November 2022

Information Dissemination and Perpetual Network

Harshit Srivastava
University of South Florida

Information Dissemination and Perpetual Network

by

Harshit Srivastava

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Ravi Sankar, Ph.D.
Nasir Ghani, Ph.D.
Ismail Uysal, Ph.D.
Manish Agarwal, Ph.D.
Inho Ra, Ph.D.

Date of Approval:
November 18, 2022

Keywords: Social Network, Cooperation Strategy, Influence Network, Optimization,
Graph Network, Binary Search Tree, Machine Learning

Dedication

This dissertation is dedicated to my parents, Shree Rajesh Kumar Srivastava and Anita Srivastava, my sister Surabhi Srivastava, my wife Sakshi and my dog Zeus and all my family and friends who supported and encouraged me during my studies.

Acknowledgments

I would like to sincerely thank my advisor, Dr. Ravi Sankar, for providing constant support and guidance with constant encouragement during my research. I have learned many lessons under his mentorship that would be helpful forever in my life. I am very grateful for the effort spent by my dissertation committee members Dr. Uysal, Dr. Ghani, Dr. Agarwal, and Dr. Ra in helping me reach my goal. I am also thankful to Dr. Wilfrido Moreno who provided valuable mentorship and support throughout these years. I would also thank Mr. Jack Dorsey and Mr. Noah Glass for providing twitter developer access, Ms. Beth Blauer and Mr. Dan Stone for giving access to John Hopkins COVID-19 data, and National Oceanic and Atmospheric Administration for providing data on Hurricane.

I am thankful to my parents and my family, for endless patience and encouragement throughout my life. I will forever be in your debt for your unconditional love, prayers and confidence in me.

Table of Contents

List of Tables

List of Figures

Abstract

Social networks have attracted increasing attention from both physical and social scientists. Social networks are essential elements in societies, serving as channels for exchanging various benefits, such as innovation, information, and social support. Moreover, research in social networks helps explain macro-level social phenomena, such as social polarization and social contagion. An understanding of social networks has significant implications, such as improving social welfare and political participation. Modeling social network formation has typically employed game theory or agent-based modeling. These studies typically propose simple and tractable micro-level rules for link formation mechanisms and show that these rules have implications for known macro-level properties. Statistics and econometrics have also used game theory to model empirical networks, but they typically have been focused on estimating and identifying the effects of interest, such as racial segregation. To date, these models have not been capable of accounting for the effects of broad heterogeneity among individuals; therefore, they lack predictive power for link formation in complex, real-world networks. This divergence is filled by cooperative techniques by applying game theory and casual inference techniques on severe weather prediction and disease spread in our work with consideration of heterogeneity, predictability of link formation and node characteristics.

The recent trend of dependence on the social network for information abstraction and propagation has a cumulative effect on critical response. The content and reliability of data are substantiated by acquiring data from a network of Twitter users. It captures the engaged multiple user behavior to formulate and diffuse the connected information across the channel. The objective is to identify a bridge between different data sources for event anomalies. This dissertation proposes a novel approach towards identifying the sublevel anomalies and predictive investigation towards the use of Twitter's social data in the extreme weather scenario and disease spread. We performed qualitative analyses by gathering data from social media and weather data websites and government websites. We also focused on a casual cooperation model outlined from social data with the help of survey data. The cooperation model encompasses cooperative attention to detect possible anomaly in an event. Various analysis methods are proposed to aggregate diffused information from the social network to generate influence data. This research also proposes the determination of spread through cooperative learning with the help of disease spread model. The analyses result further identify connected user acknowledgment for dominant information in the public domain. This information is mapped by applying a convolutional neural network for a physical sensor dataset to detect weather anomalies. Moreover, we exploited the causal inference technique to determine smart policy on influence data. The results show that our proposed method can predict critical events with high precision at the accuracy of 81% during extreme weather emergency scenarios specifically studied on hurricane IRMA.

Cooperative attention outlines the new paradigm for finding the cause of epidemic disease spread. It can be derived from the social data with the help of survey data. The cooperative attention enables the rationale to detect possible anomaly in an event by formulating the spread variable to determine disease spread rate decision score. This research proposes the determination of spread through cooperative learning with the help of disease spread model. We used game theory to define cooperative strategy and analyzed the determined dynamic states with the help of control algorithm. This model is a four-stage model to determine rewards by identifying the semantic cooperation with spread model to identify events, infection factor, location spread, and change in spread rate. Our model proposes new approach to define data cooperation by finding dynamic variable of spread and optimal cooperative strategy for the analysis of COVID 19 pandemic spread across Unites States. Our analysis successfully identified the spread rate of disease from social data with an accuracy of 81% and can dynamically optimize the decision model with $O(n^2)$.

The research also presents the development of systems for improved source selection in a process that creates real time categorization of events using only posts collected through various sensing applications that use social networks (such as Twitter or other mass dissemination networks) for reporting. The system recognizes critical instances in applications and simply views essential information from users (either by explicit user action or by default, as on Twitter) within the event and provides a textual description. As a result, social networks open unprecedented possibilities for creating sensing applications by representing a set of tweets generated in a limited timeframe as a weighted network for influence concerning users. Obtaining data from a network of social site users substantiates the quality and dependability of data. It collects many users' dynamic behavior to construct and disseminate related information across the channel. The goal is to find a link between various data sources for event abnormalities. By detecting sublevel anomalies using a convex optimization framework; the system recognizes rapid changes in the graphs' nodes and edge weights to pinpoint anomalies inside an event. This research investigates the merits of diversified data sources and developed graphical relations of information learning by correlation between social and different data sources by understanding the heterogeneity and homophily in a network with optimal accuracy.

Chapter 1: Introduction

1.1 Research Motivations and Background

A word of information cooperation and sharing is a paradigm in nature. Imagine if three people in a room connected to each other wearing headphones and can share all the information with each other. Either of the three will know if in their connection has issues but lack what event caused the connection issue. Detecting the event will not only help to solve connection issue but also can warn users. This imagination makes us dig into information dissemination over a social network where anomalies lead to events and occurrence of events leads to physical information.

Social network is attractive and have paved towards attractive structural information and relation properties to benefit in macro level social research and support in development of social tools to understand possibilities events and in making decisions during occurred events. Studies on network understanding partially fill the gap in understanding user node assortment and predictability in information linking and characteristics. This development of social media data paved new opportunities to fill the gaps through content-based retrieval with a prospect of predicting the given state of an event. This state of event detection opens paradigm of opportunities to detect trails of information for distinguishing between real or malicious events. The social world creates opening of expressing all emotions and experiences. This generates lot of emotional sensitivity of social media users on specific trend and news which creates gigantic data source in understanding behaviors. This creates a new world of news delivery of intentional or unintentional efforts to change beliefs and attitudes or behavior of users. Thus, the actors in the wide world are known as influencers who are admired or followed by other users. They can pass on the information with malicious intent or with good intent which creates instant trends. This creates a great hole in a new wide world of information and disinformation spread which sometimes can capitulate the malicious intent. Moreover, the question arises how we can detect the intent of maliciousness. Are our system able to aggregate the trends which can be used by terrorist to spread harmful information. Can we detect the social behavioral events and anomalies? To answer these questions, we must look data in a new perspective and context and explore the integrations through different data sources with the cooperative learning. Here when we talk about cooperation, we focus on node contagion, attention, and homogeneity information links by focusing only after creating social binary tree and causal inference. Where causal inference is attached with influence when a node X is influenced by node Y if there is a presence of information or factor of one or more nodes. This can be understood from causal influence as shown in Figure 1.1. A cooperative attention in a network constitute evolution and creates a behavioral reciprocity which is aligned with strategy-based game theory.

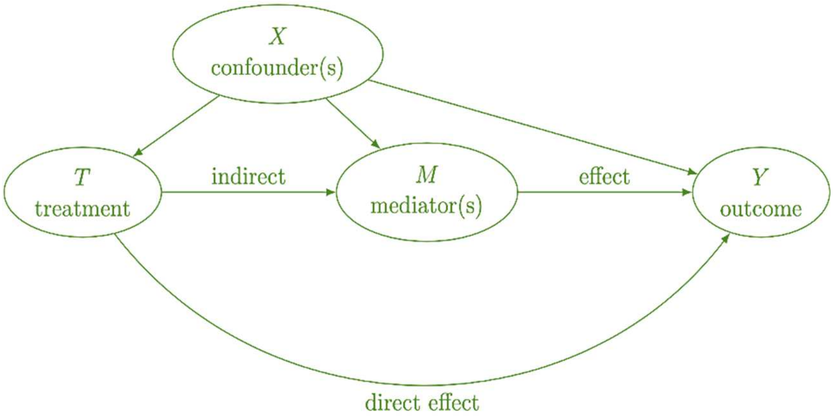


Figure 1.1 Causal Inference Example

Understanding the mechanisms of network formation is central in social network analysis. Network formation has been studied in many research fields with their different focuses; for example, network embedding algorithms in machine learning literature consider broad heterogeneity among users while the social sciences emphasize the interpretability of link formation mechanisms. A social network is a model that integrates multiple disciplines and retain both heterogeneity and interpretability. Each user encapsulates their features and use game-theoretical methods to model the utility of link formation.

The demand for social media such as Twitter as a source of current news and information grows exponentially. Several event detections approaches have been devised to deal with the velocity and volume of Twitter data streams. Engagement and understanding of social media formulation and network mechanism for the boundaries of information flow is a central process of social network analysis. Typically, users relate and interact via these spaces for example Twitter to form a relationship. Although most of these evaluate the suggested method, a comparative analysis is frequently absent. This research provides analyses and experimental examination of the state-of-the-art event detection algorithms for Twitter data streams. Several metrics are defined in this study to help the quantitative and qualitative comparisons. Microblogging service helps users send and read multiple real-time messages or news feeds getting popular as it connects with the globalized world, which we know as Twitter. The news feed composes a real-time instantaneous communication medium for everyday users. Twitter sends out more than 700 million tweets by more than 400 million everyday users. Having such a connected world tool in users' hand, event tracking such as a football game creates interest among the new generation. Consequently, users follow real-time news, business growth proposals, and stock and crypto market updates as social events. The emergent populous nature of the social world shows another interest related to emergency event tracking such as disasters, disease spread, and terrorist attacks. Users tweeting information regarding events on social media can generate a lot of perspective in events compared with the news media. However, a plethora of events and anomalies are reported every hour on

Twitter, and news portals miss the majority of personal users’ news. for example, family members update messages of their health about wellness or care for their loved ones.

Users' textual perceptions of their surrounding environmental conditions or emotions are people-centric sensing data and are often interested in detecting a sequence of crucial moments. This people-centric data sensing from social networks evolves for an event that spans time. However, due to the rate of data generation in the social world, analyzing and summarizing data topics for a specific event and sub-event anomalies is a challenge. This challenge is due to noisy data or content in heterogeneous social networks. This problem is widely studied [1]. However, the preciseness to detect all the events and subevents is low besides identifying essential moments. This shows that the fundamental requirement of social data analysis is to address unique requirements such as duration of tweets, emotions, and geographic location [2], which are dynamic in nature. Therefore, this makes a clear path for the summarization tasks consisting of two parts: (1) Detecting a stream of subevent anomalies in an event. (2) A generating module can categorize the events and provide a summary for subevents descriptions. In this thesis we propose a novel self-sufficient system that deals with the challenges mentioned above in social media event detections. Our system categorizes the data into actions, emotions, and locations and decomposes these events into time spaced graph since we assume that with time there will be the change in categories size and details (e.g., users use the same tweet to add more details). This decomposition is meta tasked through a common source, whether the users follow or are added as a friend [3]. We will also provide an overview of metrics on improving the credibility of social sensing. This social sensing can indeed provide better opportune characteristics in finding false negative rumors.

Communication is a core component in data prediction. The data analytics in recent years have seen unexpected changes from regular use in utilizing past data from events for predicting future outcomes or providing better data interpretation and analysis. Previously unnatural events were predicted from natural activities, such as company loss or gain from financial market data, whereas now, natural events found by using unnatural activities, such as social data usage to recognize emergencies. To formulate this change of utilizing virtual data in a specific depolarized way, several aspects like data aggregation, data correlation, mapping, and machine learning techniques are investigated. There are several limitations concerning data analysis and prediction, which makes the task of implementation bit tricky for short-term analysis.

The social data type enables us to collect people-centric sensing via social networking services (e.g., Twitter, Facebook). People-centric sensing data is the textual perception of users of their surrounding environmental conditions or their emotions. This data may correlate to the physical sensor information. It endorses the information provided by news agencies on social media and from weather agencies. This data not only can fill the data gaps but also give specific localized information while physical sensor data gives a more general overview of that area. It also gives information exchange between individuals to exploit optimized sensing. This optimized utility in advertising and media firms leverages consumer habits by using demographic social data [5]. The political analysts in elections have already tapped the sentiment of voters through social media to predict the results [6]. A better perspective can be gained for disaster management and prediction through data science. Even though, widely employed physical sensors can provide data in real-time with a possible tornado warning for an area about 15 minutes in advance, the same information can be obtained through social systems. For example, the tornado path was predicted through the use of social data in [7]. The data analytics have changed the pattern of interpretation in recent years through pair modelling of critical sentences, identification of paraphrases and important aspects of text entailment in many Natural Language Processing (NLP) tasks.

The important phase of these analysis is to not consider the impact of any two sentences, i.e., defining impact of each sentence separately [17] but their mutual relationship. This inherently develops limitations concerning data analysis and prediction for short-term analysis. This non-consideration of mutual influence dynamically contrasts the focus without changing the contexts. As humans if two people's arguments are presented, we extract word identities and relations to understand the entire scenario. Hence, the analysis veracity of a group of sentences becomes a challenge. This challenge entails the figurative language representation whose meanings are usually not concrete. This figurative language is engrossed as sarcasm which people use a lot on social media and represent negative feelings through positive words or vice-versa.

Detecting an event through a governed set of parameters results in detection of global event but a global event occurs because of or creates multi subevents. This detection of subevents is complicated in nature. For example, in a city pollution increases during a specific time period on every week that can be counted as a subevent and each of these can further have subevents like, increase in production of a factory or due to wind blowing or from other natural disasters. Thus, finding distinct subevents is necessary and some subevent cannot be a part of pollution but can directly or indirectly factor for the main event like, having a sports tournament, political rally, or protest. Each subevent should be detected to augment the exact map of any event.

Mapping and detecting these subevents with through contextual and sarcastic words and phrases requires a fine set of features to categorize the dynamics. This requires identification of important and influential users in a social network.

1.2 Applications Analysis

The primary purpose of both social networks and computing is to analyze the meaningful and empirical communication patterns of users. This essence begins with the inference structure of statistics and chances based on a setup of assumptions in the network. The unpredictability of the data distributions, as well as comparisons to those distributions, are generated by the topology of the network. The structure of the social network is formed by the connections between social users or between the various organizations. These connections are known as links or edges, and the nodes and vertices of the network are referred to, respectively, as vertices and nodes. These communication relationships are growing as a result of the interchange of varied information that is disseminated.

Network studies are constant in discovering patterns and connection validation in a self-organized structure where nodes establish and remove freely to reflect strong and weak connections and channels of information flow. Weak links often connect many user nodes for additional information. Therefore, in order to carry out the analysis on the network, it is necessary to ask the necessary questions and address any concerns that may arise. The most important questions are "how the network data can be sampled impartially" and "what is the probability that the observed patterns reflect the particular chance." For instance, if the users on the network are all from the same school, then the users from the other schools' networks should not be included into the chance of distributions since the probability of winning might be different for each institution [6].

When the nodes and clusters of one network are mapped onto those of another network, however, the network topology shifts in a way that might be described as a fold. Therefore, we consider the second network data to be physical data where statistical inference models are employed in order to generate stronger relationships or links between the vertices or nodes in the network. In addition, as the network expands, node-edge similarities and dissimilarities always create a shared information path for the users in the network, which in turn creates a higher chance of false-positive or vice versa analysis perspectives. This happens regardless of whether the similarities or dissimilarities are created by a node or an edge. As a result, the views of users and the engagement behavior of users reveal to be an essential cluster in differentiating the appropriate information flow. This problem is being addressed by a number of different groups using supervised learning or crowdsourcing, both of which supply limited but crucial information in a timely manner. As a result, it is very necessary to create a model so that one may obtain crucial knowledge without relying on the assistance of others.

1.3 Research Potentials

Cooperation is a vital component of social societies [1-9, 11-15], and it takes occurs when people endure difficulties in order to support the interests of others. There is evidence to suggest that individuals are influenced by their social connections, and as a consequence, emotions, ideas, and behaviors may spread across the links in their social network [15-28]. This evidence shows that people are impacted by their social contacts. As a direct consequence of this, the question of whether or not social transmission also plays a role in the evolution of cooperation has been explored. This topic is fascinating not just from a theoretical standpoint, but it also has the potential to have repercussions for therapeutic approaches that are geared at fostering cooperative behavior. On the other hand, it is notoriously difficult to differentiate homophily from contagion. Homophily is defined as "the tendency for people to develop and sustain connections with those who are similar to themselves" [23], [29], and [30]. Contagion, on the other hand, is notoriously difficult to differentiate from homophily. It is quite difficult to discriminate between homophily and contagion when utilizing observational data. This is due to the fact that homophily can occur on unobserved traits, which prevents conventional statistical control from being possible in network analysis as well as in any other context that utilizes observational data [30], [31]. In addition to this, this is why it is not possible to use homophily as a measure of similarity between individuals with different observed traits. However, these problems may be remedied by conducting investigations in a well-controlled environment such as a laboratory, where they can be closely monitored. In the controlled environment of the laboratory, the investigators have complete information and command over the interaction patterns of the subjects being studied.

Recent studies [32] have demonstrated that social contagion may, in fact, move from one person to another when people are working together on a project. The researchers used data from a well-known laboratory experiment in which participants took part in a public goods game in order to explore the social contagion of cooperation. The experiment was conducted in order to gather information on the social contagion of cooperation (a game-theoretic formalization of group social problems). At the conclusion of each round in which they were required to interact with a new group of unfamiliar people, the people who took part in the experiment were given the opportunity to decide how much money they would contribute to a collaborative project that would ultimately be of benefit to everyone in the group. When the participants were not given the option to choose their interaction partners, any possibility of their engaging in homophilic conduct was removed from the equation. Those participants, however, who were assigned to groups with other participants who had gave a sizable sum contributed considerably more in subsequent rounds when they were given the opportunity.

There is evidence to demonstrate that behavior in cooperative games is similarly contagious in fixed social networks because individuals in these networks are always forced to interact with the same neighbors, homophily cannot exist there because it would be impossible to coordinate interactions. Users deployed with multi-player techniques, repeatedly encountering the prisoner's dilemma in fixed networks with a variety of architectural layouts throughout the analogous work [33-35]. In the situation known as the prisoner's dilemma, cooperation is judged according to whether or not a person decides to cooperate or stray from the agreement. Therefore, in contrast to the public real word game, which views cooperation as a continuous variable, these games give the user node the ability to evaluate in a distinct manner whether or not selfish or cooperative actions are contagious. This is possible because the public real word game views cooperation as a continuous variable. In other words, selfish conduct was infectious; cooperators who were coupled with proportionately more defecting neighbors were more likely to flip to defection in subsequent rounds. This behavior was seen across the board in all of the simulations. On the other hand, cooperative behavior was not infectious: defections that were associated with relatively more cooperative neighbors did not raise the probability that the defector would switch to cooperation. Cooperative behavior was not viral. These literature studies provide more data suggesting that social behavior in a network in the context of cooperation may spread from user to user, and they extend the applicability of this result to fixed networks. In addition, they demonstrate that this finding is applicable to stationary network configurations. These findings also show that there may be variations among people in the degree to which behaviors of cooperation and selfishness are infectious to one another. In the experiment with the fixed network, the participants were informed not just of the decisions taken by each of their neighbors, but also of the total payoffs that were created by those choices. This was done so that the participants could better understand the implications of the experiment. It is possible that the availability of payment information impeded the development of cooperation since, on average, defectors performed better than cooperators did. Defectors who lived in areas with a large number of cooperative neighbors may have been motivated to switch to cooperation, but they may have suppressed this desire in the face of the knowledge that switching would result in a decrease in their payoff. Those who lived in areas with a large number of cooperative neighbors may have been motivated to switch to cooperation. Therefore, more research is necessary to determine whether or not cooperative behavior may be infectious even in the absence of monetary information.

On the other side, fluid social networks have another goal in mind, which is to recruit new partners for cooperative engagement. This is the case since fluid social networks are designed to be adaptable. People who (correctly) believe that they are more likely to form connections with cooperators when they themselves cooperate may be motivated to try cooperating even when their current interaction partners are relatively uncooperative. This is because they believe that cooperating will increase the likelihood that they will form connections with other people who also cooperate. This is due to the fact that there are persons who feel that if they collaborate themselves, it would make it easier for them to create relationships with others who also cooperate. Therefore, we may anticipate less of a correlation between the behavior of an individual's present neighbors and that individual's own future behavior in social networks that undergo frequent updates and where there is a big potential to recruit new cooperative partners. We may be able to anticipate something like this happening in the future.

In this article, we present a solution to test cooperative attentions to learn different actions and predict by asking how the spread of actions through different sources behaviors in social networks depends on the extent to which individual users have control over their network connections. This allows us to test these cooperative attentions to learn different actions and predict. We investigate this topic by using the most recent pandemic data of COVID 19, which was obtained from the Twitter Social Network and the Physical Data that was found on the Johns Hopkins coronavirus data source [42]. According to this study's findings, the degree to which individual nodes were granted conditional control over the network connections they were assigned differed greatly across different types of networks. Because this dataset contains comprehensive information about the history of network connections, we are able to take a longitudinal approach to identify social contagion across time even when homophily (based on network updating) is a possibility. This gives us the ability to take advantage of the information provided by this dataset. We examine how social interaction may help us predict how the spread will develop by using this dataset to explore how selfish and cooperative behaviors spread over time in social networks that have various rules governing their structural growth.

1.4 Research Contributions

If we look at social data with some boundaries in terms of events and contents, it gives a new perspective of real-time data analysis of different dynamic content through the utilization of popularity index in the information. However, when we analyze the data of physical sensors data fusion provides the knowledge about the target, and to gain better interpretation and understanding of the environment. Thus, the paradigm of fusion of physical and social information graph studied and solved where the data from physical sensor is mapped to the information of social sensor network. The physical data fusion contribution is incorporate multi-source data into the framework for the decision interpretation and provide mathematical fundamentals for data embedding, feature selection/extraction to reorganize input data and remove redundant information. Our main contribution of this work is to build tools to bridge the gap between social data information with physical data fusion with the help of graphs. This not only provides the elective representation but also user node-based process. This research provides techniques in building network tree of searching weighted categorized words, cooperative learning with causal interference, influence optimization and event and anomaly detection with the help of machine learning techniques.

We developed a graphical relations of information learning by correlation between social and different data sources. This is followed with the help of contextual information and integration of the abstracts in social data keywords. For example, a social user talking about pollution and effects in their life whereas physical data on pollution will give data points on how harmful the air is, in this both data sources talk about pollution, but one detects personal life issue and other is detecting health issues. Therefore, concept of extraction amplification perpetuated in this study and analysis. We worked on various dataset from different private and government entities of the 4-year period on different areas and focused on weather and pandemic data. A major part of the study is to study Twitter's social data and network. We built the general framework to analyze social and physical data with multi-strategy learning to maximize the learning environment and states for decision tools.

1.4.1 Extreme Weather Anomaly and Event Detection Tool

This analysis represents the predictive investigation towards the use of Twitter's social data and network utilization with the help of NodeXL [23] and NOAA's U.S. National Weather Service [24]. The approach is to combine two stages, i.e., two separate data analyses, in parallel. The first stage is social data analysis, which utilizes reinforcement learning method to quantify essential characteristics. The second stage is physical sensor data analysis, which utilizes historical data and real-time data. The critical aspects of this research are to quantify the social data with physical data gathered from Twitter [23] and NOAA [24] for hurricane disaster events like Harvey, IRMA, and Michael that occurred between 2017-18. The social data consist of approximately 2,100,000 tweets on a geographical basis, perform preprocessing and feature selection then create rules set using the defined attributes and apply classification and find the relative patterns of social data with physical data before, during, and after the events.

1.4.2 Disease Spread Anomaly and Event Detection Tool

The purpose of this research is to get an understanding of, and explore, cooperative strategy. In order to get a conclusion about the disease's dissemination, the collaborative research is carried out by connecting the diverse data sources of social networks and physical networks. We study this topic by using the most recent pandemic data of COVID 19, which was obtained from the Twitter Social Network and the Physical Data that was found on the Johns Hopkins coronavirus data source [42]. According to this study's findings, the degree to which individual nodes were granted conditional control over the network connections they were assigned differed greatly across different types of networks. Because this dataset contains comprehensive information about the history of network connections, we are able to take a longitudinal approach to identify social contagion across time even when homophily (based on network updating) is a possibility. This gives us the ability to take advantage of the information provided by this dataset. We examine how social interaction may help us predict how the spread will develop by using this dataset to explore how selfish and cooperative behaviors spread over time in social networks that have various rules governing their structural growth.

1.5 Dissertation Organization

Rest of the chapters in this dissertation are organized to provide necessary background and other information that can establish the importance of framework and the evaluation methods. In Chapter 2, the background information related to social network formulation and analysis, its development and implementation is covered with extensive comparison of event detection algorithms. It also covers the impact of social network analysis and addresses the lack of quantitative and comparative evaluation of event detection techniques by proposing several measures, both for run-time and task-based performance to detect events precisely.

In Chapter 3, the anomaly detection analysis is introduced with the objective of subevent detection using anoptimized strategy, where the influence score represents the anomaly score. This chapter provides a novel method for producing real-time categorization of events using solely Twitter tweets of all users.

In Chapter 4, the concept of information dissemination and the structure of identifying information and sentiments through location and user data profiles are presented. This relates to opportunistic sensing comprising social sensing, i.e., emotion and physical sensing, i.e., location. We introduced analytical algorithms to combine physical sensors data with social data pertaining to action, emotion, and location information of critical events during extreme weather emergencies.

In Chapter 5, a solution is provided for the interacted information which can be evolved to create a populous cooperation structure. One of the key aspects of the chapter is to understand the analysis of social network dynamics constituting cooperative attention and strategy. In this chapter. an essential way to utilize diverse data sources to find cooperativeness in a network is provided.

In Chapter 6, the evaluation and discussion of the results of this research work are summarized. The chapter also describes the overall contributions made to the research community in this topic by providing evidence of the drawbacks in the current methods and the impact and robustness of the proposed framework. We also propose future direction for improving cooperative and causal inference analysis.

Chapter 2: Review of Social Data Computation, Event Detection, and Information Dissemination Techniques for Diverse Data Sources

2.1 Background on Social Data Sources

Over the course of the past several years, a number of research papers on event recognition and tracking strategies for Twitter have been made public. As a result of this, a number of surveys have been developed to chronicle the most recent developments in the field. The research that [8] conducted covers ways for identifying natural catastrophes, traffic, diseases, and news events. In his study, Madani [9] provides solutions to the four issues of diagnosing health epidemics, detecting natural occurrences, discovering trending themes, and evaluating sentiments. The survey that was conducted by Bontcheva and Rout is a more generic one that discusses a variety of issues relating to making sense of data obtained from social media [18]. The article discusses a variety of topics, including users, networks, modeling user behavior, as well as intelligent and semantically-based information access. Under the heading "semantic-based information access," there is another part that provides an overview of event detection methods that are used in social media data streams. Techniques based on clustering, models, and signal processing are the three types of event detection approaches that are utilized by this organization. There is also a discussion of the many methods available for identifying sub-events. Last but not least, Atefeh and Khreich [19] provide a comprehensive assessment of event detection strategies. They accomplish this by providing a list of several methodologies that are arranged according to detection methods, tasks, event kinds, application areas, and evaluation metrics. Based on these surveys and the works, we are able to draw the conclusion that the majority of the existing approaches are assessed by utilizing ad hoc metrics in conjunction with manually labeled reference data sets. In addition, only a few of the researched tactics have been contrasted with other competing alternatives.

2.2 Event Detection Techniques

The detection of events is a difficult subject, especially when randomness and automation are involved. In order to recognize occurrences, it was necessary to recognize anomalies with both precision and memory. This analysis makes a comparison of the many application strategies that were tested during the Social News on the Web challenge [20] in regard to the solutions that were submitted, the metrics of accuracy and recall, readability, coherence/relevance, and variety. This examination was stratified in several manual and automatically classified categories. For the manual assessment, 11 teams were utilized. This option is one of the examples that have been provided to show event detection with regard to the various measurements and levels of complexity. In light of this, we offer a technical evaluation of various event detection and collaboration strategies found in the literature, using the Social News on the Web challenge as a basis.

Table 2.1 List of Event Detection Techniques

Applications	Papers	Measures
Disaster Management	Srivastava et al. [1]	Influence and Precision Score
	Sakaki et al. [2]	Precision and F-Score
	Li et al. [3] and Li et al. [4]	Precision Score
	Abel et al. [5]	Average Decision Score
	Adam et al. [6]	Average Decision Score
	Terpstra et al. [7]	Filtering of Data of 100K Tweets
	Nurwidyantoro et al. [8]	Survey of Techniques
	Madani et al. [9]	Survey of Techniques

Table 2.1 (Continued)

	Winarko et al. [10]	
	Aggarwal et al. [39]	Manual Tagging to get Precision Score
	Phillips et al. [23]	Average Decision Score
Disease Spread	Nurwidyantoro et al. [8]Madani et al. [9] Culotta [11]	Survey of Techniques Survey of Techniques Search of Correlation in Data
	Bodnar et al. [12]	Correlation
	Ritterman et al. [13]	Filtering of Data of 48 million Tweets
	Wakamiya et al. [14]	
	Asgari-Chenaghlu et al. [15]	
	Achrekar et al. [16]	Search of Correlation in Data
Information Spread	Alvanaki et al. [17]	
	Bontcheva et al. [18]	Survey of Techniques
	Atefeh et al. [19]	Survey on Evaluation Metrics
	Papadopoulos et al. [20]	Event Detection by Correlation
	Sankaranarayanan et al. [21]	Crawling and Spread metrics
	Walther et al. [22]	False Positive Detection
	Meladianos et al. [24]	False Positive Accuracy
	Guille et al. [25]	Precision and F-Score with Manual Tagging
	Petrović et al. [26]	Average Precision Score with Manual Tagging
	Marcus et al. [27]	Precision Score
	Popescu et al. [28]	Precision and F-Score
	Ishikawa et al. [29]	Crawling and Spread metrics
	Nishida et al. [30]	Filtering of Data of 300K Tweets
	Aiello et al. [31]	Precision and F-Score

Table 2.1 (Continued)

	Petrović et al. [32]	Manual Tagging to get Precision Score
	Osborne et al. [33]	Time Taken for Information Spread
	Ishikawa et al. [29]	Crawling and Spread metrics
Business Analytics	Benhardus et al. [34]	Precision and F-Score
	Cataldi et al. [35]	Filtering of Data
	Lee et al. [36]	Average Precision Score with Manual Tagging
	Mathioudakis et al. [37]	Crawling and Spread metrics
Others	Allan J. [38]	Filtering, Crawling and Correlation
	Aggarwal et al. [39]	Manual Tagging to get Precision Score
	Cordeiro et al. [40]	Filtering and Reduction of Noise
	Li et al. [41]	Precision Score
	Osborne et al. [42]	Time Taken for Information Spread
	Ritter et al. [43]	Manual Tagging to get Precision Score
	Bahir et al. [44]	Filtering of Data
	Martin et al. [45]	Recall Metrics of Activities
	Parikh et al. [46]	Filtering of Data and Manual Tagging
	Abdelhaq et al. [47]	Filtering of Data
	Weiler et al. [48]	Survey of Techniques
	Corney et al. [49]	Survey of Techniques
	Ifrim et al. [50]	Filtering of Data
	Zhou et al. [51]	Filtering of Data and Manual Tagging
	Thapen et al. [52]	Filtering of Data and Manual Tagging
	Monmousseau et al. [53]	Filtering and Reduction of Noise
	Blei et al. [54]	Concepts of Detection
	Hoffman *et al.* [55]	
	McCreadie et al. [56] McMinn et al. [57]	

Table 2.1 (Continued)

	Cilibrasi et al. [58]	Concepts of Detection
	Wu et al. [59]	
	Khatoon et al. [62]	
	Bellatreche et al. [63]	
	Savic et al. [64]	
	Jones [65]	
	Li et al. [66]	

The literature that is currently available on event detection approaches for Twitter is outlined in Table 2.1 and organized according to the applications. The table further categorizes the items on the list according to the many strategies, approaches, and metrics for the collecting of social data that pertain to the applications. In the measures column, a listing of the many metrics that were used to evaluate the various techniques can be seen. The majority of research investigations make use of the precision metrics for event detection (32 of 66). In addition, some studies compute the F1 score, which is comprised of the average accuracy as well as the region that is beneath the receiver-operating curve (AROC). Only two investigations, one conducted by Alvanaki et al. [17] and the other by Parikh and Karlapalem [46] used a measure (RT) to evaluate the run-time performance of their approach. This is despite the fact that measures to evaluate the task-based performance of a technique are relatively prevalent. In addition to these traditional methods of measurement, numerous brand new methods of measurement were devised. Alvanaki et al. [17], for instance, measure relative accuracy, whereas Li et al. [3-4] and Guille and Favre [25] evaluated the duplicate event rate (DER) of their respective.

These methods are entirely reliant on the collecting of data in accordance with a predetermined set of two particular rules: Users specifically direct the following: This approach of collecting data chooses a default group of users who immediately follow the users' streams of data regardless of locales, trends by worldwide or geographical region, or other factors. The data size of social media platforms where Twitter's social data was gathered is particularly determined by the guidelines outlined above. The assessment methods that were utilized in the studies that are described in Table 2.1 and Figure 2.1 were determined by the magnitude of the social data. The size of a tweet can range anywhere from one hundred thousand to one hundred million tweets. For instance, the Twitter API was used to crawl one million tweets with pre-defined tags and keywords from January 2017 until January 2018 [1], whereas Ritterman et al. [13] and Sankarnarayanan et al. [21] tracked domain-specific data and news publishing data for a period of two months each by using the user stream API of Twitter. [13] and [21] respectively.

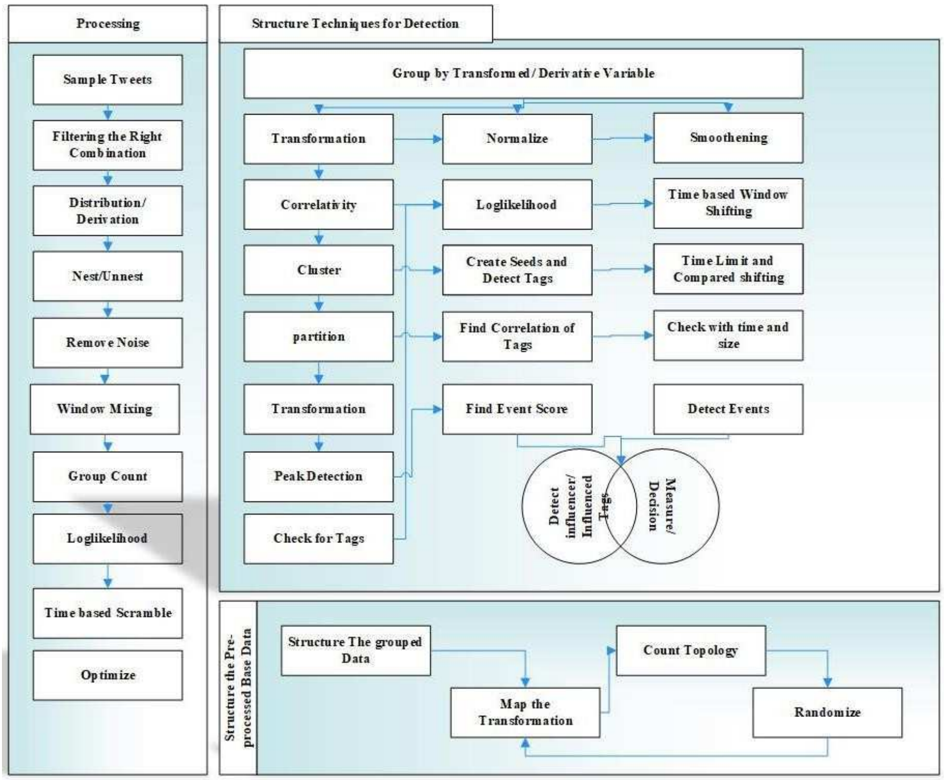


Figure 2.1 Analysis Techniques and Design Flow

This chapter is explaining the order of applications mentioned in Table 2.1, where we start by the evaluation of applications in disease spread, disaster management, information spread, business analytics, and the rest of them grouped as others. The event detection techniques presented in the surveyed papers were evaluated based on the metrics shown in the measure's column. We are not evaluating the accuracies of the current surveys as the studies presented have high sensitivities in their performances in detecting the occurrences of an event. For example, the findings are limited concerning evaluators, and most of the work is manually labeled collection of occurrences. These manually tagged occurrences are mostly voluntarily verified to examine the differences between actual and fake detections.

Cullota [11] and Bodnar et al. [12] validated the model for influenza disease detection by using regression techniques to detect the prevalence of Disease while having high sensitivity for the data sets. In [2], 500 thousand tweets were crawled through keyword-specific API for 28 months. This collected data was correlated with various proposed models to achieve an accuracy of 0.78, whereas Ritterman et al. [13] explored the hypothesis that social media encodes a belief that many people make a factual statement of utilizing the stock market prediction to predict swine flu spread. This is evaluated using the classic classification technique with regression, but the model failed to detect the noisy information leading to false event prediction. In [14], a more objective approach was utilized to detect the influenza spread by utilizing correlation of location aggregation and social media data, resulting in a high spread probability of spread detection, but the model was highly dependent on location estimation variables. Whereas Asgari-Chenaghlu et al. [15] proposed a transformer encoder to detect COVID-19 by utilizing social media data by converting tweets into universal sentences and then utilizing clustering techniques to Covid spread with a small set of data from the period of March 2020 to April 2020.

It is worth noting that keyword-specific and domain-specific event detection methods and algorithms are often evaluated by comparing the real-time data statistics. For example, the COVID statistics of John Hopkins may be used as the baseline for evaluating the spread of disease in different geographical locations or zip codes. This real-time data statistic is worded as manually labeled data collection. Achrekar et al. [16] collected the surveyed data into 1000 clusters and compared them to data that is manually labeled by the human evaluators. This categorization helped identify false positives and negatives with 68% and 32%. The eventdetection historically always contains clutters in the result due to high sensitivity, which was observed during the detection of disease spread event by correlating it to with the baseline surveydata. In [17], a model was presented to evaluate geo-location-specific tweets' event detection with crawled data of 22 million tweets. The model was highly accurate in detecting anomalies with an accuracy of 0.89 when comparing manually labeled data but could not detect the type of anomaly and failed to pinpoint the events correctly, with accuracy dropping below 0.1. This problem was solved in [1], for the detection of types of anomalies during the hurricane *IRMA*. Srivastava and Sankar [1] pointed out crucial steps to crawl data, detect events, and label them as influencers. These influencers were then processed to detect the type of events with high precision of 0.7. However, Aggarwal et al. [39] provided another perspective of disaster event detection by dividing their evaluation into two parts. In the first part, they showed a case study toevaluate the unsupervised model of their event detection technique. In the second part, they used a self-generated ground truth to evaluate the precision and recall of the supervised model for two sample events (*Japan Nuclear Crisis* and *Uganda Protest*). For the first event, they obtained a value of 0.525 for precision and 0.62 for recall), while the precision value was 1.0 and the recall value was around 0.6 for the second event. However, it is important to note that they work with highly pre-filtered data, which does not constitute a real-life evaluation of event detection techniques. Phillips et al. [23] provided a promising weather forecast decision method fortornadoes through Twitter data and showed a high correlation to detect tornadoes by just utilizingsentiment analysis and comparing with physical data. Interestingly, in [25], work on social sensors utilized the disaster detection technique in Sakaki et al. [2] to achieve better detection. The system was tailored specifically for real-time event detection; for example, ~600 samples were trained as base sample type for an earthquake and utilized classification methods to effectively detect earthquakes with the precision of 0.66. This real-time event detection and identification approach corroborate with Becker et al. [61], who trained the system for multiple weeks to evaluate the accuracy of the identified events and compared the results with manual evaluators in clusters. The evaluation produced a higher accuracy in detection with low sensitivity towards the data structure. Their work created a standard where the application detected a story as an event. The model presented high labeling capability over a dataset of 163.5 million tweets over six months. The model reduced the false positives and negatives over a streaming API.

Information spread and business analytics applications comprise tools and techniques in identifying the popularity of content and topics. The definition of popularity is identified on a real-time and hourly basis using the trends crawling technique. In [25], popular indexed topics from Twitter were extracted using an algorithm that identifies the probable popularity of a topic in the United States. The results of the event detection technique were compared with respect to Twitter trends. These trends can be constantly crawled with the help of Twitter API [4]. The technique was tested in two parts. In the first part of the test, no human interactions or manually labeled data were included, resulting in a very low average precision score of 0.25. In the second part, manually labeled data was added and the average precision score increased to 0.65. However, papers [3] and

[4] presented the evaluation on event detection with clustering of wavelet-based signals (EDCoW). The EDCoW builds signals for individual words by applying wavelet analysis on the frequency-based raw signals of the words. It then filters away the trivial words by looking at their corresponding signal autocorrelations.

Li et al. [3] utilized EDCoW to detect events daily and the evaluation of this detection technique showed high sensitivity on a highly restricted data set. The data set contained only tweets from the top 1000 users who had large followings in Singapore in 2010. This restriction was capped with a filtering technique to gather unique words. After filtering, the data setcomprised of 8140 words was used to evaluate their method which resulted in a precision of 0.76. This evaluation was compared more qualitatively rather than focusing on a quantitative evaluation. However, Li et al. [4] compared the data results and precision score by usingsegment-based event detection technique. This work tested the model on the same data set collected in [25]. The segment-based event detection technique resulted in a better precision score of 0.86, but it should be noted that [4] reported a low recall score and reduction by 25% compared to the recall score reported in [25].

TwitInfo is a tool presented in [26] for aggregating and visualizing microblogs for event exploration. Their evaluation used manually labeled events from soccer games and they also utilized geological events to detect disaster occurrences. For soccer game events, they scored 0.77 in both precision and recall (17 of 22 events found). For major disasters, the score was 0.14 (6 out of 44) for precision and 1.0 for recall (5 out of 5). Therefore, they concluded that their peak detection algorithm identifies 80–100% of manually labeled peaks. Popescu et al. [28] evaluated their work on extracting events and event descriptions from Twitter against a manually classified gold standard of 5040 snapshots, which were classified as events (2249) or non-events

(2791). Their technique, which was called Event Basic, scored 0.691 for precision, 0.632 for recall, 0.66 for the $F1$ score, 0.751 for average precision, and 0.791 for average region of convergence. Their extension of Event Basic, which was called Event Aboutness, did not show any improvements in the results with its scores being almost the same.

Alvanaki et al. [17] carried out an analysis that was predicated on the judgments of human assessors with regard to whether or not a reported result constitutes an event. For the purpose of this user research, they developed a website that presented users with the results of both their ENB and the previously established TM approach []. After that, users were given the opportunity to assess whether or not they consider a result to be an event and mark it accordingly. They were the first set of researchers to evaluate run-time performance in addition to analyzing the precision and relative accuracy of these procedures. The following are the outcomes that they achieved. ENB performed noticeably better than TM did in terms of accuracy. ENB was able to identify 2.5 out of 20 occurrences, but TM was only able to identify 0.8 on average. They assessed the connection between the parameters of the two different strategies and the increase or reduction in execution time in order to evaluate the run-time performance of the program. As a consequence of this, it is challenging to draw conclusions that are both useful and generic based on these metrics. The same can be said for the measurements of their relative accuracy, which were carried out independently for ENB and served solely to highlight the interaction of various parameters.

In their work, Osborne and colleagues [33] published the results of a first investigation on the latency for event detection based on several sources. They measured performance by calculating the average distance between each time-aligned Twitter first story and the title of the nearest neighbor Wiki page. This was done in order to evaluate how well the system worked. They came to the conclusion that there is a time lag between when events break on Twitter and when they break on Wiki, with Twitter having the edge. A technique for the extraction of open-domain events in Twitter was given by Ritter et al. [43]. In their study, they established that their method boosted precision and recall in comparison to a baseline method. This was proven by comparing their method to another method. In addition, in order to demonstrate the accuracy of the findings that were acquired, a selection of the retrieved happenings of the future was laid out in the format of a calendar.

In the context of the social sensor project, Aiello et al. [31] compared six different topic detection methods (BNGram, LDA, FPM, SFPM, Graph-based, and Doc-p) using three different Twitter data sets related to major events. These Twitter data sets differ in their time scale as well as the rate at which topics change. They came up with three different criteria for determining scores: subject recall, keyword precision, and keyword recall. They found that the BNgram approach consistently produced the best results for subject recall while still maintaining a pretty high level of precision and recall for keywords. They also noted that traditional methods of topic detection such as LDA function quite well on extremely concentrated events, but their performance was significantly worse when evaluating more 'noisy' events. This was one of their findings.

A subsequent piece of work that was presented by Martin and colleagues [45] contained an evaluation that was quite similar. In addition, they made an effort to determine the optimal slot size for the BNgram methodology and the optimal combination of clustering and topic ranking methods. As a result, it did not actually offer anything to the issue of comparative evaluations of different event detection methods. On the other hand, they found that the outcomes were different depending on which of the three data sets they utilized. One distinction stands out as particularly notable: participants' ability to recall information about football was significantly greater (over 90%) than their ability to recollect information about politics (about 60–80%). As a consequence, they came to the conclusion that the outcomes of an assessment are also dependent on the activities that take place inside the time frames that were investigated.

One of the data sets was supplied by the VAST Challenge in 2011, while the other was released in January 2013 by users residing in the United States. Parikh and Karlapalem [46] assessed their system (ET) on both sets of data. They employed the identical definitions of accuracy and recall as those found in [3-4] for the purposes of their evaluation. ET found a total of 23 occurrences in the VAST data set; however, only two of those events were considered to be unimportant or minor. As a result, the value of the accuracy was 0.91, and the recall was 21. ET retrieved a total of 15 events for the second data set, out of which only one event was not connected to any genuine occurrence, resulting in an accuracy of 0.93 and a recall of 14. It is important to keep in mind that the recall in this scenario is simply stated as the number of occurrences that were considered to be "excellent." They reported an execution time of 157 seconds to identify events from a total of 1,023,077 tweets, which translates to a throughput of 6516 tweets/seconds in order to quantify the performance of ET. This was done in order to demonstrate how well ET works.

MABED is an anomaly-based event detection approach that was suggested by Guille and Favre [25], and they called it after it. It was designed for Twitter. They carried out studies using Twitter data in both English and French simultaneously. During the course of their investigation, MABED was evaluated with ET [46] and TS [54]. The findings suggest that using MABED resulted in more accurate event recognition and greater resilience when dealing with noisy information on Twitter. In addition, when it came to the pre-filtering of data type mentions, such as the sign "@," MABED demonstrated superior performance in comparison to ET and TS. The authors also showed that MABED performed better than ET and TS in every single one of their experiments.

Meladianos et al. [24] provided a fairly stringent analysis of their methodology as it was applied to occurrences that occurred during soccer matches. They established a baseline for their analysis by searching a sports website in search of live game reports. The trials shown that their method is much superior to the baseline methods in terms of its performance on the sub-event identification job and its ability to provide quality summaries. In addition to this, their system was successful in identifying the majority of the critical sub-events that occurred throughout each match. Last but not least, Monmousseau et al. [53] gave a transportation viewpoint owing to the propagation of the illness.

Table 2.2 Cooperation and Detection Comparison

Papers	**Cooperation and Detection Techniques**	**Sensitivity/Accuracy**
Mccreadie et al. [56]	Data Distribution	Sensitivity -- 0.3
Becker [61]	Skewed Compilation	Sensitivity -- 0.43
Petrović et al. [26]	Diverse Classification	Accuracy -- 0.27
Papadopoulos et al. [20]	GT was trained	Accuracy -- 0.59
	Identifiers	Sensitivity -- 0.18
Aggarwal et al. [39], Petrović et al. [32], Allan [38], Guille et al. [25]	Event from Tagging	Accuracy -- 0.27
Allan [38], Blei et al. [54], Jones [65]	Tracking with Specificity	Accuracy -- 0.56

In this chapter, four passenger-centric indicators are extracted from Twitter in order to identify essential events that will take place between February 2020 and March 2020. This event detection is proposed by sensing empathy and mood of passengers for the transportation firms during illness spread when the precision to detect empathy was not supplied. During this time, the precision to detect empathy was not provided.

2.3 Cooperation in Event Detection Techniques

In this chapter, we address the challenge of developing universal evaluation metrics that may be implemented in a variety of event detection systems in order to make comparisons between them. Figure 2.2 provides an explanation of the evaluation method, with illustrations of the pre-processing and event detection methodologies serving as examples of the design flow of processes. The diagram illustrates the classification of influential factors and the optimization of decision-making with the assistance of a certain number of data samples. The figure blocks are packed together as a process flow in the form of three different categories: structural procedures, base processing, and pre-processing. To generate a randomization of sample sizes and proportions, preprocessing must first be organized to structure the data necessary for the process flow. In the end, these data with a set sample size were analyzed for optimum tagging in order to identify influential events and anomalous occurrences.

In the process of event detection, other attempts are applied that test out a cooperative method. The practice of human modification via tagging's on clustered data sets is used to do a quick analysis of these assessment approaches on Twitter-related analytic data works. In the next part, we will provide an explanation of a series of works that offer labeled reference data sets. These works will be presented in chronological order. For instance, McCreadie et al. [56] created a collection of 16 million tweets over the course of 14 days. [Citation needed] As a direct consequence of this, the proposed corpus consisted of an average of 50,000 tweets for each hour. We made the assumption that only 4.8 million of the corpus' tweets are in English since there was no attempt made to filter the tweets based on their language, which would have likely preserved 30 percent of the tweets. In addition, their list of reference subjects for the two weeks, which includes 49 different themes, is quite limited, and no explanation was given as to how these subjects were selected. Last but not least, due to the fact that the primary purpose of this corpus was to perform ad hoc retrieval tasks, it is not well suited for conducting large-scale evaluations of event detection strategies.

A Twitter corpus consisting of roughly 2.6 million tweets that were gathered in February 2021 was created by Becker et al. [61]. The purpose of this data collection was to only use their own approach to identify and categorize the occurrences for the purpose of conducting a study of the impacts of gender. The corpus is excessively biased in favor of their method and does not lend itself well to a comprehensive examination. In addition to this, there was no list of reference events provided, and the data set was geographically restricted to only include tweets from people based in the United States.

Petrovic et al. [32] provided a corpus of 50 million tweets that were extracted, randomly sampled, and tagged from the Twitter data stream between July and mid-September 2011 and is reused to compare the analysis results. This corpus was taken from their earlier work [26] and is used to compare the results of the analysis. The findings of this investigation led to the identification of 27 occurrences over the entirety of the time period. Because there were so few tagged events, this identification research was carried out because comparing numerous event detection algorithms is challenging, especially when the methodologies that are applied are so varied.

Three different corpora were supplied by Papadopoulos et al. [20] for the purpose of developing, training, and testing an event detection algorithm. The development data set had 1,106,712 tweets that were gathered during the presidential election in the United States in 2012 [30]. Applying filtering criteria to keyword and username combinations resulted in the creation of the training data set. The phrases "flood," "floods," and "flooding" were chosen as the keywords, and a list of "newshounds" was compiled by filtering the usernames of the users. Twitter users that often provide updates regarding recent happenings or breaking news are known as "newshounds." As a filter query, we decided to choose a total of 5000 newshounds with a concentration on the UK. For the testing data set, the same user filter was utilized; however, the keywords were modified to include the sets "Syria," "terror," "Ukraine," and "bitcoin." In addition, a ground truth comprised of 59 topics taken from UK media stories was compiled in order to facilitate the collecting of 1,041,062 tweets over the course of 24 hours.

McCreadie and colleagues [56] presented an approach for the creation of a corpus that could be used to test different event detection systems. They created a collection of potential events together with a list of connected tweets by combining Wikipedia's information with two current state-of-the-art event identification techniques [3] and [45]. The completed corpus includes around 120 million tweets and information on more than 500 events and spans a period of four weeks. However, events were reported in prose, which means that it is not possible to quickly and automatically compare the findings of the different event detection systems to the events themselves because they were presented in prose. Because Twitter's terms of service do not allow for the tweets themselves to be redistributed, all of these companies only consist of lists of tweet identifiers. This is a key point to keep in mind. In order to make use of a corpus, the matching tweets need to be crawled, which is an arduous process that can also lead to errors because certain tweets may no longer be accessible. For instance, the organizers of the 2014 SNOW challenge [20] were only able to crawl 1,106,712 of the initial 3,630,816 tweets that were included in the data collection for the 2012 US Presidential Election [30], which was described above. We employed the script method to download the corpus that was presented in McCreadie et al. [56] so that we could evaluate the usefulness of these collections of tweet identifiers. [56] The default limit for the number of queries that may be sent using the Twitter API to crawl tweets is set at 180 questions per 15 minutes. It is possible to retrieve a batch of one hundred tweets using just one search query. As a result, it is feasible to crawl 18000 tweets in each 15-minute window, and it would take about 6666 windows with an estimated total response time of 100000 minutes (approximately 1666 hours or 69 days) on a single system to crawl all of the tweets that are present in the collection.

2.4 Evaluation and Comparison of Event Detection Techniques

In order to actualize streaming implementations of cutting-edge event detection algorithms for Twitter, we made use of a technology called runtime-based Niagarino (63), which is a data stream management system that was developed and is maintained by our research group. The major objective of Niagarino is to develop research systems that are both simple to operate and capable of being expanded, specifically geared for streaming applications such as those described in the aforementioned article. Because of the operator-based processing architecture, our systems are designed to be flexible and extremely easy to configure. For instance, we can design the methods so that they report the same number of events, each of which is denoted by a primary event term and four event description terms that follow it. The results of analyzing run-time performance and memory consumption may be compared fairly if a comparable implementation is used. This is one of the advantages of utilizing a similar implementation. The following arithmetic operators are currently supported by it. In the end, query results are reported by sink operators that have no outbound streams. During the pre-processing phase, any tweets or retweets written in a language other than English will be deleted. After then, the terms associated with the remaining tweets are tokenized and unsettled. In addition to this, it gets rid of sentences that may be categorized as either stop words or noise (e.g., too short, invalid characters, etc.).

Blei et al. [54] exploited the probabilities of terms in texts to combine those phrases that had the highest possibility of being together and did so by putting them in categories according to their probabilities. It was achieved by employing Latent Dirichlet Allocation (LDA) and relative temporal association [10] together by applying its user-defined function operator. Both of these techniques were used in conjunction with one another. This is due to the fact that LDA is often utilized for topic modeling, with the purpose of connecting a topic with an event. This method allows the user to customize a variety of aspects, including the number of subjects, the amount of words allocated to each topic, and the number of rounds of the probability modeling process. Because there are so many terms that are repeated in tweets throughout the course of the time period, we consider this technique to be inadequate for event detection and have so designated it as the baseline method. After beginning with a grouping operator, the following step in the other four fundamental operations is to go on to the selection operator. Form Regroup (FR) generates 'events' by selecting five phrases at random from all of the various words included within a time frame. The major event term in a Reform Event (RE) is selected in the same manner as it is in a FR, but the associated event description terms are comprised of the four words that appear most frequently in conjunction with the primary event term. Both approaches result in N events being produced for each time frame. Both the Top N and Last N approaches are founded on the Inverse Document Frequency (IDF) [8, which extracts information and feelings from the entirety of a tweet rather than classifying it according to a predetermined set of categories. The Information Defense Force selects a single phrase from the clustered samples of tweets and assigns a score to each word in that phrase based on all of the individual words and keywords present in the time window of the samples. While Top N selects the N words that have been used the most frequently, Last N chooses the N terms that have been used the fewest times. Both report the event words that were specified, in addition to the top four keywords that occur the most frequently. Other techniques that have been presented for identifying events in Twitter data streams were added into our system in addition to these essential notions that we started with. We put all of these strategies into practice to the best of our abilities, based on the information that was presented in the preliminary research papers.

The first method, known as log-likelihood ratio (LLH), is a reimplementation of Weiler's [48] work and is carried out in the form of an LLH user-defined function that is applied to a grouped collection of words within a time frame. In contrast to the initial technique, which identified events based on pre-defined geographical regions and bigrams, we limited our analysis to single words alone. The approach computes a measure that is based on the shift in IDF values of single phrases that occur between successive sliding window pairs. At first, the IDF value of each word included within a single window is continually computed and compared to the average IDF value of all of the phrases contained inside that frame. Terms are removed from consideration if their IDF values are higher than the average. In the subsequent phase, a window with the size s1 and moving with the range r1 is going to be created in order to compute the shift from one window to the next. In this stage, the shift value is verified once again against the average shift of all terms. Terms whose shifts are greater than the average are the only ones that are preserved after this step. In the final stage, a new sliding window with dimensions of s2 and a sliding range of r2 is created. The total shift value is determined by computing the sum of all of the shift values that are associated with the sub-windows that make up this window. If this total shift value is higher than the previously defined threshold, the phrase in question is identified as an event and published alongside its top four co-occurrence terms.

In our assessment, we have pre-defined a number of factors, such as the number of events, the number of words that comprise an event, and the size of the time frames. For example, the number of terms each event consists of is three. Through a process of experimentation and a number of prolonged preliminary testing, the values for these parameters have been determined. For instance, we also examined both the accuracy and the recall measure by employing an event structure that consisted of a total of only three different phrases. During the course of this preliminary examination, we made the discovery that there is a strong likelihood that the three words are highly similar to one another, which has the potential to result in a large number of false positives. Because of this finding and the fact that other methods, such as Cordeiro [40], employ five terms, we decided to adopt this event structure instead of another one.

The configuration of the amount of events that are reported for each time window is still another obstacle. Finding a configuration that will consistently provide results that are comparable to one another is not an easy task and takes a significant amount of effort because the outcomes of most methods depend on several aspects. The common denominator of 1800 events per data set, also known as 15 events per hour, was experimentally obtained for our experimental setup by iteratively modifying the parameters of all procedures over the course of multiple rounds of testing. The size of the window is one of the most essential characteristics that has to be modified in order to be in line with this parameter setting. There is a broad range of variation in the window sizes that were utilized in the analyses that were detailed in the original papers: sample Sampled Clusters [1] reports around a week, EDCoW [3] reports approximately one month, and in [16] reports approximately one or two hours. Because the promise of identifying events in (near) real time is what drives the development of these systems, we began our experimentation by working with extremely small windows and then gradually grew them larger. Through this method, we experimentally identify one-hour periods that pertain particularly to the work [1].

The provision of measures that are stable with respect to the ranking of the various methods is our primary objective; nevertheless, we are willing to allow changes in the absolute scores that these methods receive over the course of time. It is hard to legally guarantee this attribute due to the fact that our accuracy measure relies on external services that are unable of being crawled and preserved. On the other hand, if we want to do an empirical investigation into the consistency of our precision measure and base it on this comparison, we may draw the conclusion that the results received from a search on Google and the New York Times have not altered all that much. In addition, these minor adjustments do not have an impact on the ranking that was determined using our metrics. This conclusion is positive; nonetheless, it is merely a present observation, and as such, it does not permit predictions regarding how the indexes of the search engines that were employed would develop in the future. In contrast to our accuracy measure, our recall measure allows us to discover the ground truth, and as a result, this measure is inherently stable.

2.5 Summary

In this chapter, we have addressed the lack of quantitative and comparative evaluation of event detection techniques by recommending several measures, both for run-time and task-based performance to precisely detect events. These measures can be used to evaluate the effectiveness of event detection techniques. In addition, previous research that outlined the methods for event detection concentrated mostly on offering comparative assessments of huge sets of data streams and addressing the issue of dealing with massive volumes of data. On the other hand, we concentrated our efforts on the analysis of the procedures that had been carried out in previous research and presented comparison metrics that had been used in the process of assessing the findings of huge data sets. In order to show the validity of the numerous suggested measures for the state-of-the-art event detection algorithms for Twitter data streams, these measures were devised to meet the prerequisites of an already established standard.

We want to make use of our system-based approach in the near future work that we do in order to broaden the scope of our assessments and investigate other methods. At the same time, the procedures that are now being used should have improvements made to them so that data may be processed continually. Additionally, the impact that the pre-processing has on the performance of the task-based and run-time aspects of the program might be investigated. Within the context of our system-based strategy, it is simple for us to get rid of pre-existing operators (for instance, retweet filtering) and insert brand-new operators in their stead (e.g., part-of-speech tagging or named-entity recognition). In conclusion, the development of adaptive event detection strategies may result from doing a more in-depth analysis of the ways in which the various parameters of a method influence the performance trade-off that occurs between run time and task-based performance. Additionally, it would be interesting to add a crowd-based metric to analyze how people would judge the outcomes of the various methodologies in terms of accuracy and in comparison, to the automated measurements.

Chapter 3: Case Study on Anomaly Detection Analysis[1]

3.1 Background

In recent years, social sensing has gotten a lot of attention [3]. However, much effort has been devoted to the problem of event detection [1]. We observe that the event anomalies are not studied extensively in the mentioned academia. Work strategies, relation, and mapping of social information have received little attention in any event. The use of Twitter data in predicting crime [4] is an interesting perspective of event summarization that maps crime information with the day's temperature. Social sensing has received much attention in recent years [3]. This is due to the large proliferation of devices with sensing and communication capabilities in possession of average individuals and the availability of ubiquitous and real-time data sharing opportunities via mobile phones with network connections and via social networking sites (i.e., Twitter). The machine learning approach was included as a context-aware system in [5], with two primary challenges: lowering energy use and minimizing environmental disturbance. This article describes a real-time continuous sensing system that can be implemented without affecting the reporting rate. Information spreads over the connected network in social media. This structure of recognizing information and attitudes through location and user data profiles is related to opportunistic sensing, including social and physical sensing, such as mood and location. The correlation between social and physical sensor data shown in [1]. Social sensing is hypothetically a human sensing of behaviors. One issue that arises as a result is that the obtained data is of lower quality, as people

[1] This chapter is also published as Harshit et al. "Social Network Anomaly Detection for Optimized Decision Development."IJITN vol.14, no.1 2022. (Permission Awaited)

are not as dependable as well-calibrated sensors. As a result, a large body of literature focuses on extracting valuable information from a large pool of inaccurate data. Prior to the emergence of social sensing, much of that work was done in machine learning and data mining. These methods are known as factfinders, a type of iterative algorithm that infers both the credibility of statements and the trustworthiness of sources. Given the degree of corroboration and inferred source reliability, an iterative algorithm tries to reason on this network to extract the most trustworthy information.

Web data retrieval [5], [6], [7] and query sampling [8], [9], [10] have both addressed the problem of information source selection. These efforts are based on the characteristics of sources as well as the material produced by such sources. On the other hand, ours is a content-agnostic approach that focuses solely on source relationships.

Given the ambiguity surrounding observation originality (vs dependence), we propose that diversifying sources is a beneficial method regardless of whether or not credibility evaluation can account for dependence. Our source selection approach is implemented as an online admission controller built into the Control algorithm execution pipeline as an upfront plug-in as in [7][11]. Whereas the work on diversifying sources would not be needed if one could accurately account for dependence be- tween them in data credibility assessment, we argue that, in general, estimating the degree of dependence between sources is very hard. Our optimization can both speed up data processing (by lowering the amount of data that needs to be processed) and increase credibility estimates, according to the results (by removing dependent and correlated sources) [12-14]. It is challenging to establish whether a second report is just a relay of the first or an independent measurement if one source follows another on Twitter and reports the same observation.

In this chapter, we employ the Control algorithm [2], a generic framework that may plug in many relevant algorithms for a wide range of applications. We employ the maximum-likelihood estimator [2] in the Control algorithm as the fact-finding technique. We show that applying simple strategies for social data source that are noisy, can significantly enhance subevent detection efficiency.

3.2 Conventional Network Estimation and Data Source Based Network Estimation

In this section, we provide the details of preprocessing of data harvesting for noisy social media sources which are heterogeneous in nature. Considering this into the account, we know that social data is usually infiltrated through multiple sources and unwanted content, spam and advertisements. When we collect data without knowing its context, we end up with unstructured data represented as gibberish. Therefore, the preprocessing of social data becomes an important task and should be cared. Any structured data analysis creates a simple framework to determine the important data. As a result, text analysis is used to quantify unstructured text data in relative textual and visual data for event detection and prediction (anomalies identification and feature characterization).

Providing we have unprocessed tweets as; we quantify and remove the retweets and duplicates but do not remove them from the data as we utilize the retweets and duplicates for the detection of influence [1] in a network. However, we created a sample group that mentioned the "@" in the retweeted category. Although removed other parts of "@" which were not relevant to the event under consideration. Standard text processing operations are performed on the remaining tweets, including punctuation and special character removal, and URL removal.

3.2.1 Source Selection and Representation

Identifying the right correlation for the specified categories was one of the unique aspects of word retrieval from social data. The similar access was used in [15] to determine the relationship between a word and an action. We adjusted the user weightage aspects to identify the categorized binding terms by attaching the proper features of the rule. Given the set of preprocessed Twitter data as, we provide the critical word graph of each tweet by applying approaches from [2]. We used the Twitter parsing technique [5] to identify the user's retweets and grouped them with basic terms for categorization, as shown in Figure 3.1 It is a network of Twitter users' words who have unique terms in the tweet. This graph will represent a fully connected network of words. Assuming that weight of each user is set to $1/(w-1)$, where w represents the number of users containing a unique category. The features are scaled as,

$$S_f = \frac{(f_v - l_v)}{(H_v - l_v)} \tag{3.1}$$

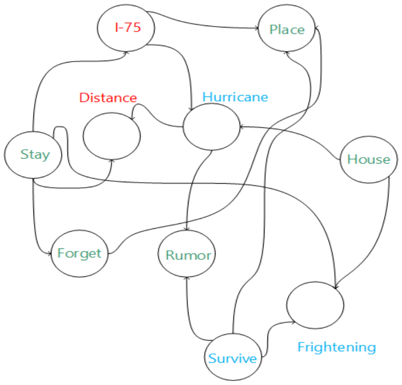


Figure 3.1 The Graph of Words

where, Sf is scaled feature under consideration, fv as feature value, lv as low value, and hv as high value appearing in the set of data [1]. Assuming that every node represents a Twitter user, and the vertices represent the relation of words to other nodes [16]. This transformation of twitter users' tweets into a single relational graph S_t where t corresponds to time interval for all the tweets occurred. Figure 3.1 represents the graph of unique words regarding users' connections illustrating the rewards as unique words are used again for the event. Assuming the words any of these graphs' vertices, edges and nodes that aren't already in S_t are added, and the weights of existing edges are boosted by the weights of the nodes in these graphs. As a result, pairings of terms that appear in a lot of tweets are likely to have a lot of edge and node weight between them.

3.2.2 Sub-Event Detection and Categorization

This section of the paper describes the proposed subevent detection and categorization approach. This approach is based on the ability for nodes to forward the required information for information sharing. The distributed means can meet all feasible needs under the quantified conditions of any event or anomaly. It represents nodes for transmitting important information regarding events or circumstances in the event of any conditions. This approach depends on the identities of (1) how much important information is posted in a windowed time frame (2) pair of words categorized with high abnormality in a widowed time (3) optimization of large network graph to detect anomalies.

Assuming that an important moment occurred during the time t next module of categorization is activated with a descriptive nature of the anomaly, thus creating a convex optimization problem as the data anomaly acts as convex sets of convex events functions. We define the dimensionality of vectors as $n^2 \in R^{nxn}$ thus the tweet matrix for N specific period is given by $adj(S) \in R^{n^2}$. This shows that each N windowed period, we add and extract information weights for pared nodes and vertices and assuming which are not connected as bias $p\ matrix$).

$$adj(S) \begin{bmatrix} | & \cdots & | \\ S_{t-N} & \ddots & S_{t-1} \\ | & \cdots & | \end{bmatrix}, where\ adj(S) now\ R^{n^2xN} \tag{3.2}$$

Therefore, the proposed anomaly detection and categorization problem is,

$$\min_{st} \frac{1}{t_p} \left|\left|adj(S)st - p\right|\right|_{t_p} \tag{3.3}$$

where, st^* be the solution of the above problem which detects the similarity of the content in a category of t periods and p way, we may argue that our approach uses the fact that when something significant occurs within an event, the vocabulary of tweets becomes more particular, and thus the weight of the edges between the related terms increases. Thus, defining the objective function as,

$$c = S^*_{t-N} \tag{3.4}$$

where, c is the event occurred probability. We are looking for combinations of terms that appear in a large number of posts in the current time period but only in a small number of posts in earlier periods. We assume that such a pair of phrases suggests an important event's progression. These terms force the objective function of the optimization problem by $\min(c_{tp}, p)$. The greater the value, the more likely it was that a significant event occurred during the current time period.

It assumed that tweets containing multiple "important" nodes and edges with the utmost of the details of an anomaly in a subevent. Let A represent the tweets with anomaly information in S and give a category to the anomaly. Thus, as we add the users and tweets to the category, the function which takes the input will not decrease with the increase in size and time. Therefore, the goal shifts to determine the policy to evaluate the inputted data streams that define the function as monotonic. This monotonic nature depends on the volume of tweets and assuming tweets vary a lot, a parameter for learning needs to be defined to approach the precise detection of anomalies in subevents.

3.2.3 Strategy and Learning Parameter

Assume that the optimal value to detect an anomaly is $\emptyset$ for each subevent. Thus the accumulation $\emptyset$ precisely detects a subevent for any event. Therefore, we need to find the threshold value of $\emptyset$. Randomly we select X events and Y events in test sets; we utilize exhaustive grid search for all the optimal threshold values with a set step size of 0.2 in between 1-10 as a set of p biases. Thus, formulating a model as,

$$\emptyset = \alpha X_t^2 * \beta X_t + \delta \tag{3.5}$$

where,X_t is the tweets for the X events and $\alpha, \beta,$ and δ are the parameters.

3.3 Proposed Binary Graph-Based Network Search

Let us assume that every node represents a Twitter user, and the vertices represent the relation of words to other nodes illustrating the reward accumulation as shown in Figure 3.2. Higher the reward, better the possibility of these words to be used again for the same occurred events These rewards are proportional to word search using a graphical model by assigning partial variable that is related to action, emotion, and location and tweet-size constrained to limit the search space. This creates a linear space to utilize depth.

In Figure 3.2, two nodes X and Y represent two users with a size of network l based on tweet size. Assume user X is having one or more category of either action, emotion or location given by H_i and user Y is having all three categories given by (H_i, v) where v has corresponding values as H_i. Given (X,Y) users, we will have a span of rewards as $\gamma = \gamma_l(X, Y)$given by equation (3.2). Therefore, as the vertices of user rewards starts increasing, the highest reward is accumulated when minimum of X, Y is picked. Hence the reward accumulation is $X, Y \leftarrow (H_i, v) * l^{\gamma}$.

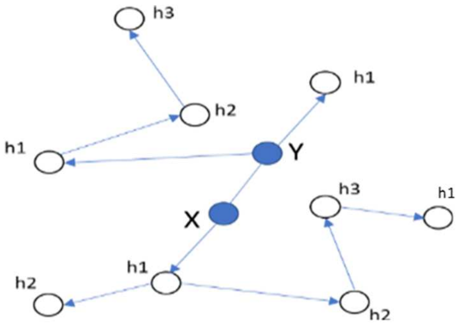


Figure 3.2 Graphical User Representation and Interaction

$$reward(\gamma) = \frac{w}{min(|X|,|Y|)}, \quad w = \frac{|XY|}{1+|XY|} \qquad (3.6)$$

Loss factor calculation given in equation 3.7 calculates the loss of rewards between any two nodes after graph search,

$$L(S,H) = \frac{w}{1+|XY|} * \frac{\gamma}{min(|X|,|Y|)} \qquad (3.7)$$

Further, we define A_γ for the accumulated reward at any state s_t where seed time *t* is zero.

$$Accumlated\ reward(A_\gamma|s) = \sum_{i=0}^{p}[(\lambda_1, \lambda_2) * L * \gamma^{k+1+t}\ |s \qquad (3.8)$$

Table 3.1 Influence and Decision Score

Data Set	Measure Metrics					
	Anomaly			Influence		
Social	Precision	Recall	Rscore	Precision	Recall	Rscore
Optimized Results	0.68	0.62	0.65	0.87	0.93	0.83

3.4 Summary

We evaluated the proposed system for the dataset gathered for a hurricane [1] Irma from September 9 to 12, 2017 when it moved across Florida. 295,621 tweets were gathered and examined over time, yielding a 286,675,237-word cloud with an average of 97 words per tweet. Raw physical sensor data, on the other hand, were collected and analyzed for 72 hours in five locations in Florida (Florida Keys, Miami, Naples, Tampa Bay, and Jacksonville) for five categories including wind, speed, pressure, rain, cyclone (56 advisories), and tornadoes.

All tweets posted within a specific time window are first preprocessed using the same approach as in the proposed system, given a stream of tweets. Following that, the current time frame's tweeting category rate is calculated. Then, if it exceeds a certain learning parameter threshold, the algorithm believes a sub-event to have occurred using a model identical to the one provided above. The threshold value was calculated separately for each data set. Finally, the model's parameters were optimized using the same set of event sets as the proposed approach.

We first test the suggested system on the objective of subevent detection using anoptimized strategy, where the influence score represents the anomaly score. We used the collection of mapped tweets with the help of the control method [1] to get these results. Standard information each are shown in Table 3.1 retrieval measurements such as precision, recall, and Rscore are used to report performance. The outcomes of the optimization approach for a set of 21 sampled datasets over a duration of 2 hours. Although our framework successfully determined sublevel events and anomalies with good accuracy, there are a few aspects that need to be investigated further. First, the framework necessitates the collection of multiple daily data points and is prone to high inaccuracy if the physical sensor's threshold values change dramatically. This is because we are modifying a few parameters to avoid biased learning.

This chapter introduced a novel method for producing real-time categorization of events using solely Twitter tweets of all users. We analyze the Twitter database for hurricanes by tracking social participation during the event period. It aided us in comprehending the qualities of social responsibilities in the aftermath of a hurricane. Our approach to hurricane events shows that our system outperforms the leading sub-event detection problem while also producing beneficial anomaly detection with an average precision of 0.65. However, a significant source of mistake in our method is that we only distinguish a single weather influence feature rather than a combination of weather impact characteristics.

Chapter 4: Case Study on Information Dissemination in Extreme Weather Scenario[2]

4.1 Background

In social media, the information propagates across the connected network. This structure of identifying information and sentiments through location and user data profiles relates to opportunistic sensing comprising social sensing, i.e., emotion and physical sensing, i.e., location. The correlation between social and physical sensor data shown in [8] effectively utilized contextual information to integrate the abstract nature of keywords. However, the expected correlation with the local conditions was not found while analyzing the historical social data [5]. Therefore, the concept of extraction amplification applied through virtual world analysis is to inherit the real experience of the physical world and its predictability. To perpetuate this, the reinforcement method to support the networked information can determine the decision process. Hence, the objective of this study is to investigate the usability of social networks during weather disasters, analyze the characteristics of a perpetual network, improve the decision and communication strategies, and facilitate the development of disaster tools.

4.2 Literature Review

During the past five years, researchers have extensively started analyzing social network data [3-4]. In any case, work techniques for relationship and mapping of social information with physical sensor information have not studied broadly. One of the thought-provoking studies is on the prediction of crime through twitter data[5]. Paper [5] talks about crime incidents and their

[2] H. Srivastava and R. Sankar, “Information dissemination from social network for extreme weather scenario,” IEEE Transactions on Computational Social Systems, 7(2), pp.319-328, 2020. Permission attached in appendix A.

to investigate crime sentiments but lacks spatial analysis and does not give augmented results. A real-time diagnostic method of symbolic aggregate approximation (SAX) [9] shown through global sensor network and predictors were used to reduce energy. One of the essential claims is that predictors behave in an unfamiliar way in the distributed system. Hence, data aggregation is needed first. It gives another dimension of analysis and reduces the need to acquire specific information. Even though this enhances data acquisition, it neither optimizes the data fusion nor utilizes the opportunistic data.

Smog disaster prediction was carried out with the help of social data and physical sensor data in [10]. This paper tries to bridge the gap between traditional web forecasting with social web data for air pollution (as smog), but the data used to predict the air pollution extracts onlythe opinion of humans. The correlation between social data and the physical sensor has ascalability and interoperability issue.

In [2], human behavior was focused on context-awareness or inference as a recommender system. This development was based on collaborative filtering or through the ranking system. However, the problem was in the sparsity of information or explicit ranking through feedback. In [1], energy demand gathered through smart physical sensors and social group feedback was used to project the demand reduction of energy. The central concept used in this paper was data acquisition through people's behavior.

In [11], the machine learning concept was being incorporated as a context-aware system with two key challenges, i.e., minimizing energy usage and minimizing the interference of environmental effects. This paper gives a real-time continuous sensing mechanism incorporated without alteration in the rate of reporting. Additionally, [12] also uses different data sources and sensors by explicitly defining the social and physical data types and came up with a method to fuse two different data for processing instead of processing them individually for data scalability. Mainly addressed the blending of social and physical data by proposing a framework that enables to analyze tweets and extracting people's mood depending on days' weather to develop a recommendation system.

In [13], real-time data monitoring and determination of critical events done by interfusion of different data models gathered from social media for smart cities to make smart decisions. In [14], an explanation of the four-folded technique for automated real-time Twitter data collection and classification by defining the correlation factor with a web interface for displaying the events with the help of environment data. However, this system works as an event monitoring framework without providing feedback or prediction warning for an emergency or disastersituation.

Importance of social network was demonstrated in [15-17] by showing the roles and warning detection through the activities in social network with a primary focus in opportunistic sensing and has been used in different applications such as transportation model [18], assistive medical services, and smart urban area mapping [19]. In [8, 20, 21], diffusion of Twitter data was incorporated to find dissemination rates. It results in faster diffusion from known nodes to external nodes.

An online social network defined as a set of social entities (people, groups, organizations) and the patterns of relationships or interactions between [22]. Social network analysis (SNA) was designed to discover the relationships established between social entities. It includes (i) computation of metrics that provide a local data (actuator) and globally (network level) description of the network, (ii) graphical visualization of the network, and (iii) community detection for understanding the structure of complex networks and finding useful information from it. Many software and tools have also been developed to fulfill the increasing need for social network data mining and visualization techniques such as R and the SNA library, JUNG, Guess, Prefuse, NodeXL, Gephi, and FluxFlow.

The objective of this research is to build a bridge between two different data sources by appropriately merging information disseminated from quantified social data with real-time physical sensor data. This analysis represents the predictive investigation towards the use of Twitter's social data and network utilization with the help of NodeXL [23] and NOAA's U.S. National Weather Service [24]. The general concept utilized in developing the methodology to generate a decision/response score is shown in Figure 4.1. The approach is to combine two stages, i.e., two separate data analyses, in parallel. The first stage is social data analysis, which utilizes reinforcement learning method to quantify essential characteristics. The second stage is physical sensor data analysis, which utilizes historical data and real-time data. However, in the final stage, multi-strategy learning is utilized to maximize the learning environment and states for a final computed decision/response score. The critical aspects of this research are mentioned below.

1. Quantify the social data with physical data gathered from Twitter [23] and NOAA [24] for hurricane disaster events like Harvey, IRMA, and Michael that occurred between 2017-18. The social data consist of approximately 2,100,000 tweets on a geographical basis.
2. For the analysis:
    a. Perform preprocessing and feature selection
    b. Create rules set using the defined attributes and apply classification
    c. Develop tiers for each type of data queries to reduce the false positivity for classification, thereby improving the efficiency.
    d. Find the relative patterns of social data with physical data before, during, and

after the events.

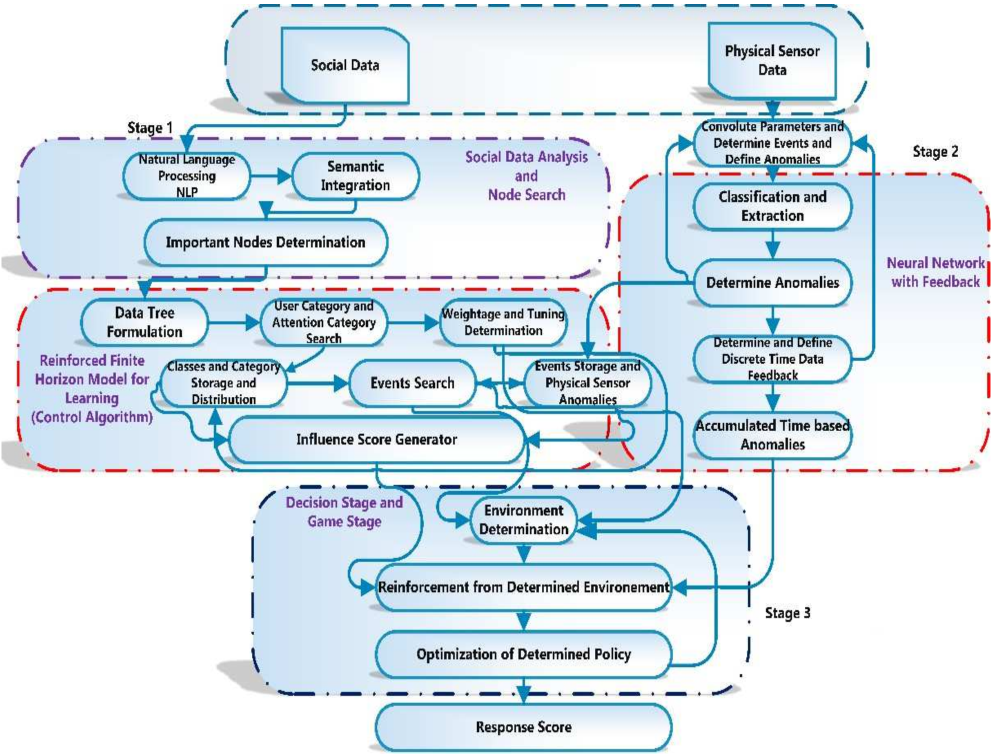
e. Analyze the information dissemination in social data within the vital context of data gathered, i.e., solely based on social analysis.

f. Predict the anomalies from social data in the context of physical data.

g. Prepare a proper state-of-the-art method to gather and analyze the quantified data. Find the utilization of event detection.

4.3 Proposed Framework

The general concept utilized in developing the methodology to generate a decision/response score is shown in Figure 4.1. The approach is to combine two stages, i.e., two separate data analyses, in parallel. The first stage is social data analysis, which utilizes However, in the final stage, multi-strategy learning is utilized to maximize the learning environment and states for a final computed decision/response score.



Social Data
Physical Sensor Data
Stage 1
Natural Language Processing NLP
Semantic Integration
Social Data Analysis and Node Search
Important Nodes Determination
Data Tree Formulation
User Category and Attention Category Search
Weightage and Tuning Determination
Reinforced Finite Horizon Model for Learning (Control Algorithm)
Classes and Category Storage and Distribution
Events Search
Events Storage and Physical Sensor Anomalies
Influence Score Generator
Convolute Parameters and Determine Events and Define Anomalies
Stage 2
Classification and Extraction
Determine Anomalies
Neural Network with Feedback
Determine and Define Discrete Time Data Feedback
Accumulated Time based Anomalies
Decision Stage and Game Stage
Environment Determination
Reinforcement from Determined Environement
Stage 3
Optimization of Determined Policy
Response Score

Figure 4.1 General Approach and Methodology

4.4 Data Acquisition

On structured information, any analysis creates a simple outline to determine relevant information. However, when we collect data without knowing its context, it develops unstructured information in terms of gibberish representation. This unstructured collection of data, when viewed in the form of words and phrases, is known as text mining. Thus, text analysisis utilized to quantify the unstructured text information in the form of relative textual, visual information for event detection and prediction (anomalies detection, feature characterization). One of the most contextual information utilizations can be observed in the field of marketing andjournalism to identify the favorite words and phrases used by a group of people and communities.

Therefore, during extreme weather or disaster situations, a decision-making tool is needed to find critical information regarding a real-time situation. This information can materialize through diffusing the preprocessed classified data from physical sensor and social media. Therefore, to better analyze the prospect on the factual occurrence and to define theproper anomaly prediction for the embryonic event, we quantify this classified information into the group of two data types as physical sensor data gathered from NOAA and social data. We re- packaged the social data using electron/Hydrator and created visual information of words as a cloud of words. Further, we used initial logistic analysis to form a group of sensor data readings for every hour with respect to the re-packaged social data. Hence, this step increases our abilityto analyze the filtered data from the cluster and help us to identify critical words and semantics.

4.5 Preprocessing and Event Detection

An event is defined as any incident that occurred. Even though physical data classified in any weather situations or anomalies from the threshold values, the main question is whether the classification or prediction of events is possible from social network word groups. To identify the social network words that represent anomalies, we grouped the data on every 2-hour basis into 12 groups for a 24-hour period as shown in Figure 4.2. Through analysis, we found that most of the critical words used by the user show their behavior correlating to the event detected from the physical sensor data. This indicates that the usage of words and phrases directly confirming the change in weather. For event detection, the physical sensor data underperformed compared to the social data when hurricane conditions including wind speed and the amount of rainfall, starts to decline at a location, while barometric pressure starts to increase. i.e., when the severity of the weather condition declines in that location and increases in the next location along its path. While considering the entire 24-hour data, it overperformed under the same conditions.

However, this analysis helped us to observe the critical words category behavior and events concerning physical sensor data. Upon analysis, we grouped the words into three categories as actions, emotions, and locations. We gave the importance to actions, followed by emotions, and then locations in terms of word-search. The reason for using this importance level is to define emotion based on actions. In the twitter data, when emotion is mentioned, the action is not and when an action is mentioned, the emotion is not. For example, the tweet: "it's going to hit my place, need to hit I-75" is grouped as action, emotion, and location. The affirmative word "need" in the tweet represents the action and emotion. The phrase "hit my place" represents a location and further with the phrase "hit I-75" a decision can be materialized. Similarly, the following tweets can be categorized. "It's just a rumor I will survive in my house" [action, emotion and location] and "It's frightening that hurricane is here" [emotion]. We noticed that these categorized groups change characteristics as the event progresses. Action and emotion categories are utilized more before the event occurred. During the event, location and emotion categories are used, and after the event, only the location category is used.

4.6 Feature Extraction

The most anomalies of word extraction from social data were to find the correct association for the given categories. In [25], the association technique was used to identify the relation of a word to the action. Thus, associating the proper aspects of the rule, we changed the user weightage aspects to identify categorized critical terms. We identified the user's retweets from the Twitter parsing technique [5] and grouped them with base words for categorization, as illustrated in Figure 4.2. It represents a scale-free network of twitter users active during the time of hurricane. The scaled features that are searched and mapped are shown in colors, with different color representing each category. The red color represents tweets of users utilizing all three categories: action, emotion, and location in one or more time. We considered this category as critical. The blue color denotes users using action words and green color denotes emotion. Through this scale-free network, we observed that most of the users were writing tweets by utilizing the same phrases already used in the network. Further, we utilized the twitter tag method to collect relevant words concerning context-based phrases [26].

$$scaled_{feature} = \frac{feature_{val} - low_{val}}{high_{val} - low_{val}} \tag{4.1}$$

where $Scaled_{Feature}$ is the scaled value of the feature under consideration, $feature_{val}$ is the original feature value, low_{val}, and $high_{val}$ is the lowest and the highest value of featuresappearing in the data set, respectively [27].

4.6.1 Binary Search Features

To explain this, let us assume that every node represents a Twitter user, and the vertices represent the relation of words to other nodes illustrating the reward accumulation as shown in Figure 3.2 in Chapter 3. Higher the reward, better the possibility of these words to be used again for the same occurred events These rewards are proportional to word search using a graphical model by assigning partial variable that is related to action, emotion, and location and tweet-size constrained to limit the search space. This creates a linear space to utilize depth. The two nodes X and Y represent two users with a size of network l based on tweet size. Assume user X is having one or more category of either action, emotion or location given by H_i and user Y is

having all three categories given by (H_i, v) where v has corresponding values as H_i. Given (X, Y) users, we will have a span of rewards as $\gamma = \gamma_l(X, Y)$given by equation 4.2. Therefore, as the vertices of user rewards starts increasing, the highest reward is accumulated when minimum of X, Y is picked. Hence the reward accumulation is $X, Y \leftarrow (H_i, v) * l^\gamma$.

$$reward(\gamma) = \frac{w}{min(|X|,|Y|)} \qquad w = \frac{|XY|}{1+|XY|} \tag{4.2}$$

Loss factor calculation given in equation 4.3 calculates the loss of rewards between any two nodes after graph search,

$$L(S, H) = \frac{w}{1+|XY|} * \frac{\gamma}{min(|X|,|Y|)} \tag{4.3}$$

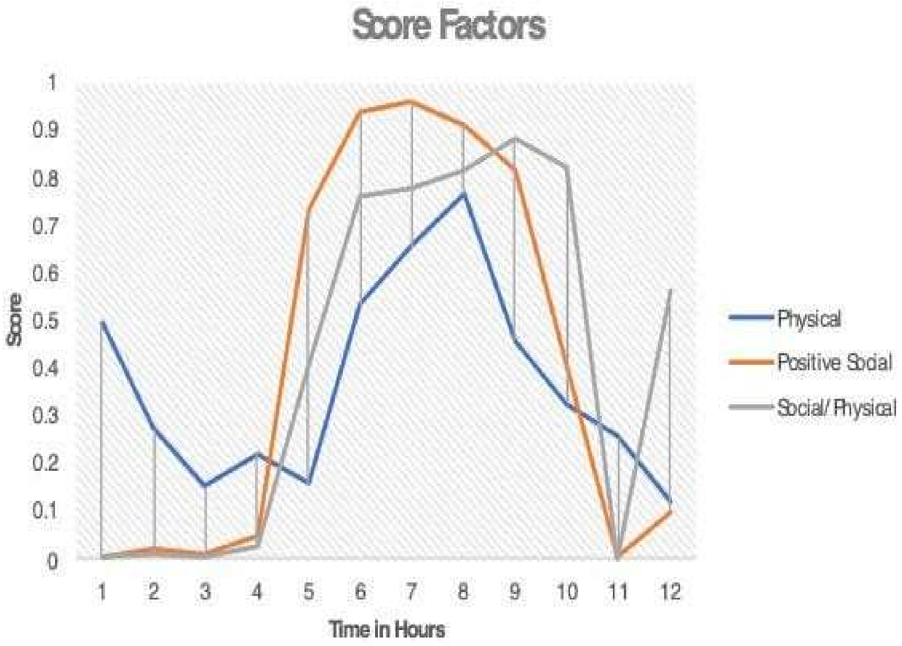


Figure 4.2 Data Analysis for 2-hour basis during IRMA Hurricane on09/10/2017

Further, we define A_γ for the accumulated reward at any state s_t where seed time *t* is zero.

$$Accumlated\ reward(A_\gamma|s) = \sum_{i=0}^{p}[(\lambda_1, \lambda_2) * L * \gamma^{k+1+t}\ |s \tag{4.4}$$

4.6.2 Proposed Control Mathematical Model and Optimization Features

The diffusion of information requires the possible means for nodes to forward the required information. The diffused means can enable all the possible needs under the quantified criteria of any event or anomalies. It represents the nodes to forward the critical information about the events or situations in possible conditions. The situations filtered from the gathered physical data location behaving abnormally from the threshold values at that time interval corresponding to social data. It leads to categorizing the words from the social data in ascending order. The first level stated as physical data as per twitter user-based importance. This data directly relates to user updates copied or correlated with weather agencies' reports. This is the raw critical information categorized from real-time data from physical sensors. The second level states the event information from social data from the weighted user. The weighted user information is extracted based on the number retweets on critical event or anomaly. The threshold value ρ has utmost importance to gather information, but an essential factor is characterized by acquiring n number of tweets of l lengths.

To understand this, we use raw data tweets $x_i = 1$, constituting pertinent information about hurricanes and storms. The symbols used are explained in Table 4.1. Equation 4.5 represents the estimated rewards accumulated for a given state to determine the policy. Equation 4.6 represents the combined evaluation of tweets in comparison to physical sensor data. This equation represents the core working of the *control algorithm*. This algorithm is modified accordingly for different analysis methods for evaluation and generation of analysis scores for comparison,

$$G^{g\in\rho}(s_{p,t}) = E_g(A_{\gamma,t}|s_t, t = 0) = CA^{g\in\rho}(s_p, a) \tag{4.5}$$

Table 4.1 Used Notations in Equations

Symbols	**Definition**
$\boldsymbol{\rho, s_t, l_i}$	Threshold values of physical sensor data, State of the event at time t, Indicated variable for time-based sensor data.
$\boldsymbol{N}$	Number of tweets considered for summarization (in the time window specified by user)
$\boldsymbol{T, m, p, s_{p,t}}$	Total time, Number of distinct content words and subevents included in the n tweets respectively, State of subevent at time t
$\boldsymbol{msize}$	Number of tweets containing distinct words
$\boldsymbol{i, j, k, a}$	Index for tweets, Content of words, Subevents, Classes, respectively
$\boldsymbol{x_i}$	Indicator variable for tweet i
$\boldsymbol{y_j}$	Indicator variable for content word j
$\boldsymbol{z_k}$	Indicator variable for subevent k
$\boldsymbol{CT_{Score(j)}}$	Feature score of content word
$\boldsymbol{SS(k)}$	The score of subevent k
$\boldsymbol{Im}$	Importance/informative score of class a
$\boldsymbol{Cat(i)}$, $\boldsymbol{lat}$	Class of tweet i, Lateral averaged required data
$\boldsymbol{\tau, g}$	Tuning parameter for sensor data, policy determination
$\boldsymbol{\lambda_1, \lambda_2}$	Tuning parameter – relative weight for the tweet, content word, and subevent score
$\boldsymbol{C_E, C_t}$	Set of categorized words and subevents present in tweets, respectively

$$CA^{g\in\rho}\left(s_{p,t},a\right)=\max_{g}.\left[\sum_{i=1}^{\rho} l_i * \tau_i * Im\left(lat(i)\right)+(1-\lambda_1-\lambda_2).\sum_{t=0,}^{T}\sum_{i=1}^{N}[x_i*(\gamma,t+1)* Im\left(Cat(i)\right)]+A_{\gamma,t}^{t+1}*\{\lambda_1*C_t*\sum_{j=1}^{m} CT_{Score(j)}*y_j.*\left(Im\left(Cat(i)\right)\right)+\lambda_2*C_E\sum_{k=1}^{p} SS\,(k)*z_k* max_{i\in T,SS_k}\left(Im\left(Cat(i)\right)\right)\}\right] \tag{4.6}$$

To understand the equation 4.6, we need to first investigate the constraints.

1. Here, the tweets lengths must be of the desired value, i.e., the length constraint by twitter platform and tweets per user defined.
2. Feature Constraint in Class $x_i m\left(Cat(i)\right)+\lambda_1*\gamma*CT_{Score(j)}.+\lambda_2*SS(k).z_k\geq 0$ $(1-\lambda_2)x_i+CT_{Score(j)}\geq 0$ and Content Constraint
3. Subevent selection Constraint

$$CT_{Score(j)}.+SS(k).z_k\geq 0$$

Thus, the analysis scores generated by the *control algorithm* for every event and the selected category. The normalization is utilized to create the loss factor in the present context to represent

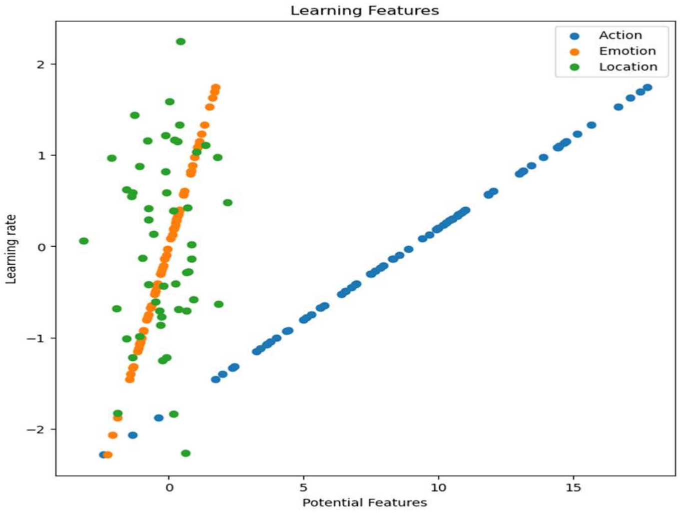


Figure 4.3 Categorization of Events

the relevant utilization factor. Whereas, the *classification algorithm* inspired by [28] helps to quantify word categories.

Table 4.2 Control Algorithm Structure

Control Algorithm

```
Input: Initialize $\boldsymbol{graph\ tree\ G_s'} = (G_1 \ldots\ldots . G_N)$ And Target $H_i \leftarrow$
$\{h_1, h_2, h_3\}\ categories$;
for $N = 1\ to\ N\ \ do$
    Choose A Graph $V$ From $G_s'$ And Select Accumulated Nodes $G$        From $G_s'$;
    Get Featured Matrix $G_o = G(V \cup\ N_V(V))$;
    Compute Featured Adjacency $G_o{}'$;
    for $t = 0\ to\ t = T - 1$ , T=Total Time Span
        Node $V_t$ From $H_i$ With $\gamma_{Highest}$ Reward Else
        Query Node $V_t$ And Observe New Graph;
        Set $lm(\rho, x_i) \leftarrow H_i\ (T - 1)\ and$ Determine H;
        Sum $(x_i) \geq \rho, y_j$;
        Set  $A_t \leftarrow arg \max_{\gamma_t \epsilon x_i} A_\gamma(s, H_i)$;
        while $y_j \leftarrow x_i$
            $\gamma_j \geq\ \lambda_1, \lambda_2 \leq \gamma_i$ Category Determination Tuning    Factor;
             Query $lm(\rho, x_i)\ at\ t = T - 1, compute\ L_{t,t-1}(S, H), \lambda_1, \lambda_2$
        for $G_o$ As $g$    Do    Compute Deep Search Node Embeddings
            Determine Strategy Query $(A_\gamma, H_i), (L_\gamma, H_i)$;
            Compute Influence Score Through Analysis Steps;
            Compute New Node Embeddings With $lm(\rho, x_i)$;
            Set $lm(\rho, x_i) \leftarrow A_{\gamma_{x_i}}\ (T - 1)$;
            With Probability $\theta\ and\ \varphi$, Select Rewards In Ascending;
            Select Node $V_{t\ with\ \theta\ ,\ \ \varphi}$    $g \leftarrow\ arg \min_{\theta\ ,\varphi \epsilon x_i} lm(\rho, x_i)$;
            Buffer $g$ With Tuned Parameter $\tau$
        end
      Update $G^{g \in \rho}(s_{p,t})$  Target Network $G_{\theta\ ,\varphi}$
        For All $g \leq 0.09\ find\ O(1/\sqrt{g_i})$ Convex Hull Optimization
        end
    end
end
```

Table 4.3 Classification Algorithm Structure

Classification Algorithm

```
Input: $\boldsymbol{L_N}\{\boldsymbol{item\ Sets}\}$
No. Of Users $\boldsymbol{L_{N-1}}$
for N=2; $\boldsymbol{L_{N-1}} \neq \{\}$  Do
    Query New Users $\boldsymbol{L_k = apriori - gen(L_{N-1})}$
    For $\boldsymbol{x_i}$ , $\boldsymbol{y_j}$   $\boldsymbol{do}$
        Find Rewards Order, $\boldsymbol{L_t = subset(L_N, t)}$
        $\boldsymbol{T_i \leftarrow (T_i, v) * l^Y}$
    end
    $\boldsymbol{L_N = \{c_E, c_t \mid (T_i, v) * l^Y > minsup\}}$
end
Activate User Nodes and Generate Rewards
```

4.7 Convolution Neural Network (CNN)

For detecting weather anomalies from the physical sensor dataset, we applied multi-layer convolution [29]. This network is fully-connected through small layers and assembled with the extension of a multi-layer perceptron model. It is to increase the rate of learning. To understand this, let us assume in our obtained dataset that we have network . Then assume we have any two distinct conditions as and . These conditions are used to train a set of physical sensors as . Hence

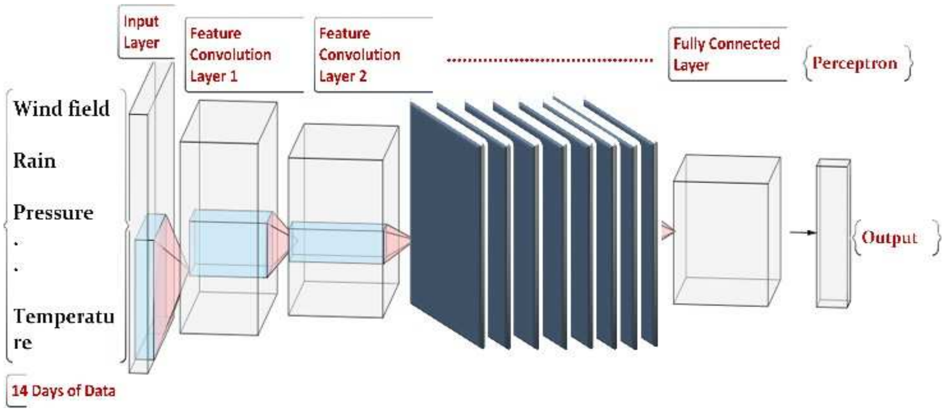


Figure 4.4 CNN Setup

for n different conditions Therefore, for physical sensor in the network, the training set is defined by equation 4.7, $S = (e_p, r_p)$. Hence for n different conditions,

$$S_n = (e_1, r_1, e_2, r_2, \dots . e_p, r_p)$$

Therefore, for i^{th} physical sensor in the network, the training set is defined by equation 4.7,

$$f_i(x) = \min_{w_i, w_j} \sum_{i=1}^{p} \sum_{j=1}^{p} w_i w_j F_{ij} \tag{4.7}$$

This is utilized to minimize the function to find appropriate weights w_i *and* w_j whereas F_{ij} is the estimated correlation coefficient for network $f_i(x)$ so our weights are defined by equation 4.8,

$$w_{ij} = \frac{cf_{ij}(x)}{\sum_{i=1,j=1}^{p} cf_{ij}(x)} \tag{4.8}$$

where c defined as a certainty of a neural network at state s,

$$c = \begin{cases} s & if \;\; s \geq 0.45 \\ 1 - s & otherwise \end{cases} \tag{4.9}$$

Therefore, as y approaches zero, we become more confident that the event is not present. Thus, we want the accumulated weight $A_{w_{ij}}$ to ensemble the network output. Assuming each network contributes to the sum of proportional certainty. For example, a value close to 0.45 would contribute less than 10 %, whereas value 0.99 will affect more certainly.

4.7.1 Application of Deep CNN

A deep neural network viewed as a supervised model with multiple hidden layers. It consists of neurons in a hidden and visible layer without connecting in the same layer. Thus, by assuming C_j the continuous state with B_j as bias with w_{ij} weight. By conditioning the state network as binary, we have activation function as,

$$C_j == \frac{1}{1-e^{-C_j}} \tag{4.10}$$

$$E = -\sum_{z=1}^{p} \sum_{i,j}^{\substack{p_x-n_x-1, \\ p_y-n_y-1}} \sum_{i,j}^{n_x,n_y} \delta_{i+r-1,j+s-1} C_j^z \, w_{rs} \tag{4.11}$$

To develop the CNN for sequential data type for physical sensor data training, thealgorithm trains by perceptron backpropagation to maximize the likelihood for training instances. Consider the created dataset network parameter is α, Then the probability of assigning a label in dataset network is,

$$p(s_p | x_i, \alpha) = \frac{\exp^{h_y}}{\sum \exp^{h_y}} \tag{4.12}$$

Then final recursion is,

$$R_t(k) = \log add \;\; s(x,p,\alpha) = h(t,k) + \log add \, (j) R_{(t-1)}(j) + A_{jk} \tag{4.13}$$

4.8 Convolution Neural Network (CNN) Algorithm

Multi-layer perceptron (MLP) is a common technique used in deep learning, but for nonlinear data with large variables, it produces inaccurate global minima’s and requires high computation time. Due to this, the hidden layers' learning process will be slower. Whereas, when we use CNN, we can convolute inputs to model variance in specific hurricane weather situations. In CNN , input topology can be changed, and the local correlation of information produces improved results. We investigated the use of CNN to achieve high classification accuracy with joint feature learning and further we have a large physical sensors data from hurricane IRMA. This is done by using greedy layer-wise learning algorithm [30] to extract prominent features.

The features and aspects of hurricane statistics depend on its surrounding atmospheric anomalies. Thus, we used a minimum of 2 anomalies around each event. We formed a window considering local features acting as a middle anomaly. This is fed as an active vector to CNN. The network contained one input layer, two pooling layers with two convolution layers, and fully connected layer of 256 hidden units. The general construction of the CNN layer is shown in Figure 4.4 and configuration in Table 4.4. The activation function for all layers is '$relu$', which provides efficient convergence of the model. The process was performed on each weather anomaly. The training completes after convolving the system using propagation. We did not use the traditional method of likelihood learning scheme since we were performing the process on each anomaly. We used 14 days of hurricane IRMA data with 7 days before it hit Florida and the next 7 days after that to reduce the risk of non-detection for low-level sub-anomalies and vice-versa, which helped to have better threshold values, i.e., variance for specific weather conditions of hurricane.

Table 4.4 CNN Initial Set up Configuration

Layer Type	**Configuration**
Convolution	Filter: $32,\ 5 \times 5 \times 2$, Stride $2 \times 2 \times 1$
Max Pooling	Kernel Size: 3×3, Stride: 3
Convolution	Filter: $64,\ 5 \times 5 \times 1$, Stride $2 \times 2 \times 1$
Max Pooling	Kernel Size: 3×3, Stride: 2
Fully Connected layer	256 Neurons
Perceptron	256 Neurons and Output Layer of 'Sigmoid' → $MLP(CNN)$

4.9 Evaluation, Result and Discussion

In this study for evaluation, we processed Twitter tweets data collected on hurricane IRMA from September 9th to 12th, 2017 when it moved across the State of Florida. The number of tweets collected and considered in this time frame is 295,621 giving 286,675,237 word-cloud with an average of 97 words in each tweet. Whereas raw physical sensor data was collected and evaluated on five categories such as wind, speed, pressure, rain, cyclone (56 advisories), tornadoes for 72 hours for 5 cities in Florida (Florida Keys, Miami, Naples, Tampa Bay, and Jacksonville). These were chosen because the greatest number of Twitter users were from the affected cities. The evaluation and analysis scores resulting from each analysis for the 2 types of datasets (social and physical sensors) with 6 stages/dataset shown in Table 4.5. The six stages are the measured hurricane intensities just before landfall and while moving across those 5 cities at agiven time period. The comparative scores computed by plugging the subevents characteristics through available physical sensor data.

4.9.1 Evaluation Points

As we mentioned, the importance factor can traverse across the categories, or it can pertain to a particular level. The constraints were set in Equation 4.5 and tailored for each of the classes. In this analysis, we acquired 9 different classes in which 8 for social and 1 for physical sensor data. In each class, the average words processed were 8,958,602.

Further, the importance of each Im (class) is set to diffused weight and for the tweet parameter (Cat). It also represents the minimum number of data points and must be included from each class, whereas τ is used as a tuning factor for physical sensor data which varies $\tau \in (\lambda 1, \lambda 2) \leq 1$. This tuning factor is measured and tailored to social data importance for equilibrium.

4.9.2 Determination of Policy

Further, the importance of each λ is set to diffused weight and for the tweet parameter (Cat). It also represents the minimum number of data points and must be included from each class. The goal of the control algorithm is to learn to evaluate all the data streams, according to policy determination g. Therefore, to determine the action to be performed, we acquire event data threshold from the physical sensor for the required environment or event represented by $s(pt)$.

The control algorithm estimates the rewards accumulated and estimates the value of g. The value of g is then compared with physical sensor state event anomaly to determine the goodness of g.

4.9.3 Evaluation Methodology

To assess the methodology for occasion-based inconsistencies, we arranged words as sets of events and action words for each subevent anomaly. To classify the important subevents concerning anomalies, we identified by sampling through t-distribution. Identification done by picking two subevents by no less than two featured words from defined categories (h_1, h_2, h_3).

This increased the normalized sampling of data with every 1000 data points, as shown in Figure 4.3. After increasing the normalized sampling, we concluded that the analysis should be done with specific conditions for comparative results. We categorized the analysis methods for time interval and synchronicity, linearity, independent data feature analysis, and regression.

1. Time analysis represents the analysis of bounded data availability in a specific interval. It analyzes different data types in divided time intervals by collecting data asynchronously with the timeframe.
2. Persistence analysis represents the time analysis with synchronous data within divided timeframes. It's a probabilistic analysis with conservative effects on the failure.
3. Asynchronous analysis does not consider time synchronicity or division of timeframes.
4. Linear analysis considers the linear behaviors of different data types without considering any time and persistent analysis.
5. Independent analysis determines the specific set of rules to determine the possible results. The social data collected should create favorable inferences in regard to an event is occurring or not by gathering event-related inferences with words. Thus, the score of the collected words represents the words available to use per hour basis without the

reference of physical sensor data. It is done independently without determining the event but by defining the subevents like wind, flood, etc. This is to bifurcate the words for a major event. This analysis gives a score for 5 cross-fold and states the narrative of subevents.

6. Regression analysis utilizes target specific subevents after detection. It is done by utilizing the scores from the independent analysis method. The steps utilized a simple methodology and considered:
    a. Identifying the probable subevents positive instances and create a random sample.
    b. Collecting individual characteristics from the random samples.
    c. Determining the categorized words inherited through this assumption.

Each of the above analyses results in an associated output score: Time Score (TS), Persistent Score (PS), Asynchronous Score (AS), Linear Score (LS), Independent Score (IS), and Regression Score (RS) for the prospect of comparative evaluation of social data with the physical sensor data.

Table 4.5 analysis score represents the mapped similarities of physical sensor data to social data for each of the analysis methods. Several insightful observations can be made from the evaluation. As the number of data increases, the score efficiency increases, and the asynchronous score decreases with time, which means that the algorithm corrects itself for the concurrent information for the given interval of time. Therefore, the values used previously are re-sampled to create a score for the current analysis. However, if the correlated sampling has a score with 10% accuracy for the categorized positive information, then the analysis score would be set to zero.

Hence, a decision is made to accept or reject the categorized words. Every evaluation stage score increases the diffusion confidence with every analysis step.

Table 4.5 Evaluation of Data

Datasets	**Analysis Score**					
	TS	**PS**	**AS**	**LS**	**IS**	**RS**
Physical Sensor	0.42	0.75	0.50	0.28	0.42	0.24
	0.67	0.68	0.17	0.23	0.25	0.37
	0.78	0.95	0.08	0.12	0.03	0.42
	0.19	0.78	0	0.13	0.15	0.36
	0.81	0.88	0.08	0.07	0.10	0.62
	0.17	0.66	0.17	0.17	0.05	0.67
Social	0.97	0.83	0.50	0.40	0.50	0.47
	0.17	0	0.50	0.23	0.17	0.325
	0.61	0.21	0	0.10	0.20	0.55
	0.94	0.86	0.43	0.10	0.07	0.45
	0.77	0.84	0.77	0.13	0.06	0.37
	0.98	0.87	0.64	0.03	0	0.43

Although the analysis scores in Table 4.5 represent the subevents, these results were not at an acceptable level. It is because the physical sensor analysis score did not influence the social data analysis score. Each city's scores were independent. Therefore, we created a framework for these subevents through influence maximization [5]. This is done by knowing the social network information and influential user nodes to maximize the relative social event influence.

The increase in confidence for event influence is achieved through reinforcement learning. This maximized confidence score is represented as the Response Factor Score as shownin Table 4.6. The results show social and physical sensor warning levels of a hurricane as the value increases, higher the possibility of an imminent threat from the hurricane. In this stage, we considered social analysis score as an opponent of physical sensor score. We defined the environment as influence propagation, but the strategy of the social opponent is not defined. It is to achieve a more realistic solution [31]. The influence metric from each analysis method helps to generate the training data. However, our goal here is to create decision/response score for the subevents and find severity in anomalies from the entire word-cloud.

4.10 Model Limitations

Although our framework successfully determined the sublevel events and anomalies with reasonable accuracies, several issues need further investigation. In this study, data cost was not considered. The framework requires the collection of several daily data and is prone tohigh error if the threshold values of the physical sensor change significantly. This is because we are adjusting a few parameters to avoid biased learning. We are also normalizing our feature selections whenever our configuration was getting an influence factor of zero.

Table 4.6 Response Factor for Datasets.

Datasets	**Decision/Response Score**						
	TS	PS	AS	LS	IS	RS	BE
Social	0.68	0.87	0.93	0.69	0.83	0.62	0.65
Physical Sensor	0.75	0.90	0.89	0.83	0.82	0.87	0.60
Social / Physical Sensor	0.85	0.92	0.93	0.77	0.94	0.52	0.57

4.11 Summary

This paper discussed the most reliable method for receiving critical data with highaccuracy. We introduced analytical algorithms to combine physical sensors data with social data pertaining to action, emotion, and location information of critical events during extreme weather emergencies. The results presented here have a combined accuracy of 86% of decision score to define emergency condition in an area or location with the help of Twitter database and national weather organization data. The observation made in this paper gives subevent (critical) information with higher accuracy byutilizing aspects of diffusion and dissemination. The reliability of disseminated information in many respects improves social awareness in the public at a faster rate. We investigated the Twitter database for hurricanes only by tracking social participation during the outcome period. It helped us to understand the characteristics of social commitment during a hurricane or natural disaster. Although in our method, a primary source of error comes from the fact that we recognize the single influence characteristic instead of combined weather influence characteristics.

For some physical sensor data, the classifier was able to predict if certain extremeweather conditions were absent or not, even though those predictions not considered in the analysis. The classification in this paper cannot be directly related to social data as the datapoints cannot be identified in respect of the exact local location of tweets and weather statistics ata given time. This method is an added feature for the current system framework to justify the real-time scenario in disaster or anomaly situation. Also, we collected sensor data from government organizations during the same period to corroborate the results. We examined the dissemination of information through multiple methods of news, weather agencies, government agencies, organizations, and the public. We have also formulated the classification in the network to determine the decision employing tweeted words. In this paper, we examined the overall importance of social data and the dissemination of physical data by categorizing the wordcloud.

The study can be extended to learn and create a virtual network from structural information of a network, even if certain real data is unavailable. This can be done by creating a model for policy determination for newly evolved networks since our framework utilizes scores to influence entire network for policy and reward determinations. We plan to utilize multiple category data for different weather situations and pollution by providing a framework to estimatedifferent scenarios. Currently, our algorithm works on a single environmental condition with one type of data stream at a time. We want to investigate further whether the trained model is transferrable for different streams rather than learning from the start. We would also want to implement Pareto optimization to handle the situation for stochastic outcomes.

Chapter 5: Case Study on Causal Cooperation and Application on Disease Spread[3]

5.1 Background

Cooperation exploitation is a viable test and poses the widespread demand in human societies constituting a network of engagement. However, how can the interacted information be evolved to create a populous cooperation structure? Accounting to the fact of possible interaction of individual users which are constituted in a network which interact to selective neighbor leading to natural reward and cooperation in the accordance of game theory model [10-12]. This conditional reward in accordance with game theory model establishes interactions to produce a clustering strategy which are termed as local interactions with nonrandom cooperation's. The results of these interaction promote cooperation with better reward systems and relative structure of these cooperation can be seen in various behavioral experiments [12-17]. One of the key things in the behavioral experiment were the dynamics in a social network. A social network if missing a dynamic, it often constitutes as biased network. Thus, a proper prior network is the one which has dynamics that can afford the clustering opportunities for different strategies in response to a conditional and non-conditional event. Hence, this opportunity to create variable action in a dynamic network can be defined as cooperative attention in a network.

However, cooperative attention in a network constitute evolution and creates a behavioral reciprocity [1,19,20] which is aligned with strategy-based game theory. This theory thus defines reciprocity as a dependent source of interactions where one's actions towards a user depends on

[3] This chapter is also published as Srivastava, Harshit, and Ravi Sankar. "Cooperation Model for Optimized Classification on Social Data." 2022 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2022.

the past reactions. Thus, creates a new problem dimension when the users are more than three [21-25] which we know as two player game theory. Hence to reduce the reciprocity, a strategy is which we know as two player game theory. Hence to reduce the reciprocity, a strategy is required. Thus, a solution is incorporated to formulate an unbiased solution where the past reactions are not only a dynamic changing factor. This is done by adding another feature of neighborhood bond which can dynamically change with or without the past interactions between the three or more users. This neighborhood bond can be defined as a link to establish reciprocity where users can engage and disengage with unbiased reward distribution.

In recent year, the game theory models showcased the reciprocity concept [36] that demonstrated the ability to link and promote the interaction in a group. These interactions can also promote cooperation in a dynamic network. The dynamic network models directly support the matched cooperation support to dynamically determine changes in theproposed node connections. This supports non-sequential link updates for the node cooperation, thereby creating failure in adaptation of the network [27,30,33]. If we dig deeper and investigate these networks, we will notice that slow and static network connection variation provides lower heterogeneity than the rapid changing network. Therefore, to have stable and proper predicted connection where non required node connection should have stable connection probability whichcan be done by creating stable bonds based on diversified data sources.

The objective of this study is to understand and explore cooperative strategy. The cooperative study exploration is done by bridging diversified data sources of social network and physical network to make disease spread decision.

The decision spread model block diagram in Figure 5.1 represents the stages of analysis of social and physical data to find the overall spread factor. The block diagram is divided into four stages. Stage 1 represents preprocessing and semantic integration of social data by creating attention-based user category in respect of physical data events. Stage 2 is physical data analysis through Susceptible Exposed Infected Recovered (SEIR) model which provides essential information like hyper parameters for events, infection factor, location spread, and change in spread rate. Stage 3 represents Cooperation spread and Control Model [5] to determine spread variable and Stage 4 represents policy determination and optimization for analysis.

Individuals incur costs to help others when cooperating and this action is a key component of a human community network. [1-4, 9-13]. Evidence indicates that people's social interactions affect them because of it, feelings, thoughts, and behaviors will disseminate through networks [15-23]. As a result, the issue of whether collaboration spreads by social contagion arises. This is an important issue with practical consequences for strategies intended to encourage cooperative activity. However, it can be hard to differentiate across spreader and homophily (the inclination

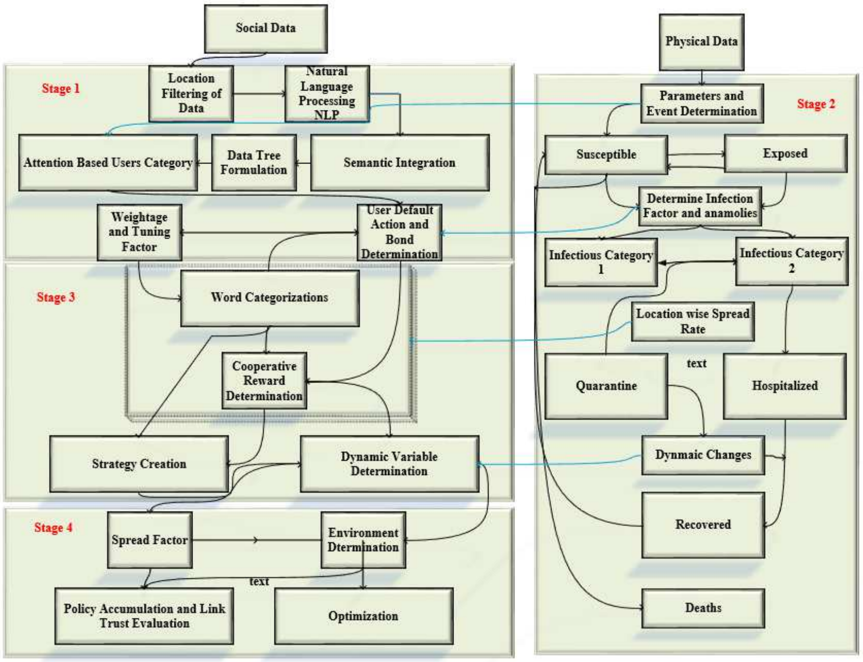

Social Data
Physical Data
Stage 1
Location Filtering of Data
Natural Language Processing NLP
Parameters and Event Determination
Stage 2
Attention Based Users Category
Data Tree Formulation
Semantic Integration
Susceptible
Exposed
Determine Infection Factor and anamolies
Weightage and Tuning Factor
User Default Action and Bond Determination
Infectious Category 1
Infectious Category 2
Stage 3
Word Categorizations
Location wise Spread Rate
text
Cooperative Reward Determination
Quarantine
Hospitalized
Strategy Creation
Dynamic Variable Determination
Dynmaic Changes
Stage 4
Spread Factor
Environment Dtermination
Recovered
text
Policy Accumulation and Link Trust Evaluation
Optimization
Deaths

Figure 5.1 Decision Spread Model Block Diagram

for people to establish and sustain relations with people who are close to them) [21], [23], [27]. Previous research has used observational data to find the relation between contagion and homophily, which resulted in observation on homophily through unobserved traits. This unobserved traits in a network are not easily observable through statistical learning [23], [27]. Therefore, controlled data experiments is used with fixed sample size to tackle the problems to differentiate the foreknowledge of their nodes interaction in specific network structure.

Social contagion develops cooperation and might occur in the context, as shown in a recent study [33]. The research took social spreading for cooperation to incorporate data from a controlled sample size to theorize the game theory model for a sample of social data. After each round of the observation, users were allocated at random to engage with new groups of stranger's nodes who chose the reward to accumulate for specific node/nodes in a network. This resulted minimization of biasedness by restricting nodes previous behavior over their current bonding. Despite these challenges, in later phases, nodes who were allocated to large groups with relatively large reward contributors shared more rewards in the network. This result shows that contribution activity will spread uniquely in a network of arbitrary users in the absence of bias.

Evidence indicates that cooperation game activity is infectious in static social networks, whereas homophily in a network is eliminated by set of nodes if nodes are pre-bonded with their specific neighbors in each sample size. In the case we have multiple users, a multi-player strategy can be utilized, this strategy mirrors the prisoner's dilemma problems in static networks with different systems to find cooperation [33-35]. This cooperation is calculated in the prisoner's dilemma by a binary choice of cooperating or fleeing. If we look into the real strategy game, we strategize cooperation and assess it as a variable. This variable should be continuous in nature as the variable value depends on each user's behavior in a binary way (cooperative or selfish) resulting in bias spread. Cooperators who were matched with more defecting neighbors in subsequent rounds were more likely to turn to defection in subsequent rounds across all simulations; poor behavior was persistent. Dissenters who were partnered with comparatively more friendly neighbors, on the other hand, were not more likely to turn to cooperation. These studies add to the growing body of evidence that social activity in a network in the sense of cooperation will spread from one consumer to the next and that this effect can be extended to fixed networks. They also propose that the degree to which cooperative and selfish behaviors are infectious can differ. Participants in the fixed network were aware not only of their neighbors' decisions, but also of their overall payout as a result of those decisions. Assuming that protester's performance in reward accumulation is more than cooperators, this can lead to improper reward distribution among the cooperators: rebels with large number of cooperative neighbors' nodes will detect extra link to swap, but this impulse link can lower the overall required reward for the node. As a result, further research is required to see whether cooperative activity will spread without knowing the specific reward distribution. Furthermore, it is essential to investigate the propagation of collaboration in dynamic networks rather than fixed networks. Networks have power over their relationships in many social experiments, since they are able to break old links and form new ones.

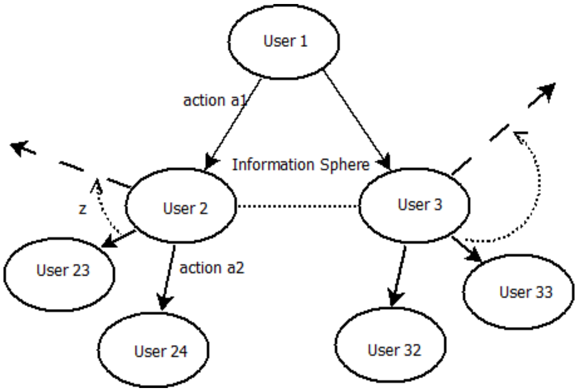


Figure 5.2 Action and Reward Cooperation

In comparison to static networks, dynamic networks provide a broader variety of techniques like cooperation, consensus mechanism and information diffusion etc. As a result, the strategic environments of fixed and dynamic networks can promote different approaches; that is, in fixed and dynamic networks, the contagion of cooperative and selfish activities may behave very differently. When active user nodes engage frequently in comparatively fixed social networks, one of their primary goals may be to reconcile the conflicting interests of effectively cooperating with others (as cooperative cooperation is superior to mutual defection) and (ii) preventing free-rider manipulation (as defecting with a defector is preferable to cooperating with a defector). Reactive tactics or responding to the interaction partners' actions by cooperating when they are agreeable and defecting when they are not, are a typical approach to this problem. People appear to defect unconditionally or play conditional tactics in repetitive cooperative games [34- 37], [41].

In comparison, another aim emerges in fluid social networks: recruiting new cooperative interaction partners. Individuals may be encouraged to attempt cooperation even though their existing relationship partners are relatively uncooperative if they feel (correctly) that cooperators are more likely to establish relations with them when they cooperate. As a result, in increasingly updating social networks where there is a significant potential to draw new cooperation partners, we will be able to anticipate less of an association between one's current neighbors' behavior and one's own future actions.

Here, we provide a solution to test these cooperative attentions to learn different actions and predict by asking how the spread actions through different sources in a social network where the individual node behaviors depend on that node social actions and connectivity with other nodes. We explored this issue using the current pandemic data of COVID 19 collected from Twitter Social Network and Physical Data available on Johns Hopkins coronavirus data source [42]. In this analysis the connection control is conditionally varied from one user node to other. This is done with the help of the dataset to find complete information, allowing us to decide on dimensional approach to separate cooperation with contagion across time even when biasness is possible as of the nature seen in homophilic structure. We utilized this dataset to separate the cooperative and selfish nodes actions and reactions in the dynamic social network. We optimized and worked on different rules to structurally define the evolving analysis to understand the social interaction in determining the information spread.

5.2 Cooperative Learning

The complete interactions and dependability formulate collective functionalities in a complex network. It is observed that the full cooperation depends on the topological structure of a network with temporal constraints on information links which constitute the dynamics of the network. The information links are reaction information which evolve with time and are registers with series of activated events at discrete time. The linked sequence of information dissemination is a state of causal flow which affects the characteristics of a social network. These characteristics redefine the network structure that includes clustering, node, controllability, and link length. These can show static and irregular patterns of inter-burst of temporal links. Thus, the systematic encapsulation of these inter-burst temporal links resolves the cooperative decision-making problem in multi agent network and require multiagent cooperation for optimized long-term collaboration.

In general, a cooperative setting the total optimized return is always summed to zero and when competitive setting is integrated the total optimal return is observed to be non-zero and has some returns. The multiagent cooperation equilibrium points are critical as adversarial nodes are always present in a network thus, require policies for nodes without increasing joint action space to reduce scalability issues [18], [30]. Assuming the nodes are choosing actions a^i after observing the system link states simultaneously to distribute the reward. Thus, the agents can make decision and accumulate reward as shown in Figure 5.2.

Assuming a variable space containing t states as s_t with n actions as a^n, hence, the probable transitions in respect of a Δs_t are stated with reward (γ) as $P = s_t \times a^n; \gamma = s_t \times a^n \times s_t$. Thus, each time t the user chooses action $a^{n,t}$ which causes $s_{t+1} \sim P\left(\frac{Event}{s_t} \times a^{n,t}\right)$ to accumulate reward as $\gamma(s_t \times a^{n,t}, s_{t+1})$, therefore the function can be defined for n nodes with tradeoff factor $\omega \in 0 \rightarrow 1$.

$$\beta(s) = E[\textstyle\sum_{n>-1}^{t\geq 0} \omega\gamma(s_t \times a^{n,t}, s_{t+1}), s_0 = s] \tag{5.1}$$

Hence, for multiple nodes,

$$\beta^i(s) = E\left[\textstyle\sum_{n>-1}^{t\geq 0} \omega^t \gamma^{i,n}(s_t \times a^{n,t}, s_{t+1}), s_0 = s\right] \tag{5.2}$$

Now, to have optimal reward distribution policy in a cooperative setting $\gamma^{i,n} = \gamma^{i,1} \ldots\ldots\ldots. \gamma^{i,N}$, and to obtain global optima, Nash equilibrium is utilized by averaging the reward, $\sum_{i\in N, t\geq 0, n>-1} \gamma(s_t \times a^{n,t}, s_{t+1})$ Thus, to have a proper distribution and link formation policy is created by applying a competitive setting $\sum_{i\in N, t\geq 0, n>-1} \gamma(s_t \times a^{n,t}, s_{t+1}) = 0$ to define link reaction policy (π)

$$\pi(a|s) \coloneqq \textstyle\bigcap_{i\in N, n>-1, t\geq 0} \pi^{i,j}\left(\frac{a^{n,i\rightarrow j,t}}{s_t}\right) \tag{5.4}$$

where, i and j are node links and x_i is the neighborhood bond. Hence, optimal reward link is,

$$\beta_{\pi^i}^{i,j}(s) = E\left[\textstyle\sum_{n>-1}^{t\geq 0, i,j\geq 0} \omega^t \gamma^{i,n}(s_t \times a^{n,t}, s_{t+1}) \left| a^{n,i\rightarrow j,t} \sim \pi(Event|s_t), s_0 = s\right.\right] \tag{5.5}$$

Now optimal joint cooperative policy $\pi^{i,j}: s_t^{i,j} \rightarrow \Delta a^{n,t}$ for i, j will be dependent on actions history $h^{i,j}$,

$$\tau_{\pi}^{t} = \prod_{h':h'a\sqsubseteq h} \pi^{p(h')}\left(a\middle|M(h \rightarrow S)\right): \tau_{\pi^{i,j}}^{t} =$$

$$\prod_{N\cup p(h')}^{i,j>0} \prod_{h':h'a\sqsubseteq h}^{n} \pi^{i,j},^{p(h'\in i,j)}\left(a^{i}\middle|M(h \rightarrow S)\right) \quad (5.6)$$

where, $p(h) \sim p(h') = P$ is the probability of an action user node takes for a policy π^{P}.

$$\beta_{\pi^{i}}^{i,j}(s) = E\left[\sum_{n>-1}^{t\geq 0,i,j\geq 0} \omega^{t}\gamma^{i,n}(s_{t} \times a^{n,t}, s_{t+1})\,\middle|a^{n,i\rightarrow j,t}, x_{i}(t)\sim\tau_{\pi}^{t}, s_{0} = s\right] \quad (5.7)$$

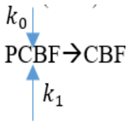
This marked departure changes the dynamical processes that occur on networks, such as cooperative evolution Identifying the temporal interaction pattern is, without a doubt, the first step in comprehending and regulating the mutual complexities of temporal networked networks.

5.3 Cooperative Learning and Strategy Creation

Cooperative learning states the variability in network with the help of reciprocity to maintain the dynamics of the information for unbiased reward distribution by creating a strategy tie for consensus in any two states. This learning strategy constitutes the continuous communication dissemination in a network with continuous and discrete data sources. For example: Consider a problem of cluster of users that are randomly interacting to another cluster of users at any instant. Additionally, the length of the action link and time is not capped (time limitation for each links). However, the reaction can change according to the reciprocity of the user node reactions thus the users link needs to come into consensus for reward distribution.

Hence to do this assume that each user node has an information link of x_i where i represent the i^{th} information of reward reciprocity, each user determine the length of time the communication occurs and sets as $x_i(0)$ and communicate through a directed graph and undirected graph (ρ, ε) [32)], where $\rho = \{1, \ldots.n\}$ user nodes and $\varepsilon \subset \rho \times \rho$ is an edge set of ordered pair of nodes . Assuming the edge $(i, j) \in \varepsilon$ denotes the user node j which obtains information from i (not vice versa!) (directed) and undirected vice versa works.

Therefore, the amount of information flow is proportional to an accumulated amount of reward link formation. This link accumulation constitutes of reactions, actions, and neighbor bond. The dynamics of accumulation of this is a dependent on a consensus breaking factor (CBF) and is designed as in the following scheme.



where, PCBF is a messenger of CBF, and k_0 and k_1 are rate of constants of influence and de-influence respectively. Hence, the establishment of links depends on the rate of reactions with the dependence of neighbor bond. Thus, neighbor bond is given as,

$$Neighbor\ Bond x_i(t) = -information\ graph\ \left(\mathbb{G}(t)\right) * \\ information\ state\left(x(t)'\right) \tag{5.8}$$

where, $\mathbb{G}(t) = [\mathbb{p}_{i,j}(t)]\epsilon\mathbb{R}^{n\times n}$ is a Laplacian communication flow (Laplacian metrics with fixedarea occurs at a metric of constant curvature and, for negative Euler characteristic, exhibited a flow from a given metric to a constant curvature metric along which the determinant increases) [24] $\mathbb{p}_{i,j}(t)\ information.$ and $\left(x(t)\right)$= [$x_1 \dots \dots x_n$] at any state,

$$communication\ flow\ rate\ k_{0,i} \\ = (rate\ of\ reaction\ coefficient)\lambda_{0,1} \\ * \exp((Activation\ state\ of\ Bond)x_{0,i}(t)/dynamic\ event\ change\ \Theta))$$

$$\beta_{\pi^i}^{i,j}(s) = E\left[\sum_{n>-1}^{t\geq 0,i,j\geq 0} \omega^t \gamma^{i,n}(s_t \times a^{n,t}, s_{t+1})\,\middle|a^{n,i\to j,t}, x_i(t) \sim \pi(Event|s_t), s_0 = s\right] \tag{5.9}$$

We developed a social network formation model with large number of parameters and latent variables. We must first allocate values to the unknown variables before we can test the model's validity. We learn the endowment vectors using real-world network observations, assuming real-world networks are at or near pairwise symmetry, to equip our model with the capacity to match real-world networks.

$$k_{0,i}^{\pi^i} = -\lambda_{0,1}\, exp(\frac{\beta_{\pi^i}^{i,j}(s)}{\min\max(\sigma,\mu^{i,j}),}) \tag{5.10}$$

where, convex and $\mu^{i,j}$ as concave joint event evaluation. Hence, the optimal accumulated strategy will be defined as, $k_{0,1\to\iota,J}^{\pi^\iota}(t) \doteq -\sum_{J=1}^{n} m_{\iota J}\left(k_{0,1\to i}^{\pi^i}(t) - k_{0,1\to j}^{\pi^i}(t)\right), i = \{1, \ldots .n\}$

where, m_{ij} defines the communication link topology.

5.4 Spread Based Analysis for Cooperative Learning

Recent pandemic situation for COVID 19 require an optimal control of spread of disease which can commune with person-to-person contact. Thus, the question arises how effectively one can track and mitigate the spread of any commutable disease. Hence, we attempted to design a spread model with the help of cooperative learning by utilizing social and physical network data.

The spread analysis model is defined by a reward optimization of a social network with the help of physical network data as shown in Figure 5.3. The model represents a basic susceptible, exposed, infectious and recovered stages in which the probable infections are defined with respect

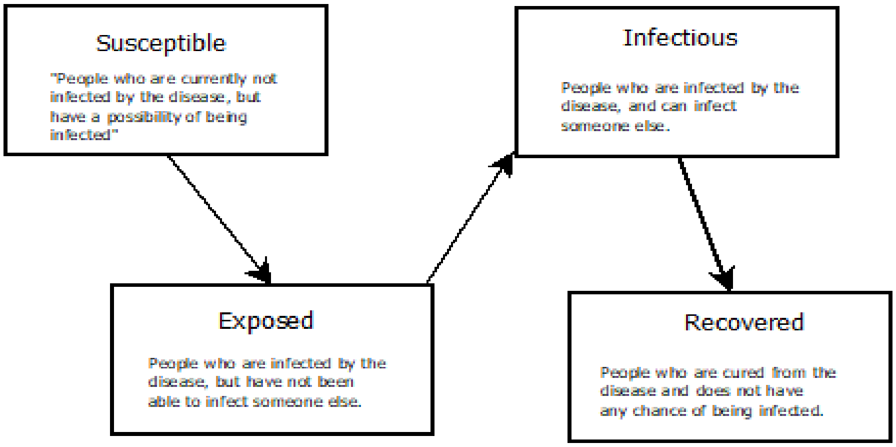

Susceptible
"People who are currently not infected by the disease, but have a possibility of being infected"
Infectious
People who are infected by the disease, and can infect someone else.
Exposed
People who are infected by the disease, but have not been able to infect someone else.
Recovered
People who are cured from the disease and does not have any chance of being infected.

Figure 5.3 General SIR Model for Disease Spread

to event, situation, or state. In addition, in a specific event, we propose that a cooperative strategy learning to be embarked as a low reward with respect to reward accumulation dynamics. However, to create a consensus in reward distribution, CBF is selected for any unmatched micro events

$$exp(\frac{\beta_{\pi^i}^{i,j}(s)}{\min\max(\sigma,\mu^{i,j}),}) \times L_t\ (cummlative\ infection\ location\ rate) \times \eta_t\ (Dynamic\ Environment\ Variable) \quad (5.11)$$

respect of macro event. Hence, the reward spread is defined as,

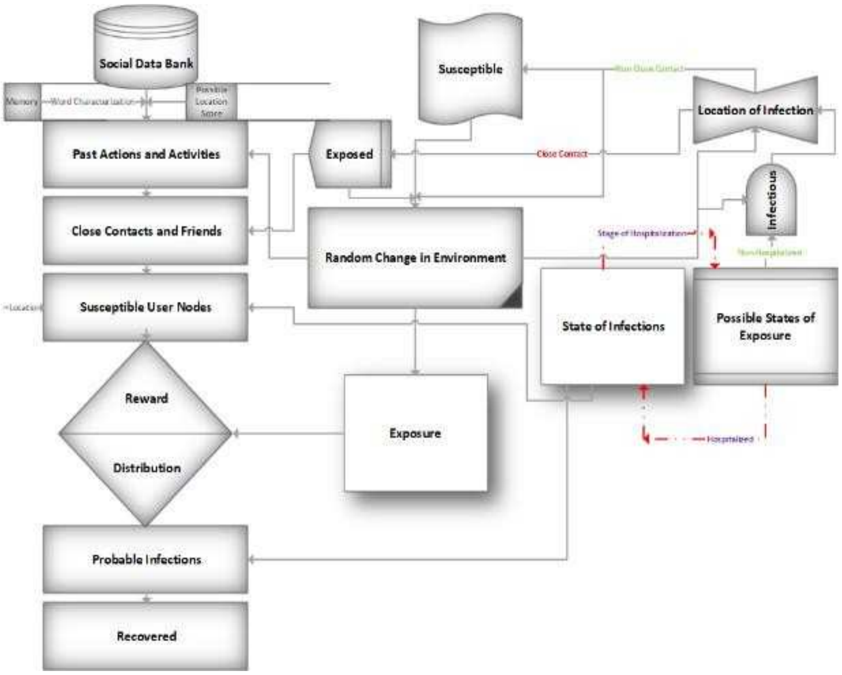


Figure 5.4 Adaptive Algorithm of Spread

Therefore, the spread process is defined as,

$$Spread\ \xi(s) = max_{\pi^i}\ [exp(\frac{\beta_{\pi^i}^{i,j}(s)}{\min\max(\sigma,\mu^{i,j}),}) \times L_t\ (cummlative\ infection\ location\ rate) \times \eta_t\ (Dynamic\ Environment\ Variable)] \quad (5.12)$$

In this, a transition to a next state is unknown, which is usually defined by a new generated state function that identifies the best actions to form links and greedily chooses the action from which it gains maximum reward. This is achieved through adaptive algorithm as shown in Figure 5.4 over all infectious state in a dynamic environment.

5.5 Proposed Variable Detection

The dynamic variables selection is best used in statistical test or criteria where automatic variable selection methods are utilized to best fit the sample according to statistical information statistical information criteria which include stepwise regression and shrinkage methods. Whenever a variable is selected, pros and cons are always considered as a potential perpetuator to define and select a model. This is done to justify the global optimal criteria in which variance of outcomes changes over time and increases the computations and we know with increase in variables the model expands exponentially. Thus, when a time series data constitutes dynamic

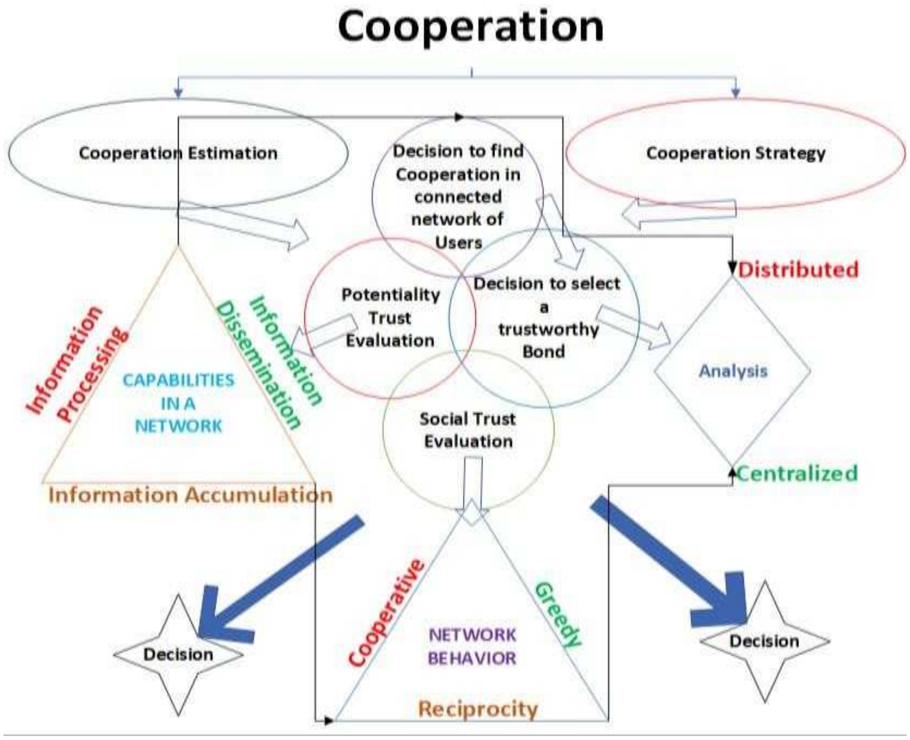


Figure 5.5 Link Trusts Evaluation after Influence Score

variable, the correlation will likely generate the nonsensical relations affecting accuracy and precision [32-34].

In the dynamic social network, the data source represents discrete time stamps for a specific period which are interconnected nodes with high probability of strong reciprocity link thus creating a dynamic cluster. However, this dynamic cluster creates a problem which makes a social network biased over the long-term as the smaller perturbation leads to major change in relations and bonds in a network which can be determined as a variance from the present state. Hence small changes we observe to higher variances of link reciprocity due a change in a network and unnecessary links (noise). Moreover, to detect these temporal links cooperative strategy is utilized for optimal information dissemination.

To maximize the objectivity of these strategy an optimization of the function compositing two or more variables results in better network topography by defining cost difference of the objective function. Assuming the linear objective function is given by the changes in nodes evolution x over the period of t, $\frac{dx}{dt} = k_i - k_{i+1}x$. If the variable environment is stationary for a given time interval the parameter k_i and k_{i+1} will result in constants resulting in,

$$x(t) = x_m - (x_m - x_0)\exp(-k_{i+1}t) \tag{5.13}$$

where, a steady state $x_m = \frac{k_i}{k_{i+1}}$ and $x_0 \sim initial\ value$.

Therefore, the variable x reaches its critical vale $x = 1$ for the first time at $t = \psi_0$ obtaining,

$$\psi_0 = \left(\frac{1}{k_{i+1}}\right) ln x_m - x_0)/(x_m - 1)] \tag{5.14}$$

Thus, the next threshold will be at, $\psi = \left(\frac{1}{k_{i+1}}\right) ln x_m)/(x_m - 1)]$ Hence, gives for T time toreach N threshold,

$$x(T) = x_m\{1 - \exp[-k_{i+1}(T - \psi_0 - (N-1)\psi)]\} \tag{5.15}$$

Hence, if $x_m \leq 1$ it will never reach the threshold value. Similarly, in equation 11 assuming variable $\eta_t \, for \; i$ and j

$$\dot{k^{\pi^\iota}_{0,1\to\iota,J}(t)} = -\sum_{J=1}^{n} m_{\iota J}\left(k^{\pi^i}_{0,1\to i}(t) - k^{\pi^i}_{0,1\to j}(t)\right) \tag{5.16}$$

Hence, the optimal cooperative spread will be defined as,

$$\eta_t{}^{i,j}_{\pi^i} = \exp\left[\frac{k^{\pi^i}_{0,1\to j}(t) - k^{\pi^i}_{0,1\to i}(t)}{\min\max(\mu^{i,j})}\right]$$

$$O_t^{\pi^i} = \sum_{t\geq 0}^{i,j\epsilon N} \frac{\left\{\eta_t{}^{i,j}_{\pi^i} - \eta_t{}^{1,0}_{\pi^0} - \cdots \ldots - \eta_t{}^{n,n+1}_{\pi^n}\right\}}{\left\{\left\{\eta_t{}^{j,i}_{\pi^j} - \eta_t{}^{0,1}_{\pi^0} - \cdots \ldots - \eta_t{}^{n+1,n}_{\pi^n}\right\}\right\}} = \left\{1 - \frac{\exp\left[-\eta_t{}^{i,j}_{\pi^i}(1-\chi_{i,j})\pi^{p\left(h^{i,j}\right)}\right]}{1-\exp(-\sum_{i,j>-1}\left[\eta_t{}^{i,j}_{\pi^i}(\sigma(t))\chi_{i,j}\right] + \eta_t{}^{i,j}_{\pi^i}(\sigma(t+1))(1-\chi_{i,j)})\pi^{p\left(h^{i,j}\right)}}\right\} \tag{5.17}$$

$\chi,$ is a critical state of action determination.

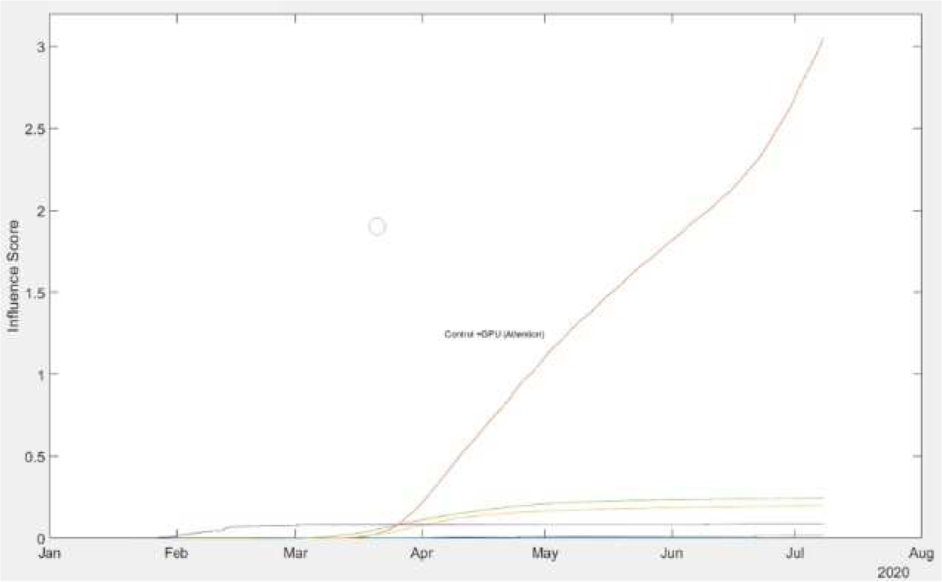


Figure 5.6 Influence Spread

The method proposed aims to find the network configuration at any time using the clusters extracted in the previous timestep. This introduces a two-stage event-based adaptive algorithm, as shown in Figure 5.4, that uses an event-tracking system. At each timestep, we use the last timestamp's collected associated components of the spread as the initial information for the state. The distribution around the seeds is determined by optimizing the ratio of the average internal external degree of information of the local cluster. The bursty existence of social networks drives the complexities of many social and economic phenomena [36]. This event- based spread implicitly recognizes that the links in a network change over the specific time period. This spread analysis drives the bursty nature of social network [36] where the dynamicsof social and economic impact is compared for spread analysis.

Social network analysis as a means of analyzing communications and relations in groups to discuss some of the various measures to determine awareness in a spread in a network. In order to determine situational spread, it is necessary to have a clear picture of who they are a cumulative infection location rate ($\boldsymbol{L}_t$). Therefore, to determine this information fusion is utilized to deal with determining the relations between the objects. In these kinds of contexts, one way to reduce the amount of knowledge provided to the user is to identify classes of objects depending on their capacities or properties [18]. However, in some cases, it might be more useful to identify the observed entities solely on the basis of their relationships with other entities. This relation is measure through a similarity computation by using Cosine model for a CF model [33].

$$L_t\left(x_i(t), a^i, u\right) = \frac{\sum_{u^j}^{i,j\epsilon N} UserSim(u^i,u^j)\times r(x_i(t),a^i,u^j)}{\sum_{r(x_i(t),a^i,u^j}^{i,j} |UserSim(u^i,u^j|} \tag{5.18}$$

where, u gives a location rating for the spread for every action. To analyze this spread of action reward influence score is observed by utilizing equation 5.7 which is determined with the help of equations 5.8 and 5.9. The important aspect to analyze in equation 5.7 is to find the probable location score for the susceptible nodes and optimized cooperative spread with strategy. This is done by analyzing twitter data on COVID 19 and physical data of positivity from Johns Hopkins.

The social data was analyzed by grouping the word categories into COVID positive emotions of users to determine the susceptible nodes. The analysis created a sampled space which with actions and spread homogeneity but does not comprehensively show the probable locations effecting it. To determine this, we utilized the physical data during of the same sample timestamps as of susceptible users to determine the rate of positivity in specific city with zip codes and gave similarity scores to map the of exposed user's information. The results can be seen in Table 5.1, in which action influence shows a relationship with location influence and spread influence shows a cumulative relationship. This influence space score is used to analyze the influence spread as can be seen in Figure 5.6. The influence spread sampled space is generated according to the action space.

5.6 Discussion

In the current analysis, we utilized physical location-based datasets of COVID-19 United States and social data set from Twitter to identify the measured changes in information dissemination and behavior. The social network behaves as a behavior function and this behavior is derived from the probabilistic change in actions of user nodes during the change in states in social network. We examine cooperation and non-cooperation states in terms of transition probabilities. This examination of transition probability is then utilized to encapsulate the cooperating strategies. This approach to examine behaviors is focused by measuring the spread behavior from diverse data sources in a social network to adapt the action frequencies by employing an infectious disease framework to study social contagion [5,27,28].

As we know the social and physical network establishes a multiagent system which is based on decentralized cooperation. It represents the nodes to forward the critical information about the events or situations in possible conditions. The critical information in the form of words is categorized in ascending order to incorporate restructuring of the physical data, which isfiltered with respect to situations or events by correlating it with location variable at sample time interval of corresponding social data.

Table 5.1 Influence Space Score

Sampled Space	Action Influence	Location Influence	Spread Influence
13356	0.11	0.63	0.09
597	0.62	0.42	0.12
21036	0.31	0.75	0.45
3166	0.33	0.31	0.81
49756	0.71	0.76	0.16
18570	0.07	0.87	0.77
999	0.16	0.51	0.81

The data structuring of physical data is done in respect of social data also indicated the importance of actions and reactions. This is the raw critical information categorized from real time data from physical data. Assuming each user network node $u \in X$, where X is a set of all the network nodes is and have a capability perceive a local and global directed and directed path for link reciprocity. Each node receives x_i observation as y^i via a noisy observation link $A^i: s \rightarrow p(y^i)$ such that the node i observes $y^i \sim A^i(Event|s)$ a random variable for the environment state. Thus, the collective information links is defined as,

$$I_t^i = \left\{A^i, \left(O_t^{\pi^i}\right)_{j \epsilon N} \left(Spread\ \xi(s)^{j \rightarrow i}\right): (t = 0 \ldots t - 1)\right\} \cup \{A^i{}_t\} \qquad (5.19)$$

$$
\begin{aligned}
&\mathrm{K}^{g\epsilon p}\left(s_{p,t}, I_t^i = \rho\right) \\
&\quad = \max_{g}.\Bigg[\sum_{i=1}^{\rho} l_i * \tau_i * Im\big(lat(i)\big) + (1 - \lambda_1 \\
&\quad - \lambda_2).\sum_{t=0,}^{T}\sum_{i=1}^{N}[x_i * (\gamma, t+1) * Im\big(Cat(i)\big)] + A_{\gamma,t}^{t+1} * \{\lambda_1 * C_t \\
&\quad * \sum_{j=1}^{m} CT_{Score(j)} * y_j.* \Big(Im\big(Cat(i)\big)\Big) + \lambda_2 * C_E \sum_{k=1}^{p} SS\,(k) * z_k \\
&\quad * max_{i \in T, SS_k}\Big(Im\big(Cat(i)\big)\Big)\}\Bigg]
\end{aligned}
$$

notice the information size increase as t increases, constituting the memory utilization issue which is solved by defined the latent space by weight sharing and attention mechanismwith defined policy according to the multi agent environment.

In this study, we processed Twitter tweets data collected on COVID 19 from January 2020 to October 2020, while focusing on the result analysis on New York, Florida, and San Francisco. The number of tweets collected and considered in this time frame is 185,755 giving word-cloud of 76,781 with an average 65 words in each tweet. Whereas the physical data was collected and evaluated on SIR model [31] to define rate of infection and map accumulatedlocation infection rate.

5.6.1 Observation Model

The observation is set up for cooperative setting with partial observability. This setting is decentralized in nature where nodes share reward with all the other nodes in respect of reward function and transition model with the difference of neighbor bonds except the nodes in the network has local observation for any state s. This model starts with no reciprocity link to other agents and does not maintain a global belief vector. This model is solved by finding the influence data to find observation points for all the nodes and then optimized by defining policies using the influence data [5] which maps lo cal observation histories to actions to find predict spread. Monte Carlo Tree Search (MCTS) and sampling for policy iteration.

Table 5.2 Symbols Illustrations

Symbols	**Definition**
$\boldsymbol{\rho}, \boldsymbol{s_t}, \boldsymbol{l_i}$	Threshold values of physical sensor data, State of the event at time t, Indicated variable for time-based sensor data.
$\boldsymbol{N}$	Number of tweets considered for summarization (in the time window specified by user)
$\boldsymbol{T}, \boldsymbol{m}, \boldsymbol{p}, \boldsymbol{s_{p,t}}$	Total time, Number of distinct content words and subevents included in the n tweets respectively, State of subevent at time t
$\boldsymbol{msize}$	Number of tweets containing distinct words
$\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}, \boldsymbol{a}$	Index for tweets, Content of words, Subevents, Classes, respectively
$\boldsymbol{x_i}$	Indicator variable for tweet i
$\boldsymbol{y_j}$	Indicator variable for content word j
$\boldsymbol{z_k}$	Indicator variable for subevent k
$\boldsymbol{CT\_Score}(\boldsymbol{j})$	Feature score of content word
$\boldsymbol{SS}(\boldsymbol{k})$	The score of subevent k
$\boldsymbol{Im}$	Importance/informative score of class a
$\boldsymbol{Cat}(\boldsymbol{i}), \boldsymbol{lat}$	Class of tweet i, Lateral averaged required data
$\boldsymbol{\tau}, \boldsymbol{g}$	Tuning parameter for sensor data, policy determination
$\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$	Tuning parameter – relative weight for the tweet, content word, and subevent score
$\boldsymbol{C_E}, \boldsymbol{C_t}$	Set of categorized words and subevents present in tweets, respectively

The MCTS actions for multi agent network are done in either predefined or in default offline state by defining action space but for our model we used search process in which actions are repeated for flexible operation for a sampled hierarchical system. In this user nodes choose actions simultaneously without the knowledge of future actions of other nodes to receive immediate intermittent reward for transition to another consecutive state though the transition states are dependent on each node's actions. Each state agent tries to maximize the cumulative reward to follow optimal policy as by using -greedy search algorithm [32],

$$\pi_t^{i,j}\left(a^i \middle| s_t\right) = \begin{cases} 1 - \varepsilon + \frac{w}{min(|X|,|Y|)} & if\ a = argmax_{a\epsilon A} = \mathrm{K}^{g\epsilon p}\left(s_{p,t}, I_t^i = \rho\right) \\ \frac{w}{min(|X|,|Y|)} & otherwise \end{cases} \tag{5.20}$$

The search space for the multi node network follows a key challenge in respect ofasynchronous decisions, flexibility, and extensive cooperation.

1. Asynchronous Decisions: The multi nodes in a network will have variable links durations and will end asynchronously which can lead to different reward.
2. Flexibility: Policy inside the network should be trained and learned to allow flexible adaptations under an event or anomaly as predefined nodes can make system biasedwith primitive actions.
3. Extensive Cooperation: While making the system flexible, the network can ignore all the primitive actions that are unsuitable for any real time change in information. Hence an option policy is initiated with respect of action space when $\pi(s|a) = 1$ and flexibility termination is 1 then $\varphi \rightarrow s$.

5.6.2 Evaluation Points with Results

The evaluation is done by simulating the network structure with the goal of learning nodes, quicker convergence and finding robust solution for cooperative and non-cooperative links. Each scenario is indexed with initial variable to a desired index.

1. algorithm requires to select what all the other nodes will link the action to communicate. In all the instances a strategy is chosen if the nodes have exchanged the messages. Hence,a random selection of starting action link assumption is made by the algorithm so that the computational cost is reduced.
2. Generating Macro Actions by Learning Scenarios: The scenario shown in Figure 5.7 is utilized by clustering the social data on specific key sentiments and utilized as algorithm's ability to simultaneously learn which nodes to be picked and how to execute the actions. The figure 5.7 represents the action of spread where all the nodes are represented as yellow green and red. The yellow color represents the susceptible nodes, green represents exposed nodes and red represent infected nodes. While if we

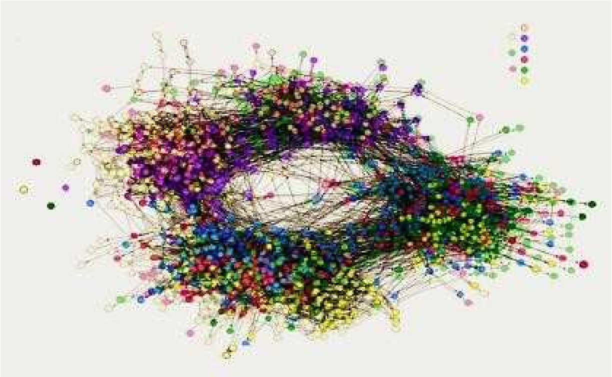


Figure 5.7 Action Spread

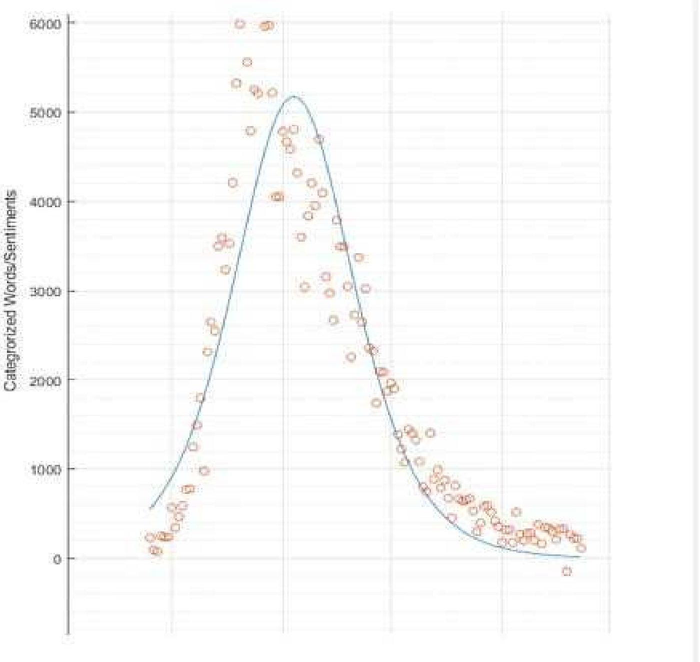


Figure 5.8 Policy Improvement and Macro Action for Decisions

look the Table 5.1, it defines the scenarios, and all the nodes' messages are controlled by MCTS with a cooperation factor. The step duration is set to 3s and a total of 5000 steps are executed with a maximum of 100 steps for planning link. This can be observed in Figure 5.7 of macros action spread in respect of categorized words that represent COVID sentiments. We can see the categorized word space increase as the word space increases with the increase in increase in action links. Further, we observe that the initial node will try to first take a lead but as soon the second node is activated, it makes cooperative decision by sharing after observation.

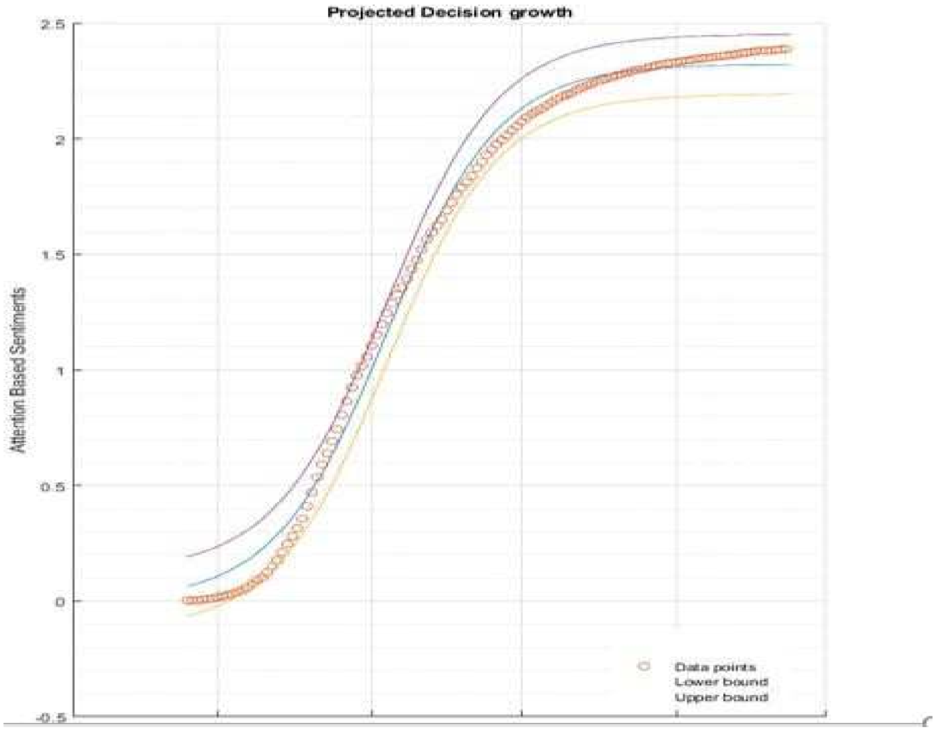


Figure 5.9 Single Node Decision Growth

3. Decision Convergence: The speed at which the algorithm makes decision is important to replicate the reward distribution. This helps to remove the action merging case for multiple nodes. Theaction merging is a state in which one or more nodes share similar actions, but the rewarddistribution is different. For example, in a scenario: assume that two nodes have 3 different decision links for reward dissemination in which the other group cluster is blocking the link. This will reduce the effective states of the other nodes in a network. To resolve this decision blocking, we implement greedy policy in which control modeutilizes cumulated tree depth reward at every iteration and determines the upper and lower level of reward thresholds within the different clusters before determining the link decision. This improves the algorithm performance though will increase the number of iterations by the 10x fold as can be seen in Figure 5.8 in which as the number of nodes increase

the decision depth of reward distribution increased.

4. Policy Accumulation: Policy accumulation is termed here as path to distribute the spread decision efficiently in the network. However, a single node at initial state will always try all the possible links for optimal information path. This possible linking before transitioning can lead to grow the search depth. Although in the current algorithm the decentralization is utilized for multiple node network in which nodes can make joint actions before depth increase in which decision link establishment is independent and leading to asynchronous decisions. The unbiasedness in the policy is proposed through non-prior information but dependence is only on random prior bond. The prior bond initiates the policy at specific stage.

To demonstrate this, we used the COVID social and physical data from the month of April 2020 to August 2020 for the state of Florida and normalized the data information for every two months for policy determination. We observed through our algorithm when policy transition at initial stage cares more about the next stage transition and the node tries to weigh the dependence of other nodes, but as the number of accumulated policies increase the policy follows the trends and predicts the policies for the day-by-day period. This policy accumulation can be observed in the Figure 5.9 and when we compare the policy accumulation with the spread of information of COVID, we observe the directional relationship with the accuracy of 67%. Additionally, policy accumulation rate is observed in Figure 5.8 and Figure 5.9 which shows the policy follows the classical trend to spread of disease and which optimizes the disease spread rate estimation as rate of spread of disease. The policy accumulation is measured with a bounded sample space and depends directly on the influence score.

The spread analysis described in this paper gives a comprehensive objectivity on reliance of social data which when effectively utilized gives a correlative analysis to understand the possible spread of a disease. The analysis is based on multi agent cooperation where every node in a network is cooperatively affecting the possible results, and this has been active and significant research topic in reinforcement learning to make sequential decision as per actions and information. The multi agent analysis requires tools from game theory and non-trivial optimization techniques which are effectively proven in different setting and applications where cooperative learning is utilized following the gradient policy and minimizing the mean square error. The analysis in this paper is attention based cooperative learning with strategy utilizing the diverse data sources by utilizing partially observed settings in which states and actions typically modelled as a stochastic game with a common reward which require generational steps to find optimal information state with all possible policies. The difficulty during this analysis lies in nodes making their own observations and making decentralized decisions although the learning curve was centralized. This led to the Nesting issue ad increased the computational cost of the analysis. The convergence result followed the vanilla policy of gradient and to avoid these few assumptions were made to verify the link quality setting in respect of neighbor bond.

Table 5.3 Decision Scores

Sample Space< 49999	Influence Score	Policy Accumulation (x1000)	Spread Decision Rate
25655 (1st)	0.035	0.50	0.05
35967 (2nd)	0.10	0.89	0.10
44500 (3rd)	0.19	0.97	0.16
48799 (4th)	0.25	1.5	0.17
49756 (5th)	0.26	1.94	0.14
18570 (6th)	0.28	2.17	0.21
999 (7th)	0.31	2.56	0.32

The study in this paper provides an essential way to utilize diverse data sources to find cooperativeness in a network. This cooperativeness is a perpetual sourced network with a permanent link of neighbor bond. We proposed a cooperative algorithm, cooperative learning, and dynamic variable to predict the spread of an information which showed direct correlation of disease spread. The correlation varies from 45% to 81% which depends on the policy accumulation with an accuracy of 67%. We found that multi node cooperative problem can be utilized to solve location determination. Future work is required in this area by expanding the objectives and inclusion of more data sources. This will be requiring more optimization so that the tuning of the system parameters can be done for higher performance.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

This dissertation encompasses the development of a framework in analyzing diverse data sources with an objective of subevent detection using an optimized strategy, where the influence score represents the anomaly score. In social media, the information propagates across the connected network. This structure of identifying information and sentiments through location and user data profiles relates to opportunistic sensing comprising social sensing, i.e., emotion and physical sensing, i.e., location. The correlation between social and physical sensor data shown in [8] effectively utilized contextual information to integrate the abstract nature of keywords. The objective of this research study is to investigate the usability of social networks during weather disasters, analyze the characteristics of a perpetual network, improve the decision and communication strategies, and facilitate the development of disaster tools. the concept of extraction amplification applied through virtual world analysis is to inherit the real experience of the physical world and its predictability.

The observation made in this dissertation gives subevent (critical) information with higher accuracy by utilizing aspects of diffusion and dissemination. The reliability of disseminated information in many respects improves social awareness in the public at a faster rate. We investigated the Twitter database for hurricanes only by tracking social participation during the outcome period. It helped us to understand the characteristics of social commitment during a hurricane or natural disaster. Although in our method, a primary source of error comes from the fact that we recognize the single influence characteristic instead of combined weather influence characteristics. This thesis discussed the most reliable method for receiving critical data with high accuracy. We introduced analytical algorithms to combine physical sensors data with social data pertaining to action, emotion, and location information of critical events during extreme weather emergencies. The results presented here have a combined accuracy of 86% to define weather or a factual emergency condition from the Twitter database and national weather organization data.

For some physical sensor data, the classifier was able to predict if certain extreme weather conditions were absent or not, even though those predictions not considered in the analysis. The classification in this thesis cannot be directly related to social data as the data points cannot be identified in respect of the exact local location of tweets and weather statistics at a given time. This method is an added feature for the current system framework to justify the real-time scenario in disaster or anomaly situation. Also, we collected sensor data from government organizations during the same period to corroborate the results. To analyze categorical features of wind, temperature, rain etc., to gets threshold mapping information with an accuracy of 94%. We examined the dissemination of information through multiple methods of news, weather agencies, government agencies, organizations, and the public. We have also formulated the classification in the network to determine the decision employing tweeted words. In this research, we examined the overall importance of social data and the dissemination of physical data by categorizing the word cloud.

Other aspect of our work focuses on cooperation model by exploiting cooperation structure. The structure is accounted for possible interaction of individual users which are constituted in a network which interact to selective neighbor leading to natural reward and cooperation in the accordance of game theory model. One of the key things in the behavioral experiment were the dynamics in a social network. A social network if missing a dynamic, it often constitutes as biased network. This study provides an essential way to utilize diverse data sources to find cooperativeness in a network. This cooperativeness is a perpetual sourced network with a permanent link of neighbor bond. We proposed a cooperative algorithm, cooperative learning, and dynamic variable to predict the spread of an information which showed direct correlation of disease spread. The correlation varies from 45% to 81% which depends on the policy accumulation with an accuracy of 67%.

6.2 Future Work

In the current analysis, we utilized physical location-based datasets of Weather and Disease Spread and of the United States and social data set from Twitter to identify the measured changes in information dissemination and behavior. The social network behaves as a behavior function and this behavior is derived from the probabilistic change in actions of user nodes during the change in states in social network This approach to examine behaviors is focused by measuring the spread behavior from diverse data sources in a social network to adapt the action frequencies by employing an infectious disease framework to study social contagion.

However, the more data diversifies the cooperative problem becomes complex in nature and we found that multi node can be utilized to solve the complexity with work required in this area by expanding the objectives and inclusion of more data sources. This will be requiring more optimization so that the tuning of the system parameters can be done for higher performance.

The study can be extended to learn and create a virtual network from structural information of a network, even if certain real data is unavailable. This can be done by creating a model for policy determination for newly evolved networks since our framework utilizes scores to influence entire network for policy and reward determinations. We plan to utilize multiple category data for different weather situations and pollution by providing a framework to estimate different scenarios.

Currently, our algorithm works on a single environmental condition with one type of data stream at a time. We want to investigate further whether the trained model is transferrable for different streams rather than learning from the start. We would also want to implement Pareto optimization to handle the situation for stochastic outcomes.

References

[1] H. Srivastava and R. Sankar, “Information dissemination from social network for extreme weather scenario,” IEEE Transactions on Computational Social Systems, 7(2), pp.319-328, 2020.

[2] N. H. Goddard, J. D. Moore, C. A. Sutton, J. Webb, and H. Lovell, "Machine learning and multimedia content generation for energy demand reduction," in 2012 Sustainable Internet and ICT for Sustainability (SustainIT), 2012: IEEE, pp. 1-5.

[3] D. Zelenik and M. Bieliková, "Context inference using correlation in human behaviour," in 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization, 2012: IEEE, pp. 3-8.

[4] J. Chen, H. Chen, G. Zheng, J. Z. Pan, H. Wu, and N. Zhang, "Big smog meets web science: smog disaster analysis based on social media and device data on the web," in Proceedings of the 23rd international conference on world wide web, 2014: ACM, pp. 505-510.

[5] R. V. Kulkarni and G. K. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: A brief survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 2, pp. 262-267, 2011.

[6] Z. Guo, H. Wang, and L. Li, "Business applications on potential lightning prediction," in 2011 7th Asia-Pacific International Conference on Lightning, 2011: IEEE, pp. 546-550.

[7] W. D. Phillips and R. Sankar, "Improved transient weather reporting using people centric sensing," in 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC), 2013: IEEE, pp. 920-925.

[8] Nurwidyantoro, A. and Winarko, E., "Event detection in social media: A survey." In International Conference on ICT for Smart Society, pp. 1-5. IEEE, 2013.

[9] Madani, A., Boussaid, O. and Zegour, D.E., "What’s happening: a survey of tweets event detection." In Proc. Intl. Conf. on Communications, Computation, Networks and Technologies (INNOV), pp. 16-22. 2014.

[10] Winarko, E. and Roddick, J.F., "ARMADA–An algorithm for discovering richer relative temporal association rules from interval-based data." Data & Knowledge Engineering 63, no. 1 (2007): 76-90.

[11] Culotta, A., “Towards detecting influenza epidemics by analyzing Twitter messages,” in Proceedings of the First Workshop on Social Media Analytics, Washington D.C., District of Columbia, 2010, pp. 115–122.

[12] Bodnar, T. and Salathé, M., "Validating models for disease detection using twitter." In Proceedings of the 22nd International Conference on World Wide Web, pp. 699-702. 2013.

[13] Ritterman, J., Osborne, M. and Klein, E., "Using prediction markets and Twitter to predict a swine flu pandemic." in 1st international workshop on mining social media, vol. 9, pp. 9-17. 2009.

[14] Wakamiya, S., Kawai, Y. and Aramaki, E., , “Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study,” vol. 4, no. 3, pp. e65, 2018.

[15] Asgari-Chenaghlu, Meysam, Narjes Nikzad-Khasmakhi, and Shervin Minaee. "Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder." arXiv preprint arXiv:2009.03947 (2020).

[16] H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu and B. Liu, "Predicting Flu Trends using Twitter data," 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2011, pp. 702-707.

[17] Alvanaki, Foteini, Sebastian Michel, Krithi Ramamritham, and Gerhard Weikum. "See what's enBlogue: real-time emergent topic identification in social media." In Proceedings of the 15th International Conference on Extending Database Technology, pp. 336-347. 2012.

[18] Bontcheva, K. and Rout, D., "Making sense of social media streams through semantics: a survey." Semantic Web 5, no. 5 (2014): 373-403.

[19] Atefeh, F. and Khreich, W., "A survey of techniques for event detection in twitter." Computational Intelligence 31, no. 1 (2015): 132-164.

[20] Papadopoulos, S., Corney, D. and Aiello, L.M., "Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media." In Snow-dc@ www. 2014.

[21] Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. "Twitterstand: news in tweets." In Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems, pp. 42-51. 2009.

[22] Walther, M. and Kaisser, M., "Geo-spatial event detection in the twitter stream." In European conference on information retrieval, pp. 356-367. Springer, Berlin, Heidelberg, 2013.

[23] Phillips, W. D. and Sankar, R., "Improved transient weather reporting using people centric sensing", Proc. IEEE 10th Consum. Commun. Netw. Conf. (CCNC), pp. 920-925, Jan. 2013.

[24] Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y. and Vazirgiannis, M., "Degeneracy-based real-time sub-event detection in twitter stream." In Proceedings of the international AAAI conference on web and social media, vol. 9, no. 1, pp. 248-257. 2015.

[25] Guille, A. and Favre, C., "Mention-anomaly-based event detection and tracking in twitter." In 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), pp. 375-382. IEEE, 2014.

[26] Petrović, S., Osborne, M. and Lavrenko, V., "Using paraphrases for improving first story detection in news and Twitter." In Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 338-346. 2012.

[27] Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S. and Miller, R.C., "Twitinfo: aggregating and visualizing microblogs for event exploration." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 227-236. 2011.

[28] Popescu, A.M., Pennacchiotti, M. and Paranjpe, D., "Extracting events and event descriptions from twitter." In Proceedings of the 20th international conference companion on World wide web, pp. 105-106. 2011.

[29] Ishikawa, Shota, Yutaka Arakawa, Shigeaki Tagashira, and Akira Fukuda. "Hot topic detection in local areas using Twitter and Wikipedia." In ARCS 2012, pp. 1-5. IEEE, 2012.

[30] Nishida, Kyosuke, Takahide Hoshide, and Ko Fujimura. "Improving tweet stream classification by detecting changes in word probability." In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 971-980. 2012.

[31] Aiello, Luca Maria, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. "Sensing trending topics in Twitter." IEEE Transactions on multimedia 15, no. 6 (2013): 1268-1282.

[32] Petrović, Saša, Miles Osborne, and Victor Lavrenko. "Streaming first story detection with application to twitter." In Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, pp. 181-189. 2010.

[33] Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C. and Ounis, I., "Bieber no more: First story detection using twitter and wikipedia." In Sigir 2012 workshop on time-aware information access, pp. 16-76. 2012.

[34] Benhardus, J. and Kalita, J., "Streaming trend detection in twitter." International Journal of Web Based Communities 9, no. 1 (2013): 122-139.

[35] Cataldi, M., Di Caro, L. and Schifanella, C., "Emerging topic detection on twitter based on temporal and social terms evaluation." In Proceedings of the tenth international workshop on multimedia data mining, pp. 1-10. 2010.

[36] Lee, R. and Sumiya, K., "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection." In Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks, pp. 1-10. 2010.

[37] Mathioudakis, M. and Koudas, N., "Twittermonitor: trend detection over the twitter stream." In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 1155-1158. 2010.

[38] Allan, J. ed., Topic detection and tracking: event-based information organization. Vol. 12. Springer Science & Business Media, 2012.

[39] Aggarwal, C.C. and Subbian, K.,"Event detection in social streams." In Proceedings of the 2012 SIAM international conference on data mining, pp. 624-635. Society for Industrial and Applied Mathematics, 2012.

[40] Cordeiro, M., "Twitter event detection: combining wavelet analysis and topic inference summarization." In Doctoral symposium on informatics engineering, vol. 1, pp. 11-16. 2012.

[41] Petrović, S., Osborne, M. and Lavrenko, V.,"Using paraphrases for improving first story detection in news and Twitter." In Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 338-346. 2012.

[42] Weiler, A., Grossniklaus, M. and Scholl, M.H.,"Survey and experimental analysis of event detection techniques for twitter." The Computer Journal 60, no. 3 (2017): 329-346.

[43] D. Zelenik and M. Bieliková, "Context inference using correlation in human behaviour," in 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization, 2012: IEEE, pp. 3-8.

[44] R. V. Kulkarni and G. K. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: A brief survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 2, pp. 262-267, 2011.

[45] M. Okamoto and M. Kikuchi, "Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries," in Asia Information Retrieval Symposium, 2009: Springer, pp. 181-192.

[46] S. Takahashi, D. L. Sallach, and J. Rouchier, Advancing social simulation: The first world congress. Springer, 2007.

[47] Z. Guo, H. Wang, and L. Li, "Business applications on potential lightning prediction," in 2011 7th Asia-Pacific International Conference on Lightning, 2011: IEEE, pp. 546-550.

[48] C. Klüver, J. Klüver, and D. Zinkhan, "A self-enforcing neural network as decision support system for air traffic control based on probabilistic weather forecasts," in 2017 International Joint Conference on Neural Networks (IJCNN), 2017: IEEE, pp. 729-736.

[49] Ş. Kolozali, D. Puschmann, M. Bermudez-Edo, and P. Barnaghi, "On the effect of adaptive and nonadaptive analysis of time-series sensory data," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 1084-1098, 2016

[50] Weiler, Andreas, Michael Grossniklaus, and Marc H. Scholl. "Survey and experimental analysis of event detection techniques for twitter." *The Computer Journal* 60, no. 3 (2017): 329-346.

[51] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: privacy-aware people-centric sensing," in Proceedings of the 6th international conference on Mobile systems, applications, and services, 2008: ACM, pp. 211-224.

[52] S. R. Yerva, H. Jeung, and K. Aberer, "Cloud based social and sensor data fusion," in 2012 15th International Conference on Information Fusion, 2012: IEEE, pp. 2494-2501.

[53] N. Farajidavar, S. Kolozali, and P. Barnaghi, "Physical-cyber-social similarity analysis in smart cities," in 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), 2016: IEEE, pp. 484-489.

[54] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in Fifth international AAAI conference on weblogs and social media, 2011.

[55] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in 2012 Proceedings IEEE INFOCOM, 2012: IEEE, pp. 1701-1709.

[56] D. Ediger, S. Appling, E. Briscoe, R. McColl, and J. Poovey, "Real-time streaming intelligence: Integrating graph and nlp analytics," in 2014 IEEE High Performance Extreme Computing Conference (HPEC), 2014: IEEE, pp. 1-6.

[57] T. Giannetsos, T. Dimitriou, and N. R. Prasad, "People□centric sensing in assistive healthcare: Privacy challenges and directions," Security and Communication Networks, vol. 4, no. 11, pp. 1295-1307, 2011.

[58] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 4, pp. 1624-1639, 2011.

[59] T. Giannetsos, T. Dimitriou, and N. R. Prasad, "People□centric sensing in assistive healthcare: Privacy challenges and directions," Security and Communication Networks, vol. 4, no. 11, pp. 1295-1307, 2011.

[60] M. Okamoto and M. Kikuchi, "Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries," in Asia Information Retrieval Symposium, 2009: Springer, pp. 181-192.

[61] P. Svenmarck et al., "Message dissemination in social networks for support of information operations planning," TNO DEFENCE SECURITY AND SAFETY SOESTERBERG (NETHERLANDS), 2010.

[62] S. Navadia, P. Yadav, J. Thomas, and S. Shaikh, "Weather prediction: A novel approach for measuring and analyzing weather data," in 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2017: IEEE, pp. 414-417.

[63] D. Puiu et al., "Citypulse: Large scale data analytics framework for smart cities," IEEE Access, vol. 4, pp. 1086-1108, 2016.

[64] T. Data. "https://developer.twitter.com." https://developer.twitter.com (accessed 2018, 2017).N. C. F. E. Information. "NOAA." https://www.ncdc.noaa.gov/data-access (accessed).

[65] A. Gámbaro, E. Parente, A. Roascio, and L. Boinbaser, "Word association technique applied to cosmetic products–A case study," Journal of sensory studies, vol. 29, no. 2, pp. 103- 109, 2014.

[66] X. Zhao, J. Jiang, J. He, Y. Song, P. Achanauparp, Ee-P. Lim, X. Li., "Topical keyphrase extraction from twitter," in Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, 2011: Association for Computational Linguistics, pp. 379-388.

[67] M. Dash and H. Liu, "Feature selection for classification," Intelligent data analysis, vol.1, no. 1-4, pp. 131-156, 1997.

[68] J. Z. Kolter and M. A. Maloof, "Learning to detect and classify malicious executables in the wild," Journal of Machine Learning Research, vol. 7, no. Dec, pp. 2721-2744, 2006.

[69] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[70] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in Advances in neural information processing systems, 2007, pp. 153-160.

[71] J. Von Neumann and O. Morgenstern, Theory of games and economic behavior (commemorative edition). Princeton university press, 2007.

[72] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," IEEE Signal Processing Magazine, vol. 31, no. 5, pp. 32-43, 2014.

[73] M. Srivastava, T. Abdelzaher, and B. Szymanski, "Human-centric sens- ing," Philosophical Transactions of the Royal Society, A, vol. 370, no. 1958, pp. 176–197, 2012.

[74] D. Wang, H. Le, T. Abdelzaher, and L. Kaplan, "On truth discovery in social sensing: A maximum likelihood estimation approach," in Proc of IPSN, 2012.

[75] Akbari, Mohammad, Xia Hu, Nie Liqiang, and Tat-Seng Chua. "From tweets to wellness: Wellness event detection from twitter streams." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1. 2016.

[76] Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond trending topics: Real-world event identification on twitter." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, no. 1, pp. 438-441. 2011.

[77] Chakrabarti, Deepayan, and Kunal Punera. "Event summarization using tweets." In Proceedings of the International AAAI Conference on Web and social media, vol. 5, no. 1, pp. 66-73. 2011.

[78] Meladianos, Polykarpos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. "Degeneracy-based realtime sub-event detection in twitter stream." In Proceedings of the international AAAI conference on web and social media, vol. 9, no. 1, pp. 248-257. 2015.

[79] N. H. Goddard, J. D. Moore, C. A. Sutton, J. Webb, and H. Lovell, "Machine learning and multimedia content generation for energy demand reduction," in Proc. Sustain. Internet ICT Sustainability (SustainIT), Oct. 2012, pp. 1–5.

[80] D. Zelenik and M. Bielikova, "Context inference using correlation in human behaviour," in Proc. 7th Int. Workshop Semantic Social Media Adaptation Personalization, Dec. 2012, pp. 3– 8.

[81] R. V. Kulkarni and G. K. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: A brief survey," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 41, no. 2, pp. 262–267, Mar. 2011.

[82] C. Kluver, J. Kluver, and D. Zinkhan, "A self-enforcing neural network as decision support system for air traffic control based on probabilistic weather forecasts," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2017, pp. 729–736.

[83] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: Privacy-aware people-centric sensing," in Proc. 6th Int. Conf. Mobile Syst., Appl., Services (MobiSys). New York, NY, USA: ACM, 2008, pp. 211–224.

[84] D. Ediger, S. Appling, E. Briscoe, R. Mccoll, and J. Poovey, "Realtime streaming intelligence: Integrating graph and NLP analytics," in Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC), Sep. 2014, pp. 1–6.

[85] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layerwise training of deep networks," in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 153–160.

[86] A. Gámbaro, E. Parente, A. Roascio, and L. Boinbaser, "Word association technique applied to cosmetic products—A case study," J. Sensory Stud., vol. 29, no. 2, pp. 103–109, Apr. 2014

[87] Nikolentzos, Giannis, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. "Shortest-path graph kernels for document similarity." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1890-1900. 2017.

[88] Srivastava. H and Sankar. R, "Information Dissemination from Social Network for Extreme Weather Scenario," in IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 319-328, April 2020.

[89] Sakaki, T., Okazaki, M. and Matsuo, Y., “Earthquake shakes Twitter users: real-time event detection by social sensors,” in Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, 2010, pp. 851–860.

[90] Li, R., Lei, K.H., Khadiwala, R. and Chang, K.C.C.,"TEDAS: A Twitter-based Event Detection and Analysis System," 2012 IEEE 28th International Conference on Data Engineering, 2012, pp. 1273-1276, doi: 10.1109/ICDE.2012.125.

[91] Li, C., Sun, A. and Datta, A., "Twevent: segment-based event detection from tweets." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 155-164. 2012.

[92] Abel, F., Hauff, C., Houben, G.J., Stronkman, R. and Tao, K., “Twitcident: fighting fire with information from social web streams,” in Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 2012, pp. 305–308.

[93] Adam, N., Eledath, J., Mehrotra, S. and Venkatasubramanian, N., "Social media alert and response to threats to citizens (SMART-C)." In 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp. 181-189. IEEE, 2012.

[94] Terpstra, T., Stronkman, R., de Vries, A. and Paradies, G.L., "Towards a realtime Twitter analysis during crises for operational crisis management." In Iscram. 2012.

[95] Ritter, A., Etzioni, O. and Clark, S., "Open domain event extraction from twitter." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1104-1112. 2012.

[96] Bahir, E. and Peled, A., "Identifying and tracking major events using geo-social networks." Social science computer review 31, no. 4 (2013): 458-470.

[97] Martin, C., Corney, D. and Goker, A., “Mining newsworthy topics from social media." In Advances in social media analysis, pp. 21-43. Springer, Cham, 2015.

[98] Parikh, R. and Karlapalem, K., "Et: events from tweets." In Proceedings of the 22nd international conference on world wide web, pp. 613-620. 2013.

[99] Abdelhaq, H., Sengstock, C. and Gertz, M., "Eventweet: Online localized event detection from twitter." Proceedings of the VLDB Endowment 6, no. 12 (2013): 1326-1329.

[100] Weiler, A., Grossniklaus, M. and Scholl, M.H., "Event identification and tracking in social media streaming data." In EDBT/ICDT, pp. 282-287. 2014.

[101] Corney, D., Martin, C. and Göker, A., "Spot the ball: Detecting sports events on Twitter." In European Conference on Information Retrieval, pp. 449-454. Springer, Cham, 2014.

[102] Ifrim, G., Shi, B. and Brigadir, I., "Event detection in twitter using aggressive filtering and hierarchical tweet clustering." In Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014. ACM, 2014.

[103] Zhou, X. and Chen, L., "Event detection over twitter social media streams." The VLDB journal 23, no. 3 (2014): 381-400.

[104] Thapen, N., Simmie, D. and Hankin, C., "The early bird catches the term: combining twitter and news data for event detection and situational awareness." Journal of biomedical semantics 7, no. 1 (2016): 1-14.

[105] Monmousseau, P., Marzuoli, A., Feron, E. and Delahaye, D., "Impact of Covid-19 on passengers and airlines from passenger measurements: Managing customer satisfaction while putting the US Air Transportation System to sleep." Transportation Research Interdisciplinary Perspectives 7 (2020): 100179.

[106] Blei, D.M., Ng, A.Y. and Jordan, M.I., "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[107] Hoffman, M., Bach, F. and Blei, D., "Online learning for latent dirichlet allocation." advances in neural information processing systems 23 (2010).

[108] McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I. and McCullough, D., "On building a reusable twitter corpus." In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 1113-1114. 2012.

[109] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance," in IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 370-383, March 2007, doi: 10.1109/TKDE.2007.48.

[110] Xiao, L., Zheng, Z. and Peng, S., "Cross-Domain Relationship Prediction by Efficient Block Matrix Completion for Social Media Applications." Int. J. Perform. Eng. 16, no. 7 (2020): 1087-1094.

[111] Firoozeh, N., Nazarenko, A., Alizon, F. and Daille, B.,"Keyword extraction: Issues and methods." Natural Language Engineering 26, no. 3 (2020): 259-291.

[112] Zurbenko, I.G. and Smith, D., "Kolmogorov–Zurbenko filters in spatiotemporal analysis." Wiley Interdisciplinary Reviews: Computational Statistics 10, no. 1 (2018): e1419.

[113] Becker, Rolf. "Gender and Survey Participation: An Event History Analysis of the Gender Effects of Survey Participation in a Probability-based Multi-wave Panel Study with a Sequential Mixed-mode Design." methods, data, analyses 16, no. 1 (2022): 30.

[114] Khatoon, S., Asif, A., Hasan, M.M. and Alshamari, M., "Social Media-Based Intelligence for Disaster Response and Management in Smart Cities." In Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities, pp. 211-235. Springer, Cham, 2022.

[115] Bellatreche, L., Ordonez, C., Méry, D. and Golfarelli, M., "The central role of data repositories and data models in Data Science and Advanced Analytics." Future Generation Computer Systems 129 (2022): 13-17.

[116] Savic, N., Bovio, N., Gilbert, F., Paz, J. and Guseva Canu, I., "Procode: A Machine-Learning Tool to Support (Re-) coding of Free-Texts of Occupations and Industries." Annals of Work Exposures and Health 66, no. 1 (2022): 113-118.

[117] Jones, Karen Spärck. "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation (2004).

[118] Li, J., Maier, D., Tufte, K., Papadimos, V. and Tucker, P.A., "No pane, no gain: efficient evaluation of sliding-window aggregates over data streams." Acm Sigmod Record 34, no. 1 (2005): 39-44.

[119] Srivastava, H., Sheybani, E., & Sankar, R. (2022). Social Network Anomaly Detection for Optimized Decision Development. International Journal of Interdisciplinary Telecommunications and Networking (IJITN), 14(1), 1-8. http://doi.org/10.4018/IJITN.309697

[120] H. Srivastava and R. Sankar, " Cooperative Attention Based-learning Between Diverse Data Sources," in IEEE Transactions on Computational Social Systems (Review).

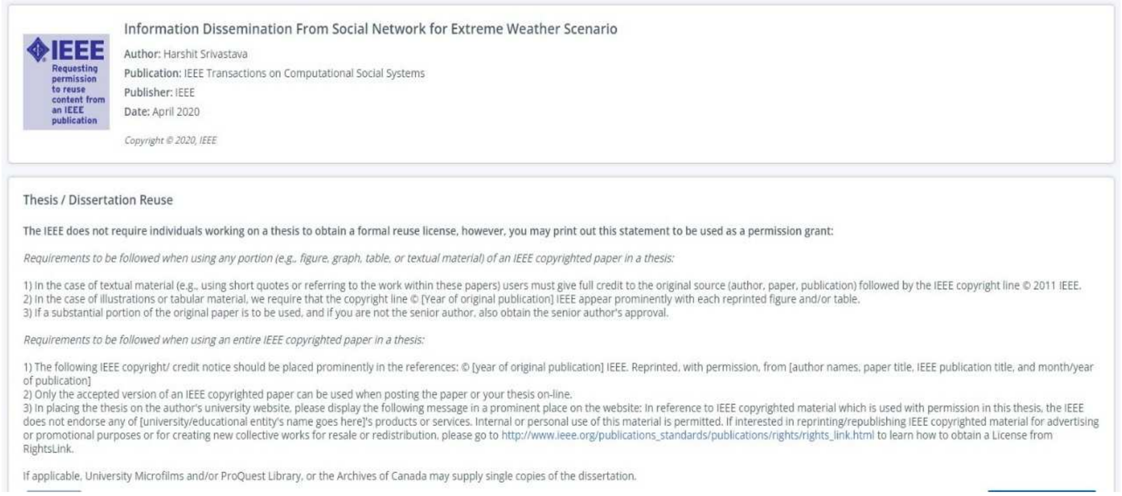
[121] H. Srivastava and R. Sankar, " A Detailed Survey on Social Sensors and Use in Detecting Events," in ACM Transaction of social computing (Review).

[122] Harshit Srivastava, In-Ho Ra, and Ravi Sankar. 2021. Cooperative Influence Learning. In The 9th International Conference on Smart Media and Applications (SMA 2020). Association for Computing Machinery, New York, NY, USA, 433–437. https://doi.org/10.1145/3426020.3426159

[123] H. Srivastava and R. Sankar, "Cooperation Model for Optimized Classification on Social Data," 2022 IEEE Symposium on Computers and Communications (ISCC), 2022, pp. 1-5, doi: 10.1109/ISCC55528.2022.9913047.

Appendix A: Copyright Permissions

The permission below is for the use of material in Chapter 4.


IEEE
Requesting permission to reuse content from an IEEE publication

Information Dissemination From Social Network for Extreme Weather Scenario

Author: Harshit Srivastava
Publication: IEEE Transactions on Computational Social Systems
Publisher: IEEE
Date: April 2020

Copyright © 2020, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

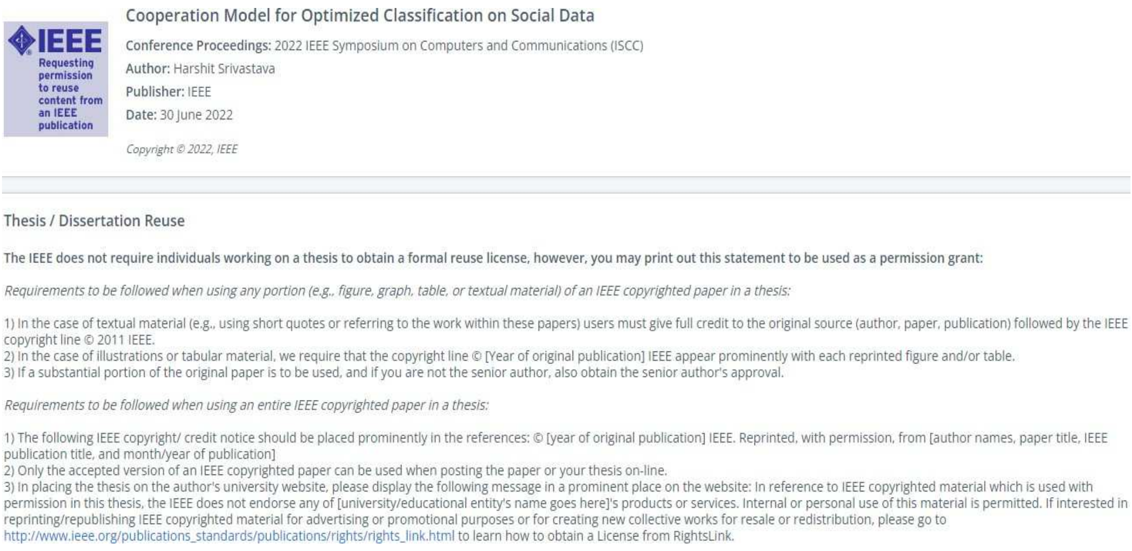
1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

The permission below is for the use of material in Chapter 5.


IEEE
Requesting permission to reuse content from an IEEE publication

Cooperation Model for Optimized Classification on Social Data

Conference Proceedings: 2022 IEEE Symposium on Computers and Communications (ISCC)
Author: Harshit Srivastava
Publisher: IEEE
Date: 30 June 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.