and M. Faezipour, “Mapreduce and spark-based analytic framework using social media data for earlier flu outbreak detection,” in *Advances in Data Mining. Applications and Theoretical Aspects*, pp. 246–257, 2017.

- [109] G. P. Haryono and Y. Zhou, “Profiling apache hive query from run time logs,” in *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 61–68, 2016.
- [110] A. Verma, A. H. Mansuri, and N. Jain, “Big data management processing with hadoop mapreduce and spark technology: A comparison,” in *Proceedings of the IEEE Symposium on Colossal Data Analysis and Networking (CDAN)*, pp. 1–4, 2016.
- [111] R. Chunara, S. Aman, M. Smolinski, and J. S. Brownstein, “Flu near you: an online self-reported influenza surveillance system in the usa,” *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.
- [112] J. O. Rawlings, S. G. Pantula, and D. A. Dickey, *Applied regression analysis: a research tool*. Springer Science & Business Media, 2001.
- [113] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, pp. 199–222, Aug 2004.
- [114] M. H. Zweig and G. Campbell, “Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine,” *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993.
- [115] G. Louppe, “Understanding random forests: From theory to practice,” *arXiv preprint arXiv:1407.7502*, 2014.

APPENDIX A: More Application Data

More application data is collected for validation purposes. This data includes a set of Twitter posts that are collected during the period of four consecutive weeks from October 29, 2018 to November 25, 2018 within the boundary box of the state of Connecticut. The collected posts were passed on to the proposed framework for weekly flu-rate estimation. Figures A.1 and A.2 show the output of the framework based on different regression models. Using the linear regression model, the predicted weekly rates produced by the framework are highly correlated to the gold standard data (CDC) with 94.2% Pearson Correlation.

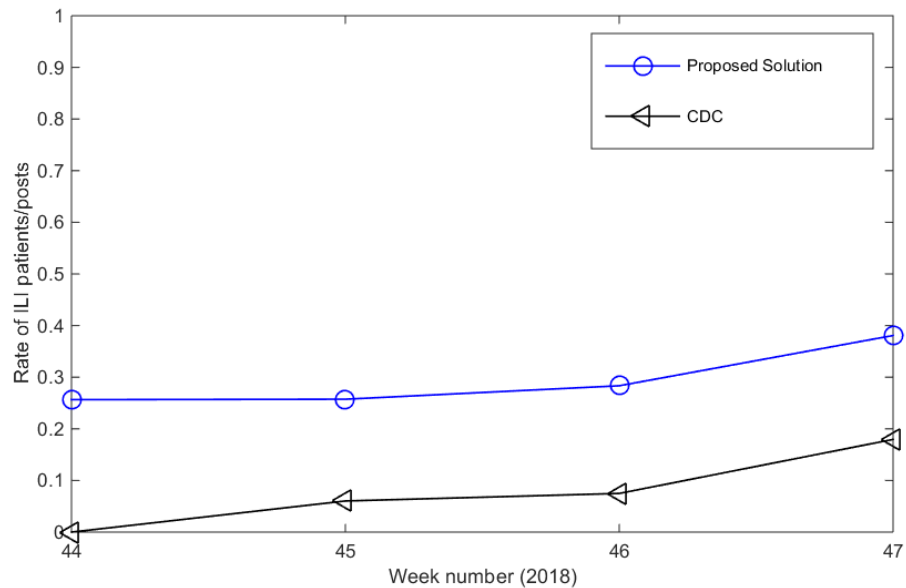


Figure A.1: Correlation between the proposed framework and CDC ILI rates (Nov. 2018)

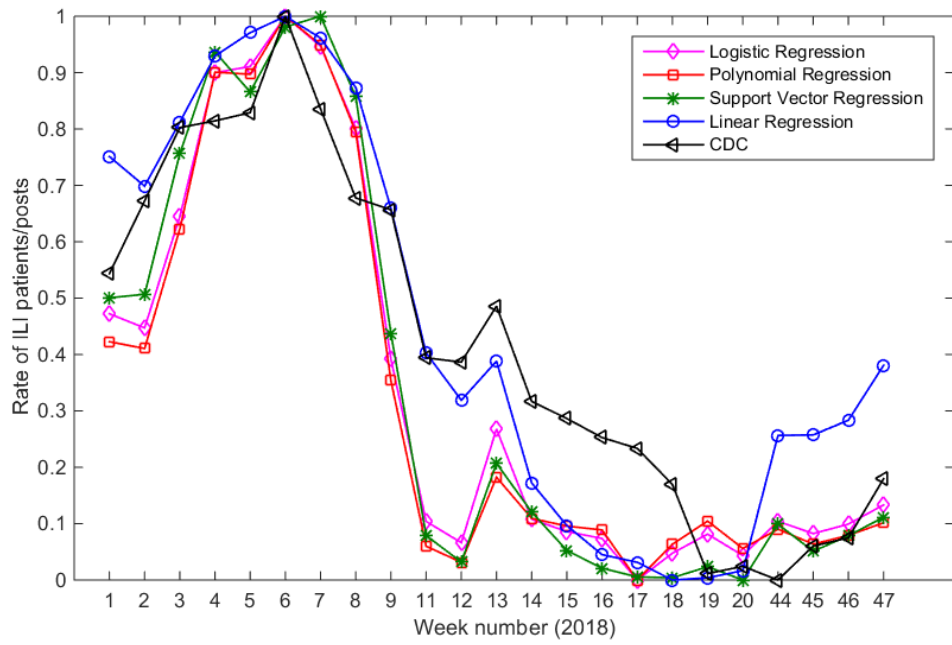


Figure A.2: Correlation between the proposed framework and CDC ILI rates using different regression models (Jan.-May, Nov. 2018)