AUTOMATIC DETECTION OF CYBER SECURITY EVENTS FROM TURKISH TWITTER STREAM AND TURKISH NEWSPAPER DATA

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF INFORMATICS

OF

THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGÜR URAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF SCIENCE

IN

CYBER SECURITY

MAY 2019

Approval of the thesis:

**AUTOMATIC DETECTION OF CYBER SECURITY EVENTS FROM TURKISH TWITTER STREAM AND TURKISH NEWSPAPER DATA**

Submitted by **ÖZGÜR URAL** in partial fulfillment of the requirements for the degree of **Master of Science in Cyber Security Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin

Dean, **Graduate School of Informatics**

Prof. Dr. Aysu Betin Can

Head of Department, **Cyber Security**

Assist. Prof. Dr. Cengiz Acartürk

Supervisor, **Cognitive Science Dept., METU**

Assoc. Prof. Dr. XXX

Co-Supervisor, **Computer Engineering Dept., METU**

**Examining Committee Members:**

Prof. Dr. XXX (\*)

Computer Engineering Dept., METU

Prof. Dr. XXX (\*\*)

Medical Informatics Dept., METU

Prof. Dr. XXX

Cognitive Science Dept., METU

Prof. Dr. XXX

Computer Engineering Dept., Boğaziçi University

Prof. Dr. XXX

Computer Engineering Dept., Bilkent University

**Date: \_***Write your defense date!*

*\*Write the name of the head of the examining committee in the first row.*

*\*\*Write ten name of the supervisor in the second row.*

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Last name : ÖZGÜR URAL**

**Signature :**

# ABSTRACT

AUTOMATIC DETECTION OF CYBER SECURITY EVENTS FROM TURKISH TWITTER STREAM AND TURKISH NEWSPAPER DATA

Ural, Özgür

MSc., Department of Cyber Security

Supervisor: Assoc. Prof. Dr. Cengiz Acartürk

May 2019, XX pages

The Internet has different kinds of timely information. Every day, security experts scan the internet and face security events that affect people, institutions, and governments. An information security analyst regularly examines sources to stay up to date on security events require. This situation may lead to a heavy workload for the information analysts if they have not proper tools. For example, an information analyst may want to stay aware of cybersecurity incidents, such as a DDoS attack on a government agency website. The earlier they detect and understand the threats, the longer time remaining to alleviate the obstacle and to investigate the incident through forensics analyses methods. Therefore, information security analysts need to establish situational awareness of the security events and their likely effects. However, due to the large volume of information flow, it may be difficult for security analysts and researchers to detect and analyze security events appropriately. There are academic and commercial successful researches to solve this problem. Information security analysts actively use such technical solutions in their daily routine.

On the other hand, the internet community uses different languages to share information. For instance, information about security events, which effects citizens, institutions, and the government in Turkey, are shared on the internet, mainly in Turkish. The present thesis investigates the automatic detection of security incidents in Turkish by processing Twitter and news media. It proposes an automatic, Turkish specific software system that can detect cybersecurity events in real time.

Keywords: Cyber Security, Event Detection, Turkish, Twitter, Hurriyet Newspaper.

# ÖZ

TÜRKÇE TWİTTER AKIŞI VE TÜRKÇE GAZETE VERİLERİNDEN SİBER GÜVENLİK OLAYLARININ OTOMATİK TESPİT EDİLMESİ

Ural, Özgür

Yüksek Lisans, Siber Güvenlik Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Cengiz Acartürk

Mayıs 2019, XX sayfa

İnternet çeşitli konularda ve zamana bağlı bilgiler içerir. Güvenlik uzmanları her gün interneti taramakta ve insanları, kurumları ve hükümetleri etkileyen güvenlik olaylarıyla karşı karşıya kalmaktalar. Bir bilgi güvenliği analisti güvenlik olayları hakkında güncel kalabilmek için çeşitli kaynaklarını düzenli olarak takip eder. Bu durum, uygun araçlara sahip olmadıkları takdirde bilgi güvenliği analistleri için ağır bir iş yüküne yol açabilmektedir. Örneğin, bir bilgi analisti, bir devlet kurumu web sitesine yapılan DDoS saldırısı gibi siber güvenlik olaylarının farkında olmak isteyebilir. Tehditleri ne kadar erken saptarsa ve olası tehditleri anlarlarsa, problemleri hafifletmek ve adli olarak soruşturmak için o kadar uzun süreye sahip olurlar. Bu nedenle, mevcut güvenlik olayları ve olası etkileri hakkında durum bilgisine sahip olmaları gerekmektedir. Ancak, çok sayıda siber güvenlik olayı gerçekleşmesi nedeniyle, güvenlik analistlerinin ve araştırmacıların bu bilgi akışını yeterli şekilde ele alması zor olabilmektedir. Bu sorunu çözmek için akademik ve ticari başarılı araştırmalar bulunmaktadır. Bilgi güvenliği analistleri bu tür teknik çözümleri günlük rutinlerinde aktif olarak kullanmaktadırlar.

Öte yandan, internette bilgi paylaşmak amacıyla farklı diller kullanılmaktadır. Örneğin, Türkiye'deki vatandaşları, kurumları ve hükümeti etkileyen güvenlik olayları hakkındaki bilgiler internette çoğunlukla Türkçe olarak paylaşılmaktadır. Bu tez, Türkçe kullanılan Twitter akışlarını ve haber medyalarını işleyerek, güvenlik olaylarının otomatik olarak algılanmasını araştırmaktadır. Siber güvenlik olaylarını gerçek zamanlı olarak algılayabilen otomatik, Türkçe'ye özgü bir yazılım sistemi önermektedir.

Anahtar Sözcükler: Siber Güvenlik, Olay Tespiti, Türkçe, Twitter, Hürriyet Gazetesi.

# DEDICATION

To My Family

# ACKNOWLEDGMENTS

First of all, I would like to express …..

Besides my supervisor, I would like to thank …..

I would also like to thank all of colleagues from …..

To my wife, …..

# TABLE OF CONTENTS

[ABSTRACT iv](#_Toc9467474)

[DEDICATION vii](#_Toc9467476)

[ACKNOWLEDGMENTS viii](#_Toc9467477)

[TABLE OF CONTENTS ix](#_Toc9467478)

[LIST OF TABLES xii](#_Toc9467479)

[LIST OF FIGURES xiii](#_Toc9467480)

[LIST OF ABBREVIATIONS xv](#_Toc9467481)

[1 INTRODUCTION 16](#_Toc9467482)

[1.1 Motivation 16](#_Toc9467483)

[1.2 Research Question and Objectives 19](#_Toc9467484)

[1.3 Use Cases 20](#_Toc9467485)

[1.4 Outline 20](#_Toc9467486)

[2 BACKGROUND INFORMATION 22](#_Toc9467487)

[2.1 Routine Tasks of an Information Security Analyst 22](#_Toc9467488)

[2.2 What is Natural Language Processing? 23](#_Toc9467489)

[2.3 What is Text Mining? 24](#_Toc9467490)

[2.4 Twitter Social Network 25](#_Toc9467491)

[2.4.1 What is Twitter? 25](#_Toc9467492)

[2.4.2 Twitter API 26](#_Toc9467493)

[2.5 Hürriyet Turkish Newspaper as a Data Source 27](#_Toc9467494)

[2.5.1 What is Hurriyet Newspaper? 27](#_Toc9467495)

[2.5.2 Hurriyet Newspaper API 27](#_Toc9467496)

[2.5.3 What is OData 27](#_Toc9467497)

[2.5.4 Hürriyet API - OData Usage 27](#_Toc9467498)

[2.6 Python Programming Language 29](#_Toc9467499)

[2.7 Istanbul Technical University NLP API 29](#_Toc9467500)

[2.8 SQLite Database 30](#_Toc9467501)

[2.9 Other Technologies Used in the Thesis Project 30](#_Toc9467502)

[3 LITERATURE REVIEW 31](#_Toc9467503)

[3.1 Weakly Supervised Extraction of Computer Security Events from Twitter 31](#_Toc9467504)

[3.2 Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an Example 33](#_Toc9467505)

[3.3 DDoS Event Forecasting Using Twitter Data 33](#_Toc9467506)

[3.4 Prediction of Drive-by Download Attacks on Twitter 34](#_Toc9467507)

[3.5 SONAR: Automatic Detection of Cyber Security Events Over the Twitter Stream 34](#_Toc9467508)

[3.6 Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media 36](#_Toc9467509)

[4 SYSTEM ARCHITECTURE AND DESIGN 38](#_Toc9467510)

[4.1 Approach 38](#_Toc9467511)

[4.2 Data Collection 42](#_Toc9467512)

[4.3 Data Preprocessing 42](#_Toc9467513)

[4.4 Cybersecurity Related Event Detection 43](#_Toc9467514)

[5 IMPLEMENTATION 44](#_Toc9467515)

[5.1 Multi Process Architecture 44](#_Toc9467516)

[5.1.1 Twitter API Stream to Database Process 44](#_Toc9467517)

[5.1.2 Hurriyet API Stream to Database Process 45](#_Toc9467518)

[5.1.3 ITU NLP API Normalization Process 45](#_Toc9467519)

[5.1.4 Security Events Web Portal Process 45](#_Toc9467520)

[5.2 Microservice Architecture 46](#_Toc9467521)

[5.3 Database Architecture of the System 46](#_Toc9467522)

[5.3.1 Source Column of the Database 46](#_Toc9467523)

[5.3.2 Date Column of the Database 47](#_Toc9467524)

[5.3.3 UserName Column of the Database 47](#_Toc9467525)

[5.3.4 Title Column of the Database 47](#_Toc9467526)

[5.3.5 Text Column of the Database 47](#_Toc9467527)

[5.3.6 Status Column of the Database 48](#_Toc9467528)

[5.4 User Interface of the System 48](#_Toc9467529)

[5.5 Implementation Details of the Code Files 48](#_Toc9467530)

[5.5.1 manager.py File 48](#_Toc9467531)

[5.5.2 config.py File 49](#_Toc9467532)

[5.5.3 hurriyetApiToDb.py File 58](#_Toc9467533)

[5.5.4 ituNlpPipeline.py File 59](#_Toc9467534)

[5.5.5 securityEventsWebPortal.py File 59](#_Toc9467535)

[5.5.6 sqliteOperations.py File 60](#_Toc9467536)

[5.5.7 twitterStreamToDb.py File 60](#_Toc9467537)

[5.5.8 userInterface.html File 61](#_Toc9467538)

[6 RESULTS 62](#_Toc9467539)

[6.1 Successful Cybersecurity Event Detection Samples 63](#_Toc9467540)

[6.1.1 WhatsApp Spyware Attack 63](#_Toc9467541)

[6.1.2 Vulnerabilities in Remote Patient Tracking System Applications 64](#_Toc9467542)

[6.1.3 Other Successful Detection Examples 64](#_Toc9467543)

[6.2 Unsuccessful Cybersecurity Event Detection Samples 69](#_Toc9467544)

[6.2.1 Sample False Positive Cybersecurity Event Detection 69](#_Toc9467545)

[6.2.2 Sample not Useful Cybersecurity Event Detection 69](#_Toc9467546)

[7 CONCLUSION AND FUTURE WORK 70](#_Toc9467547)

[7.1 Conclusion 70](#_Toc9467548)

[7.2 Future Work 71](#_Toc9467549)

[REFERENCES 72](#_Toc9467550)

[APPENDICES 75](#_Toc9467551)

[APPENDIX A 75](#_Toc9467552)

[APPENDIX B 76](#_Toc9467553)

# LIST OF TABLES

[Table 1: Example high-confidence events extracted using the system published within this paper. 31](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467555)

[Table 2: Example of high-weight features. Context words other than nouns and verbs are replaced with their part of speech tags for better generalization. 32](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467556)

[Table 3: Seed Instances for DDoS Attacks. 32](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467557)

[Table 4: Tweet Examples with Attack Targets. 33](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467558)

[Table 5: Sample Table after the Analyze 40](#_Toc9467559)

[Table 6: Sample Information Sharing Plan Table 40](#_Toc9467560)

[Table 7: Sample Detected Events Table 40](#_Toc9467561)

# LIST OF FIGURES

[Figure 1: Tweets in Turkish After the Turktrust Vulnerability Announcement on 3 January 2013 17](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467600)

[Figure 2: Hürriyet Newspaper News after the Turktrust SSL Vulnerability is Detected 18](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467601)

[Figure 3: Research results of IBM Security Lab about Cyber Security Analysts 23](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467602)

[Figure 4: A Simple diagram to explain what NLP does. 24](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467603)

[Figure 5 Sample Turkish Tweets Related with a Security Incident 26](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467604)

[Figure 6: Architecture of the Keyword Finder Component. 35](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467605)

[Figure 7: Technical Overview of Sonar. 35](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467606)

[Figure 8: A Schematic Overview of Cybersecurity Event Detection System from The Publication. 36](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467607)

[Figure 9: Streamgraph Showing Normalized Volume of Tweets (September 2015 through October 2016) Tagged with Data Breach (red), DDoS Activity (grey) and Account Hijacking (blue) Types of Cybersecurity Events. 36](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc9467608)

[Figure 10: Normalizer Sample from ITU NLP API 39](#_Toc9467609)

[Figure 11: The General Overview of the System 41](#_Toc9467610)

[Figure 12: Sample Stored Twitter Data in the Database File 44](#_Toc9467611)

[Figure 13: Sample Stored Hurriyet Newspaper Data in the Database File 45](#_Toc9467612)

[Figure 14 User Interface of the Cybersecurity Detection Software 62](#_Toc9467613)

[Figure 15 WhatsApp Spyware Attack Detection 63](#_Toc9467614)

[Figure 16 STM Warns about Remote Patient Tracking System Applications 64](#_Toc9467615)

[Figure 17 Supply Chain Attack Targets ASUS Computers Through Backdoored Update 64](#_Toc9467616)

[Figure 18 MuddyWater Attack 65](#_Toc9467617)

[Figure 19 MegaCortex Ransomware 65](#_Toc9467618)

[Figure 20 Using Popular TV Shows to Spread Malware 65](#_Toc9467619)

[Figure 21 RTM Banking Trojan 65](#_Toc9467620)

[Figure 22 Mother's Day Cyberattack Alerts 66](#_Toc9467621)

[Figure 23 Valentine’s Day Phishing Attacks 66](#_Toc9467622)

[Figure 24 Pirate Matryoshka Virus 66](#_Toc9467623)

[Figure 25 Cyber Attacks Using Smart TVs 66](#_Toc9467624)

[Figure 26 Smart Home Cyberattacks 66](#_Toc9467625)

[Figure 27 Valletta Bank Cyberattack 67](#_Toc9467626)

[Figure 28 Business Email Compromise Attack 67](#_Toc9467627)

[Figure 29 Cyberattacks Which Use Netflix Brand 67](#_Toc9467628)

[Figure 30 Cyberattacks Targeting Job Seekers 67](#_Toc9467629)

[Figure 31 Phishing Attacks Targeting Social Media Users 68](#_Toc9467630)

[Figure 32 Zombie Cookies 68](#_Toc9467631)

[Figure 33 Millions of Email Addresses Infiltrated the Internet 68](#_Toc9467632)

[Figure 34 The USA and China Cyberwar Started Tweet 68](#_Toc9467633)

[Figure 35 Angela Merkel and Hundreds of German Politicians Hacked 69](#_Toc9467634)

[Figure 36 Sample False Positive Cybersecurity Event Detection 69](#_Toc9467635)

[Figure 37 Sample, not Useful Cybersecurity Event Detection 69](#_Toc9467636)

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **DDOS** | Distributed Denial of Service |
| **DOS** | Denial of Service |
| **REST** | Representational State Transfer |
| **API** | Application Programming Interface |
| **HTTP** | Hyper-Text Transfer Protocol |
| **OData** | Open Data Protocol |
| **JSON** | JavaScript Object Notation |
| **IDE** | Integrated Development Environment |

**CHAPTER 1**C

# INTRODUCTION

## Motivation

Security awareness tools help security analysts to protect an institution's sensitive and mission-critical data from being stolen, damaged, or compromised by attackers. The duration between the disclosure of a new vulnerability and the moment when the security analyst becomes aware of it is crucial for taking appropriate countermeasures in a timely manner.

On 3 January 2013, Google Inc. announced a security vulnerability which allowed spoofing using fraudulent digital certificates issued by Turktrust Inc.(Langley, 2013) Other companies such as Microsoft and Mozilla which may be affected by this vulnerability followed Google and announced the vulnerability, shared their affected software and devices and suggested actions. After these announcements, Twitter and Turkish newspapers showed a quick reaction. As shown in Figure 1, Twitter users shared the news on the same day immediately after the announcement on 3 January 2013. Since Turktrust certificates were a significant part of certificate use market in Turkey, numerous Tweets circulated in Turkish related to the vulnerability.

According to Statistia, Twitter has 321 million monthly active users worldwide(Twitter, 2019). Turkey is the fifth country in the list of leading countries with nearly 9 million active users, as of January 2019.(“Countries with most Twitter users 2019 | Statistic,” 2019) Twitter users can tweet in any languages they select. Although there are no statistics about the use of Turkish by Twitter users from Turkey, it is very likely that most of the Turkish Twitter users share their tweets in their native language.

A review of the literature and recent state of technology reveals that most of the research conducted on security event detection focus on English. As of our knowledge, research is lacking on real-time security event detection in Turkish language streams.



Figure 1: Tweets in Turkish After the Turktrust Vulnerability Announcement on 3 January 2013

According to wearesocial.com's 2019 Global Digital Report, Turkey has 82,4 million population. Internet usage penetration in Turkey is %72 with 59.36 million internet users, and active social media penetration in Turkey is %63 with 52 million people.(“Global Digital Report 2019 | Free Download | We Are Social UK,” n.d.) With emerging internet adoption in Turkey, there are much timely information shared in Turkish. Event detection systems which developed for English texts are not useful for Turkish texts mining. Therefore, in order to use Turkish texts at detection of cyber security events, we should add the Turkish language-specific methods and algorithms to the event detection systems and automate such systems.

Social media is not the only option to extract information as such. A security analyst has a wide range of sources available such as the specialized press, blogs, forums, news agencies, newspapers, and so on to gather cyber threat information. However, their initial source of information for detecting such security events is usually social networks. After the emergence of a trending event, users increasingly share posts about it on social media. For instance, a DDOS attack to a service or a website is usually recognized and reported by social media users first, and they share the information on online platforms, by posting tweets such as "X website is unreachable."

An alternative way to extract information about security events is newspapers. After the Turktrust SSL vulnerability in 2013, the newspapers also share that information fast. Figure 2, shows an excerpt from Hurriyet newspaper related to the vulnerability.(“Yanlış sertifika Google’dan döndü - Teknoloji Haberleri,” n.d.)



Figure 2: Hürriyet Newspaper News after the Turktrust SSL Vulnerability is Detected

An autonomous system which can use various data sources for security event detection has the potential to be beneficial for a security analyst. We designed and developed a software system capable of detecting and monitoring cybersecurity-related events over the Twitter Stream in Turkish. It can technically process millions of documents per day and detect security events. To gain more accurate results, we added the Hürriyet Turkish newspaper stream to analyze and detect security events. The software solution’s infrastructure supports adding new data resources, thus providing flexibility. For example, we can add LinkedIn, Facebook, Eksisozluk website streams to gain more accurate results.

## Research Question and Objectives

The objectives of this thesis are the followings.

* Current cybersecurity event detection tools are developed for extracting meaningful data from English texts. Cybersecurity event detection rate will be low when they are adapted to Turkish as they are due to the linguistic characteristics of Turkish. What can be done to make the accuracy of a tool developed for Turkish as high accuracy as the tools developed for English in terms of cybersecurity event detection? It is our research question. This thesis also answers this research question by proposing a methodology and its implementation.
* We research and review “state of the art” studies and software systems in real-time event detection, as reported in the literature review chapter.
* We investigate potential data sources to determine the most suitable ones to use it for real-time event detection with Turkish-text.
* We investigate methodologies and API's related to NLP(Natural Language Processing) to use it for normalization[[1]](#footnote-1) of Turkish texts.
* We design and develop a software system for real-time cybersecurity event detection using Turkish texts.
* We design the system as a framework to make useable it for further researches. Turkish datasets are used in various research areas like text classification, author detection, automatic question answering. However, finding datasets in Turkish is difficult since there are limited accessible datasets online. By means of this thesis software framework, researchers will be able to access datasets in Turkish. Moreover, they will be able to select and modify their queries by changing keyword vectors, thus changing the concent of information to be extracted from online sources.
* We validate the proposed approach using several detected events already shared in Turkish-in online platforms.

## Use Cases

Cybersecurity is an emerging topic in Turkey, just like the rest of the world. There exists limited research about automated security event detection systems recently. However, these studies focus on data mining in the English language. Although the available cybersecurity event detection systems can be beneficial for detecting global level events, such systems cannot be used with other languages like Turkish, because NLP data mining is language specific. Security analysts who work in Turkey, or just interested in local security events in Turkey can use data in Turkish to detect such events. By means of automatic event detection systems, a security analyst establishes situation awareness in cyberspace and take countermeasures against new threats. For example, a security analyst who is working for a Turkish institution may use local websites APIs like Ekşisözlük API e-Devlet API or libraries/frameworks developed for focused Turkish people. If these API's, libraries or frameworks have vulnerabilities, and someone discovers them, they are probably discussed and announced within social media like Twitter in Turkish. It is likely that Turkish newspapers publish it as breaking news too. To detect such events automatically, the software system must listen to Turkish data sources and process the text in Turkish. Our research aims at meeting these requirements by proposing a software system and framework for security event detection.

## Outline

* Chapter 2 presents background information for the thesis. A review of available methods, terminology, and standard terms are presented in this chapter.
* In Chapter 3, we present relevant literature.
* In Chapter 4, we introduce the software system in terms of its architectural and design perspectives.
* In Chapter 5, we present the software system in terms of its implementation and evaluation perspectives.
* In Chapter 6, we discuss thesis results.
* Finally, in Chapter 7, we present the conclusion and propose possible future works.

**CHAPTER 2**

# BACKGROUND INFORMATION

## Routine Tasks of an Information Security Analyst

According to the careerexplorer website(“What does an information security analyst do? ‐ CareerExplorer,” n.d.), an information security analysts the primary responsibility is to take countermeasures for protecting organizational-level, mission-critical and sensitive information, as well as being prepared for cyber attacks. To be prepared for a cyber attack, they use various tools and systems. One of their responsibility is to analyze data and to recommend changes to higher-ups. However, security analysts are not authorized to implement changes. Their main job is to keep cyber attacks out.

In practice, a security analyst spends approximately one hour per a working day to get caught up on the latest security news through bulletins, forums, news, social networks an so on to identify new threats. They further spend two to three hours by repeated investigation of potential security incidents using online resources. They spend the rest of their daily time with manually copying and pasting information from disparate and siloed tools to correlate data. They generally face with ten to twenty challenges daily such as monitoring security access, analyzing security breaches to identify the root cause, verifying the security of third-party vendors and collaborating with them to meet security requirements and so on.(“What is a Security Analyst? Responsibilities, Qualifications, and More | Digital Guardian,” n.d.) Their investigation time gives cyber attackers advantages if it is long enough, and it is challenging for a security analyst to keep up with threats.

In Figure 3(Borrett, 2017) below, statistics are shown about the security analysis, which motivates why security analysts need automated systems.



Figure 3: Research results of IBM Security Lab about Cyber Security Analysts

It is not sustainable without automated systems. To make it sustainable, automated NLP analysis tools and Text mining methods need to be used. In the following sections 2.2 and 2.3, we will explain the concepts of NLP and Text Mining.

## What is Natural Language Processing?

NLP is “the ability of machines to understand and interpret human language the way it is written or spoken”(*International Journals of Management, IT et Engineering IJMIE.*, n.d.). In Figure 4(“Overview of Artificial Intelligence &amp; Role of NLP in Big Data - XenonStack Blog,” n.d.), can be seen as a simple explanation of What NLP does. In this thesis, we used a few NLP techniques and Istanbul Techincal University’s NLP API(“ITU Turkish Natural Language Processing Web Interface,” n.d.) for normalization of the texts.



Figure 4: A Simple diagram to explain what NLP does.

In order to develop automated systems, NLP is one of the actively used concepts in text mining. According to the data-flair website, “The role of NLP in text mining is to deliver the system in the information extraction phase as an input.”(“Text Mining in Data Mining - Concepts, Process & Applications - DataFlair,” n.d.) In the following section 2.3, the concept of text mining introduced.

## What is Text Mining?

The Oxford English Dictionary defines text mining as “the process or practice of examining large collections of written resources to generate new information, typically using specialized computer software.”(Stephanie Prato, 2013) Text mining consists of a broad variety of methods and technologies. In this thesis, we used Keyword-based technologies and statistics technologies. According to expertsystem website, Keyword-based technologies definition is “The input is based on a selection of keywords in text that are filtered as a series of character strings, not words nor concepts.”(“Text mining vs data mining: discover the differences -,” n.d.) Also, statistics technologies definition is “Refers to systems based on machine learning. Statistics technologies leverage a training set of documents used as a model to manage and categorize text.”(“Text mining vs data mining: discover the differences -,” n.d.)

In the sections above, NLP and text mining concepts were presented. The text required for text mining for cybersecurity event detection purposes is gathered from online platforms, as presented in Chapter 1. The following section provides information about Twitter, a popular online platform.

## Twitter Social Network

### What is Twitter?

Twitter is an online social networking service, which was created in October 2006 by Jack Dorsey, Even Williams, and Biz Stone. People use Twitter for various purposes(Huberman, Romero, & Wu, 2008).

First of all, One of its usage examples is as a social messaging service. Users can interact with the other users, communicate with their friends and family, and share details of their lives. Secondly, users can use it as a microblogging service for sharing details of a person’s life. Thirdly, users can use Twitter as a marketing tool for public relations. Many celebrities and politicians use Twitter for interacting with their audience. Lastly, Twitter is an information platform on which users can get news via broadcasting agents’ or journalists’ accounts fast and efficiently. Moreover, there are Twitter bots created by developers for a precise function like Bitcoin ticker bot will tweet every hour the price of Bitcoin in Turkish Lira.

According to the first quantitative study on Twitter “What is Twitter, a Social Network or a News Media?” which is published in 2010(Kwak, Lee, Park, & Moon, n.d.), Twitter is more an information sharing network than a social network. They found that result while working on Twitter follower graph. They decided that because of the low rate of reciprocated ties. People tend to use Twitter as a news feed by following multiple online news media, but other Twitter users will only follow “real” users.

Twitter users can post a short message called tweet, which is limited to 280 characters, or retweet another user tweet. Photos, videos, or URLs can be added to the tweets. Users can follow other accounts and creates their networks. They can mention each other or reply to each other within their tweets. To identify what the tweet is about, users use word preceded by a hash sign (#). Twitter uses these hashtags to define trending topics, both locally and globally. Users use the trending topic lists to identify favorite subjects at that time on Twitter.

In default settings, all Twitter accounts are public. Users can interact with each other like replying other user's tweets, sending a private direct message, and so on.



Figure 5 Sample Turkish Tweets Related with a Security Incident

### Twitter API

The Twitter API is a set of URLs. The URLs cant take parameters and let users access Twitter features like finding tweets which contain a set of specific words and so on.

Twitter provides several APIs to get tweets:

Twitter’s REST API allows users to get tweets or search terms, which includes specific parameters. It is useful for analytics on historical data because this API does not give users live data. Moreover, this API cannot retrieve tweets older than seven days.

Twitter’s Streaming API gives users access to live data on Twitter and keeps sending it until asked it to stop. Developers can access only 1% sample of all the tweets. Its typical usage is when doing analytics over live campaigns on Twitter, and so on.

In the present thesis, we use two different data source, and one of them is Twitter. We gather the unstructured data as Twitter text(tweets) and analyze them to detect cybersecurity events. Its details will be presented in Chapter 4 and Chapter 5. Our second data source, Hurriyet Newspaper, is introduced in the following Section.

## Hürriyet Turkish Newspaper as a Data Source

### What is Hurriyet Newspaper?

Hürriyet is one of the major Turkish newspapers, founded in 1948. As of January 2018, it had the highest circulation of any newspaper in Turkey at around 319,000.(“Tiraj | MedyaTava - Yazmadıysa Doğru Değildir,” n.d.)

### Hurriyet Newspaper API

Hürriyet API is an interface which enables the usage of Hürriyet data programmatically in web, mobile, or desktop applications. It is a free service. With Hürriyet API, developers can reach news, columns, writers, photo galleries, and pages. Hürriyet API has a RESTful-based, resource-oriented architecture. Developers can access Hürriyet newspaper data via standard HTTP requests. The resultant set of results is in JSON format. Requests via the API are limited to 5 per second and 500 per hour to prevent abuse.(“Hurriyet Developers API v1.0 Docs — Hürriyet Public API,” n.d.)

In the present thesis, we use two different data source, and Hurriyet Newspaper is one of them. We fetch the unstructured data as news text and analyze them to detect cybersecurity events. The process details will be presented in Chapter 4 and Chapter 5.

### What is OData

OData is a REST-based data source using the HTTP protocol is a global protocol for querying services. With OData standards, developers do not waste much time on basic standards such as to request and response headers, status codes, HTTP methods (GET, POST, and so on), and query options. Developers can only create RESTful APIs by building business logic.

Consuming OData services is easy. Client - interpretable can quickly render OData metadata. Therefore, developers can quickly integrate it into robust and expandable client applications.

### Hürriyet API - OData Usage

The OData structure has a unique query structure. Below are some of the most basic query keywords and their functionality briefly outlined:

$ select: Limits the columns/properties in the response set from the query. Example use;

* https://api.hurriyet.com.tr/v1/articles?$select=Title

To limit relational properties such as Files, RelatedNews; it is necessary to use $ select filter with $ expand. Example use;

* https://api.hurriyet.com.tr/v1/articles?$select=Files&$expand=Files

$ filter: By adding a filter to the query, the answer set can be limited. Example use;

* https://api.hurriyet.com.tr/v1/articles?$filter=Path eq '/gundem/'

Users can also use these keywords together to increase the number of filters in the result set and make it easier to reach the desired result set.

Using OData protocol on Hürriyet API service, these can be queried and used in applications.

* Articles in the system
* Columns in the system
* In-system photo galleries
* The pages in the system and the pages assigned to the pages
* Folders in the system
* Writers

Requests via the API are restricted to block abuse. These limits are five requests per second and 500 requests per hour.

In order to develop such a system, a software implementation is required. We used Python Programming Language to implement the system. Python Programming Language and its selection criteria are introduced in the following section.

## Python Programming Language

“Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.”(Python, 2017) It is a multi-paradigm programming language and supports so many paradigms like object-oriented programming, structured programming, functional programming, and so on. It has enough frameworks and API to work on cognitive science, text mining, NLP like areas. It is fast enough, and learning it is also fast. Most big companies use Python in data mining projects. To illustrate, according to a 2014 article in Fast Company magazine, Facebook chooses to use Python for data analysis because it was already used so widely in other parts of the company.(“Businesses Can Now Use The Same Stats Language As Universities, Thanks,” n.d.) In this thesis, we use Python version 3.6.6.

In the following section, a Turkish language specific NLP API, which we use in the present thesis is explained.

## Istanbul Technical University NLP API

Turkish NLP Tools and APIs developed by the Natural Language Processing group at Istanbul Technical University. The program is available at “tools.nlp.itu.edu.tr” website. (“ITU Turkish Natural Language Processing Web Interface,” n.d.) The API is free to use for academic purposes. To be able to use the API, we need access token and an account for the token. In order to get them, we sent an email to briefly explain who we are, why we need to access the API and our affiliation. Our application seems okay for them. Therefore, they give us credentials.

The platform operates as a Software as a Service and provides the researchers and the students the state of the art NLP tools in many layers: preprocessing, morphology, syntax, and entity recognition. (Eryiğit, 2015). It is a web API; developers can access it with an HTTP request and can use GET or post method.

The ITU NLP API components for stand-alone usage are the followings;

• Tokenizer

• Deasciifier

• Vowelizer

• Spelling Corrector

• Normalizer

• isTurkish

• Morphological Analyzer

• Morphological Disambiguator

• Named Entity Recognizer

Twitter API can also filter Turkish Tweets, and Hürriyet is a Turkish newspaper. Therefore, we do not need an “isTurkish” component of the API for the thesis. Currently, we only use the “Normalizer” component of the ITU NLP API.

After fetching the data, during the processing and presentation steps, we need to store information in a database. In the following section, we explain which database technology we used to store the data.

## SQLite Database

According to SQLite.org website, SQLite is an in-process library that implements a serverless, self-contained, zero-configuration, transactional SQL database engine. Using both commercial and private is free. SQLite is the most widely deployed database in the world, including high-profile projects. (Sqlite.org, 2013) It is an embedded database engine. Unlike most other SQL databases, SQLite reads and writes directly to ordinary disk files. SQLite does not have a separate server process. In the thesis project, we do not need the server side. Therefore, we choose SQLite to use in the thesis project.

## Other Technologies Used in the Thesis Project

We used Visual Studio Enterprise 2017 as IDE. It is handy, especially for debugging the code. Moreover, we used JSON as a data-interchange format. We use gif for version service with GitHub web-based hosting service. Our repository on GitHub is currently private, but we are planning to make it public as an opensource project when we finish the thesis.

**CHAPTER 3**

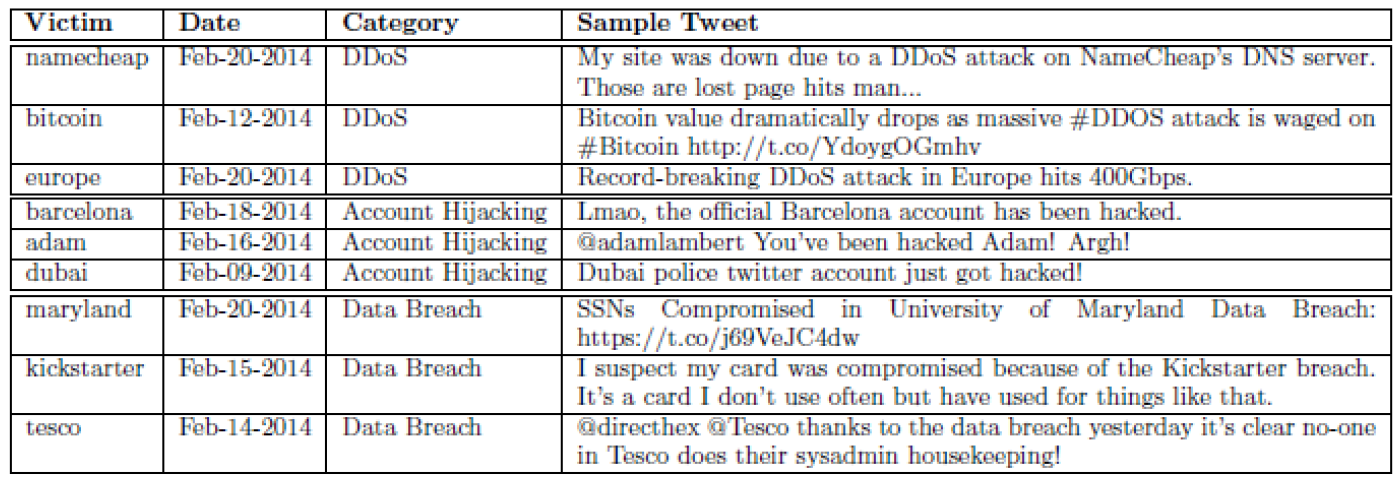
# LITERATURE REVIEW

In this section, we share the results of the literature review. We introduce the most relevant researches with our research and explain how they give shape to our research. Most of the following researches focus event detection and try to answer how can we obtain valuable information from streaming data.

## Weakly Supervised Extraction of Computer Security Events from Twitter

Research on identifying victims affected by attacks in these categories as output, using the Twitter data and adding categories to the user without being dependent on fixed categories.(Ritter, Wright, Casey, & Mitchell, 2016)

Table 1: Example high-confidence events extracted using the system published within this paper.



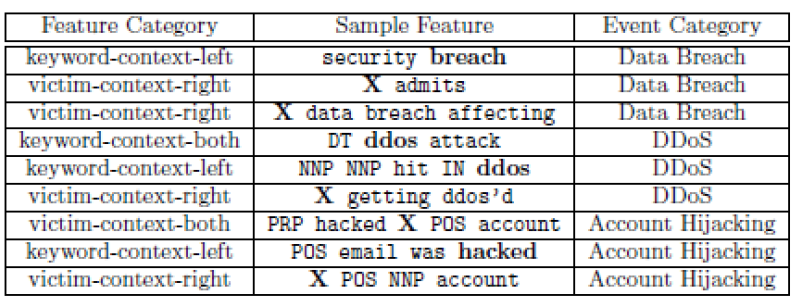
They determine candidate events as in Table 2.

Table 2: Example of high-weight features. Context words other than nouns and verbs are replaced with their part of speech tags for better generalization.

Then they are trying to find the victim, institution, or program affected by these events.

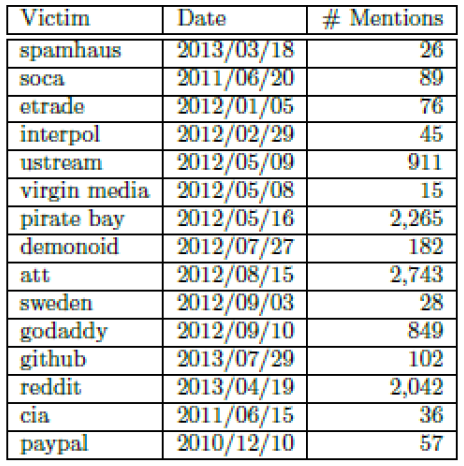


Table 3: Seed Instances for DDoS Attacks.

In this research, they focus on cybersecurity events detection using only Twitter. On the other hand, we use both Twitter and Hurriyet Newspaper to detect cybersecurity events in the present thesis. Moreover, they choose to use English texts as a data resource, while we use Turkish texts as the data resource. Furthermore, they programmatically detect victims. On the other hand, we use predefined vector sets to detect victims in our research.

## Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an Example

That paper(Aslan, Sağlam, & Li, 2018) is a joint publication of a Ph.D. student from Yıldırım Beyazıt University with one of the professors from his university and a professor at the University of Kent Canterbury form the UK. Using machine learning techniques, they investigated to find a method of whether social media accounts related to cybersecurity.

To prepare their dataset to use in their research, they develop a crawler with Twitter API using Python programming language. We also use Twitter crawler with Python programming language.

## DDoS Event Forecasting Using Twitter Data

It is a publication(Wang & Zhang, 2017) to estimate the DDoS attacks that have not yet taken place by processing Twitter data.

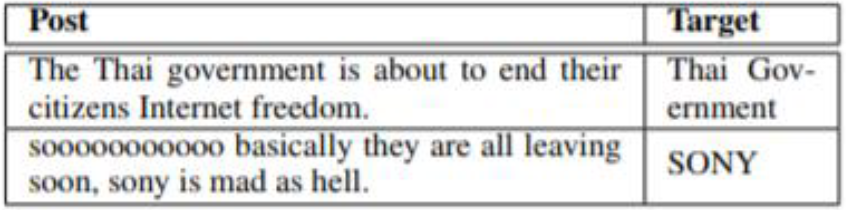


Table 4: Tweet Examples with Attack Targets.

They tried to obtain this information using six popular supervised classification models. To illustrate, one of the models which they used is the “negative term count.”. Neg-Term-count is the baseline sentiment-based model. They count the negative words from tweets each day, forecasting an attack if the number of negative words is more significant than a threshold α, which is the average number of negative words on training data.

Their research helps us to recognize a future work field which can be added to our research. In the future, we can try to detect cybersecurity events that have not yet taken place by processing streaming data using Turkish.

## Prediction of Drive-by Download Attacks on Twitter

Some cyber attackers use the URL abbreviation method to show malicious websites as if a harmless website and share them on twitter as an abbreviated URL. Some Twitter users may believe in this deception and click on such website abbreviations, and these links can harm the users. They have explored what we can do to prevent such malicious websites from being clicked like a safe website due to this kind of abbreviation. They try various methods such as detecting malicious software infection from the increase in the use of CPU or RAM with using Honeypot. (Javed, Burnap, & Rana, 2019)

Our thesis research may be useful in informing security experts from current cybersecurity events. The security experts may also want to inform such malicious URL’s. Therefore, we may also try to add such functionality to our research as future work. However, we will try to detect such an attack using Turkish Tweets.

## SONAR: Automatic Detection of Cyber Security Events Over the Twitter Stream

They developed a self-learning framework called Sonar.(Petersen, 2017) Sonar can automatically capture events related to cybersecurity by processing twitter data. Developers give the system some keywords to follow. The system can find other keywords to followed related to cybersecurity with the help of previously given keywords.

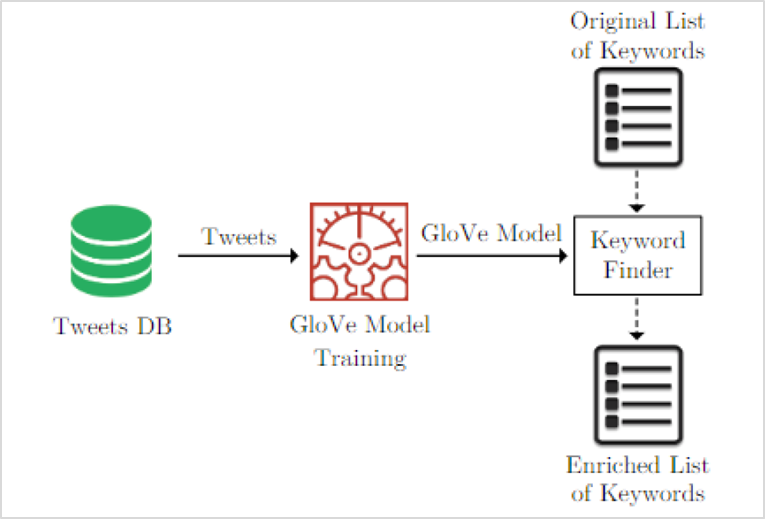


Figure 6: Architecture of the Keyword Finder Component.

They have also benefited from many big data technologies to do their researches.

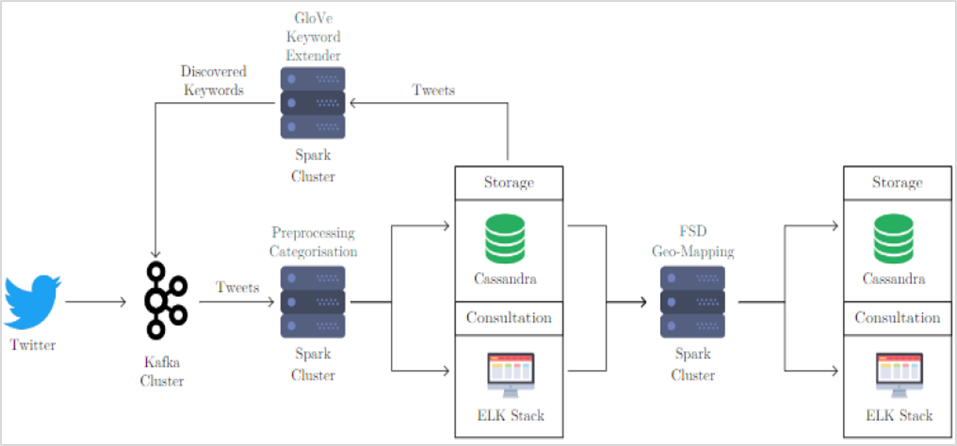


Figure 7: Technical Overview of Sonar.

In means of the architectural design of our system, we use this research in our present thesis.

## Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media

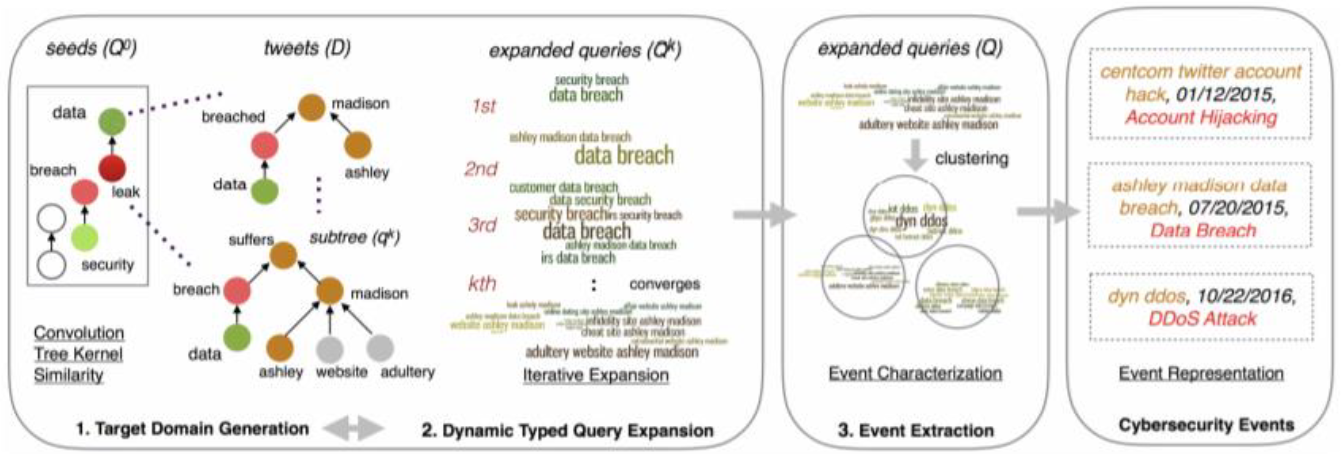
It is another study on detecting cybersecurity attacks by processing Twitter data. They acknowledge that their work is similar to that of previous studies, but they claim to have more successful results. (Khandpur et al., 2017)

Figure 8: A Schematic Overview of Cybersecurity Event Detection System from The Publication.

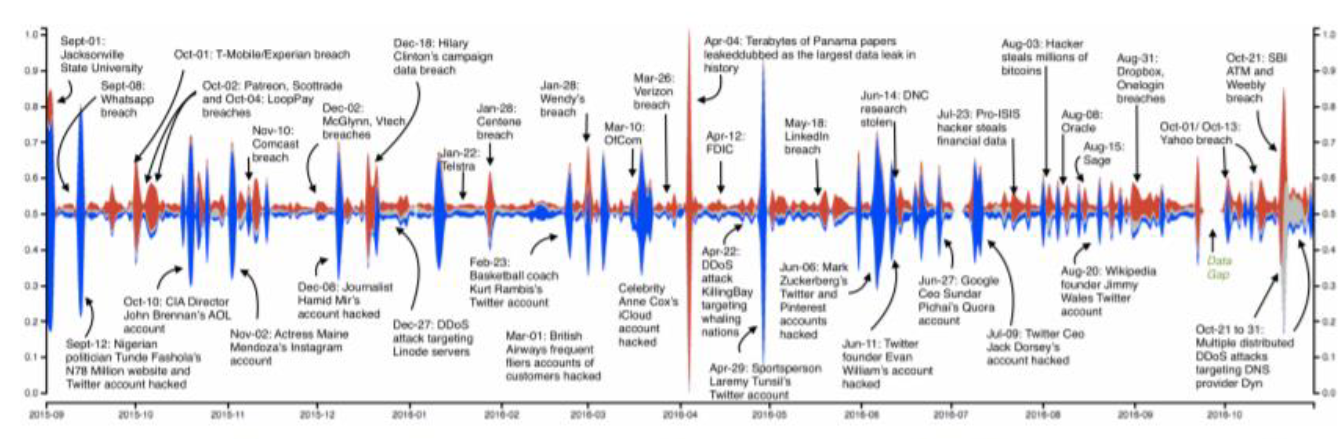


Figure 9: Streamgraph Showing Normalized Volume of Tweets (September 2015 through October 2016) Tagged with Data Breach (red), DDoS Activity (grey) and Account Hijacking (blue) Types of Cybersecurity Events.

This research is one of the “state-of-art project” in the cyber attack detection field. We also use this research to detect the boundary points of our research. They use Tweets in English to detect cyber attacks. On the other hand, we focus on both Tweets and newspaper data in Turkish. Moreover, while they research detecting cyber attacks, we research detecting cyber events in the present thesis.

**CHAPTER 4**

# SYSTEM ARCHITECTURE AND DESIGN

In this chapter, we explain the software system’s architecture and design.

## Approach

Our software system has a configuration file. The configuration file includes constant values such as Twitter API and ITU NLP API constants, logger constants, string vectors for named entity recognition, and so on. The software system is developed as generic as possible for the researchers can use them as a framework with just changing the configuration file.

In order to establish a Twitter stream connection, the software uses configuration file values, which are statically defined in the configuration file. We use a cybersecurity-related Turkish keyword vector to gather useful Twitter stream for our research. Moreover, we use the language filter feature of the Twitter API in order to fetch only Turkish Tweets.

To establish the Hurriyet Newspaper stream connection, the software also uses configuration file values in its codes. We use the cybersecurity-related Turkish keyword vector to gather useful newspaper data.

The mentioned keyword vector includes terms related to cybersecurity such as Turkish synonyms of "hacked, DDoS". Turkish synonyms of "hacked, DDoS" are "heklendi, erişilemiyor". Unfortunately, there is no Turkish cybersecurity terms dictionary. Therefore, we research Turkish cybersecurity terms to create a cybersecurity-related keywords vector.

The architecture of the software system is implemented considering new data sources may be wanted to add. Before writing the fetched data to the database, both fetched data of Hurriyet Newspaper and Twitter are formatted to a suitable form for writing database.

After writing them to the database, the texts in the database are sent ITU NLP API to normalize them. In Figure 10(Eryiǧit & Torunoǧlu-Selamet, 2017) a sample Turkish text normalization pipeline is shown.



Figure 10: Normalizer Sample from ITU NLP API

After the normalization step, we move forward to Named Entity Recognition[[2]](#footnote-2) step of our pipeline. In this state, we use the predefined string vector, which currently includes institution names, government organization name, and country names. These strings represent the potential victims of security events.

After that step, the software counts the number of mentions of the potential victims with searching the predefined string vector elements in the normalized texts which are stored in the database.

With this analysis, we will get a table like as shown in the following. The period of the table is one day. Every new day, the number of mentions per victim is set and start with zero.

Table 5: Sample Table after the Analyze



We add daily thresholds. (i.e., 10 per day) If the number of mentions is more than the thresholds value, we will share this information within a user interface like the following Figure.

Table 6: Sample Information Sharing Plan Table



Permanently, the software checks the database and analyze new texts for detecting new cybersecurity events. If one of the possible victim’s numbers of mentions in the cybersecurity-related database texts exceeds the threshold limit per day, the software system adds them to the table too like the following figure.

Table 7: Sample Detected Events Table



We will show this information in a local HTML file. In the future, we can show them on a web page. Security Analysts can see the detected security events from there.



Figure 11: The General Overview of the System

## Data Collection

Both Hürriyet API and Twitter API need seed keywords to query them. In order to collect Turkish stream data, we need Turkish cybersecurity terms. However, we cannot find a Turkish cybersecurity terms dictionary. Therefore, we research the Turkish cybersecurity terms and gather them as a list to use them in the query. However, we face a problem at that step. More than half cybersecurity-related terms have no Turkish synonym. Even in Turkish tweets and Turkish newspaper texts, English expressions of the cybersecurity-related terms may be used. Therefore, we decided to add both English and Turkish version of the cybersecurity-related terms to our keyword list and use them in our query. The resulting stream contains one or more elements of the keyword list per each Tweet or news. We can get up to %1 of the Twitter stream, which is approximately one million Tweets per day due to Twitter API limitation. Moreover, we can 12000 request per day in Hurriyet Newspaper API. Therefore, the keyword list is essential to get relevant data in the result streams.

## Data Preprocessing

Before writing the streaming data to our database, we need to format to texts. Firstly, we should select the needed keys from JSON streams of Twitter API and Hürriyet API. For example, Hurriyet API requests return related news in a JSON which has “Title of the News” key. The key can be useful for representing the detected event. On the other hand, there are unrelated or unuseful data in the JSON too, so we filter them and do not write in our database. We filter the Twitter API stream’s JSON keys too and select the useful and relevant keys too.

In our database, we have a ‘Status’ column. When we first write the texts to our database, we set the text’s status with ‘0’. ‘0’ means that the text is not processed yet, and it is raw data. We sent the raw data to ITU NLP API to normalize it. After the normalization step, we update the text with normalized text and update the Status column of the row which has the text with “1”. After the row is processed to detect cybersecurity events, the Status column is set with “2”. “2” means that the data processed before and there is nothing to do with that row of the table.

## Cybersecurity Related Event Detection

From the previous steps of the software system, we get the possible cybersecurity-related texts from different sources. Then preprocess them and store them in our database. In order to detect the events and find the possible victim of those events, we prepared a named entity vector. This vector includes possible victims which we want to track. Currently, this list includes institution names, government organization names, and country names. The vector can be updated from changing the configuration file to change tracked entities.

**CHAPTER 5**

# IMPLEMENTATION

* 1. Multi Process Architecture

We use multi-processed system architecture in the implementation of the project. There are four processes as described in the subchapters below.

* + 1. Twitter API Stream to Database Process

This process continually gathers Twitter API stream. Then preprocess the data and write them to the database.

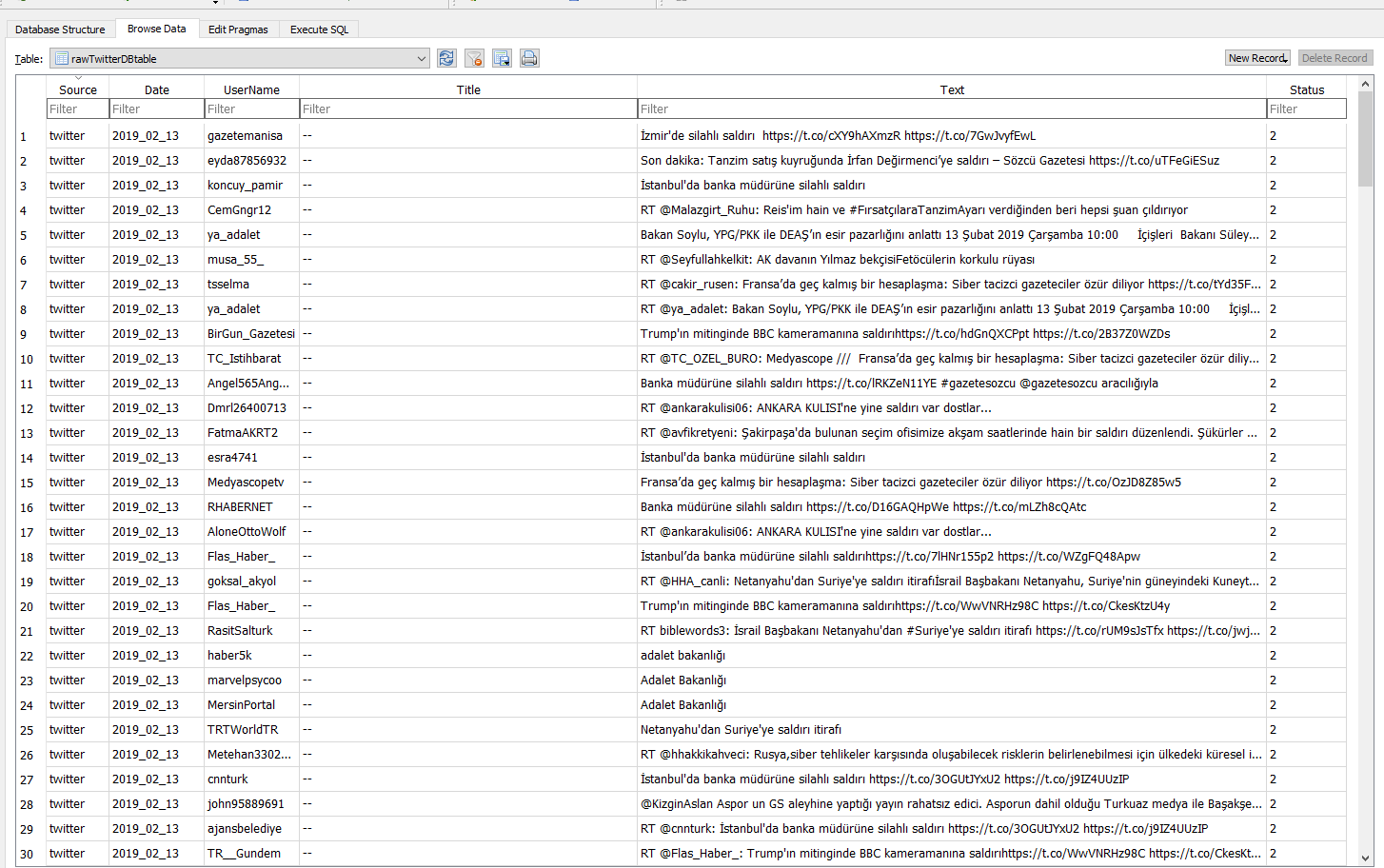


Figure 12: Sample Stored Twitter Data in the Database File

* + 1. Hurriyet API Stream to Database Process

This process continually gathers Hurriyet API stream. Then preprocess the data and write them to the database.

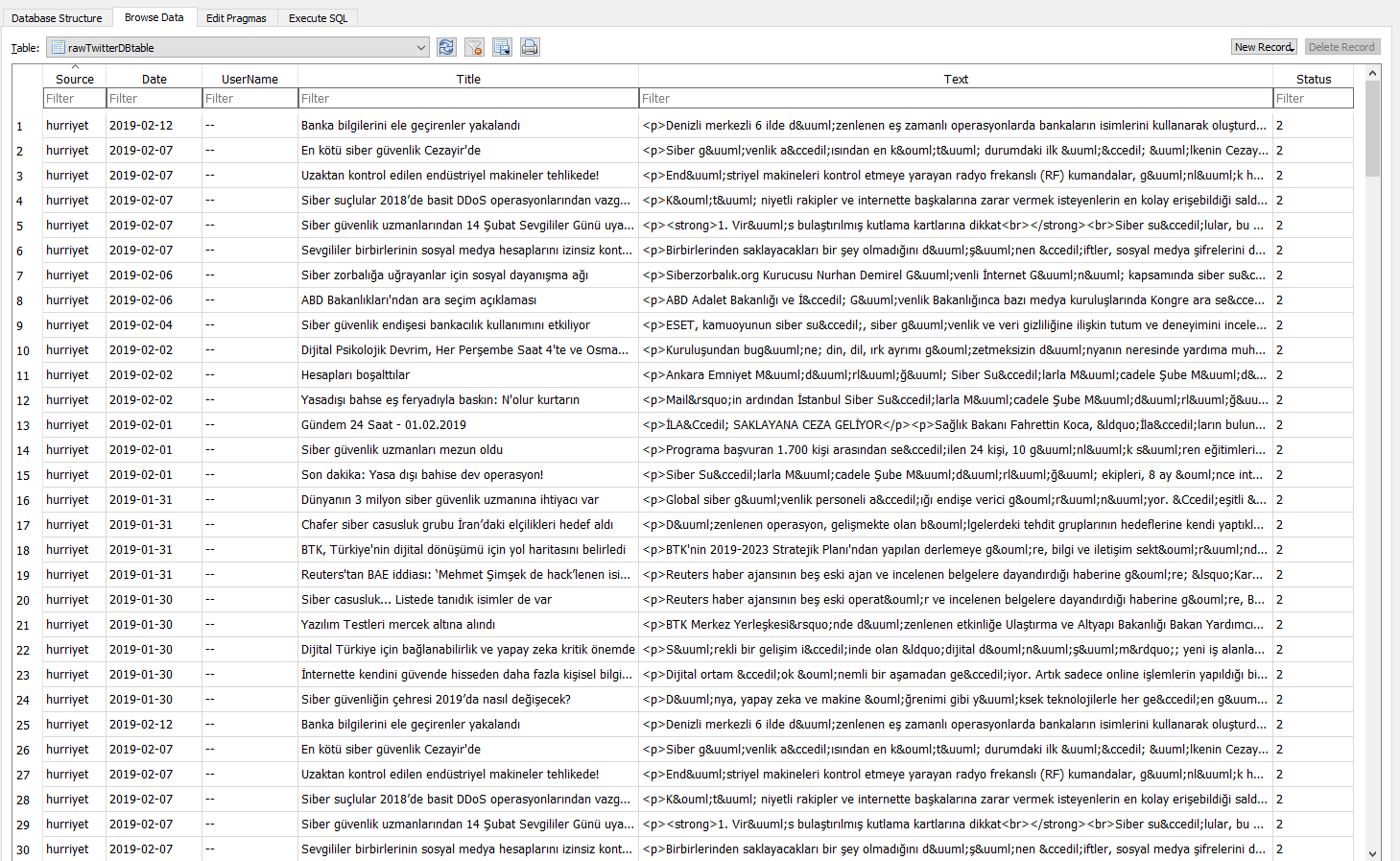


Figure 13: Sample Stored Hurriyet Newspaper Data in the Database File

* + 1. ITU NLP API Normalization Process

This process continually checks the database. If the process can find columns with status 0, then sent the columns to ITU NLP API servers to normalize them. After the normalization, the process writes back the texts to the database and update their status row with “1”.

* + 1. Security Events Web Portal Process

This process continually checks the database to find columns with status row set with “1”. If it can find, it processes them to add the HTML page which security analysts can monitor the events from that page.

* 1. Microservice Architecture

Microservices are small, and independent services focus on doing a task at a time and ability to work together. Because the project has the potential to grow, we design it with following microservice architecture. With this design, our software became resilient. Failure in one service does not impact the other services of our project. For example, assume that ITU NLP API service stops to work for a while and does not respond to our project’s requests. Due to the microservice architecture of our software, the other services can continue to work even if our software has monolithic or bulky service errors in one service. Hurriyet API can still gather the streaming data, preprocess them, and write them to the database; Twitter API can still gather the streaming data, preprocess them, and write them to the database and so on.

Moreover, it has scalability. For example, if our database technology becomes insufficient for our software, we can easily change the database technology with a more suitable one.

Furthermore, our software has less dependency and easy to modify its code and test them. Our software can easily understand by other developers since the processes represent the small piece of functionality. It is vital because our software solution will be an open source project and will be used by other developers and researchers. Lastly, this architecture method gives us the freedom to choose technology. We can choose the best-suited technology for each of functionalities.

## Database Architecture of the System

There are six columns in our database. Their properties are explained in the subsections below.

### Source Column of the Database

In our current design, this column must be filled. The column must be “twitter” if the source of the row is the Twitter API stream. The column must be “Hurriyet” if the source of the row is Hurriyet newspaper API stream. If new sources added to our system in the future, new unique texts could be set to define the source of the raw data.

### Date Column of the Database

This column must be filled. The period of the time is one day for our system. It means that our software system counts how many entities exist in texts per each day. For example, assume that Middle East Technical University is hacked, and people share tweets; newspapers share news about that hacking event. Let us say the first day after the hacking event; our software system can detect 100 tweets/news about that event. Let us say the second day after the hacking event; our software system can detect 50 tweets/news about that event. Due to both of the first day and the second detections are detected in separate days, our HTML portal which shows the detected evens shows this detection information with two separate detection because the period of our software solution is one day.

### UserName Column of the Database

It is an optional column of our database file because users share tweets on Twitter, and the users have a username. However, there is no user in Hurriyet newspaper data.

We store this value to control each Twitter user can affect the system with only one tweet per day. For example, if a Twitter user shares one thousand Tweet about an event in a day, our system allows only one of the user’s Tweet to write in our database for each day.

### Title Column of the Database

It is an optional column of our database file because Tweets do not have a title while news has one.

We show a representative Tweet or news text in the HTML portal to present information to security analysts. However, News texts can be very long to represent. Instead of the full text of a news, we only represent the title of the news for more clear representation

### Text Column of the Database

This column must be filled because both newspapers and Twitter stream data has text. For Twitter data, this column is filled with Tweets. On the other hand, for Hurriyet newspaper data, this column is filled with news texts.

### Status Column of the Database

This column represents the instant status of that row data. The meaning of each status number is explained in the table below.

|  |  |
| --- | --- |
| Status | Meaning of the Status |
| 0 | Text in that row is not processed yet, and it is raw data. |
| 1 | Text in that row was sent to ITU NLP API to normalize it, and the text of that row was updated with normalized text. |
| 2 | Text in that row was processed before, and there is no work remains to do on that row of the table. |

* 1. User Interface of the System

It is a simple dynamically generated HTML page which will be used by security analysts as a portal page of the system. A process continuously checks the database per minute to detect new data and use them to show the new cybersecurity events in this user interface.

## Implementation Details of the Code Files

### manager.py File

It is the code file which starts the whole software system. It is written with python. “multiprocessing” library is imported for multi-process usage. This manager defines four processes like below and starts them.

* p1 = Process(target=startTwitterStreamToDb)
* p2 = Process(target=startHurriyetApiToDb)
* p3 = Process(target=startItuNlpApi)
* p4 = Process(target=securityEventsWebPortalStart)

The processes always work in parallel.

### config.py File

It is the configuration file of the system. Static constants and configuration items are stored in there.

|  |
| --- |
| Twitter API Related Constants Stored in the Config File |
| CONSUMER\_KEY |
| CONSUMER\_SECRET |
| ACCESS\_TOKEN |
| ACCESS\_TOKEN ACESS\_SECRET |

|  |
| --- |
| ITU NLP API Related Constants Stored in the Config File |
| CONSUMER\_NAME |
| CONSUMER\_PASSWORD |
| ITU\_NLP\_API\_TOKEN |
| PIPELINE\_ENCODING |
| DEFAULT\_TOOL |

The cybersecurity keyword list which used as a keyword during fetching Turkish texts is also stored in “config.py”. The list is as below.

* keywords\_query = 'siber, hacklendi, fidye yazılımı, zafiyet, ddos, erişilemiyor, zararlı yazılım, malware, virüs, casus yazılım, fidye yazılımı, oltalama, phising, güvenik duvarı, firewall, hacker, sistem açığı, cyberattack, cyberwarfare, hacking, kimlik avı, spear-phishing'

The list which will be used for named entity recognition step is as below. It is also stored in “config.py”.

|  |  |  |
| --- | --- | --- |
|  | |  |
| **Named Entity Vectors in Turkish** | | |
| **Country Names** | **Institution and Government Organization Names in Turkey** | |
| "Türkiye", | | "Adalet Bakanlığı", |
| "ABD", | | "Adıyaman Üniversitesi", |
| "Amerika Birleşik Devletleri", | | "Adli Sicil ve İstatistik Genel Müdürlüğü", |
| "Afganistan", | | "Adli Tıp Kurumu Başkanlığı", |
| "Almanya", | | "Afet ve Acil Durum Yönetimi Başkanlığı", |
| "Andora", | | "Ağrı İbrahim Çeçen Üniversitesi", |
| "Angola", | | "Ahmet Yesevi Üniversitesi", |
| "Anguilla", | | "Aile ve Sosyal Politikalar Bakanlığı", |
| "Antarktika", | | "Aile ve Toplum Hizmetleri Genel Müdürlüğü", |
| "Barbuda", | | "Akaryakıt İkmal ve Nato Pol Tesisleri İşletme", |
| "Arjantin", | | "ANT", |
| "Arnavutluk", | | "Amelebirliği", |
| "Aruba", | | "Anadolu Ajansı Genel Müdürlüğü", |
| "Avrupa Birliği", | | "Anadolu Üniversitesi", |
| "Avustralya", | | "Anayasa Mahkemesi Başkanlığı", |
| "Avusturya", | | "Ankara Ticaret Odası", |
| "Azerbaycan", | | "Antalya Bilim Üniversitesi", |
| "Bahamalar", | | "Artvin Çoruh Üniversitesi", |
| "Bahreyn", | | "Asker Alma Dairesi Başkanlığı", |
| "Bangladeş", | | "Askerî Yargıtay Başkanlığı", |
| "Barbados", | | "Askerî Yüksek İdare Mahkemesi", |
| "Belçika", | | "Ataşehir Adıgüzel Meslek Yüksekokulu", |
| "Benin", | | "Atatürk Araştırma Merkezi", |
| "Beyaz Rusya", | | "Atatürk Kültür, Dil ve Tarih Yüksek Kurumu Başkanlığı", |
| "Bermuda", | | "Atatürk Kültür Merkezi", |
| "Bhutan", | | "Atatürk Orman Çiftliği", |
| "Birleşik Arap Emirlikleri", | | "AB Bakanlığı", |
| "Birleşik Krallık", | | "Avrupa Birliği Eğitim ve Gençlik Programları Merkezi Başkanlığı", |
| "Bolivya", | | "Avrupa Birliği ve Dış İlişkiler Genel Müdürlüğü", |
| "Bosna Hersek", | | "Balıkçılık ve Su ürünleri Genel Müdürlüğü", |
| "Botsvana", | | "Bandırma Onyedi Eylül üniversitesi", |
| "Bouvet", | | "Bankacılık Düzenleme ve Denetleme Kurumu", |
| "Brezilya", | | "Basın İlan Kurumu", |
| "Brunei", | | "Basın Yayın ve Enformasyon Genel Müdürlüğü", |
| "Bulgaristan", | | "Başbakanlık", |
| "Burkina", | | "Başbakanlık Güvenlik İşleri Genel Müdürlüğü", |
| "Burundi", | | "Başbakanlık İdareyi Geliştirme Başkanlığı", |
| "Cape Verde", | | "Batman Üniversitesi", |
| "Cebelitarık", | | "Bilgi Teknolojileri ve İletişim Kurumu", |
| "Cezayir", | | "Bilim, Sanayi ve Teknoloji Bakanlığı", |
| "Cibuti", | | "Bingöl Üniversitesi", |
| "Çad", | | "Bitkisel Üretim Genel Müdürlüğü", |
| "Çek Cumhuriyeti", | | "Bitlis Eren Üniversitesi", |
| "Çin", | | "Boru Hatları İle Petrol Taşıma A.Ş.", |
| "Danimarka", | | "Bursa Teknik Üniversitesi", |
| "Dominik Cumhuriyeti", | | "Bütçe ve Mali Kontrol Genel Müdürlüğü", |
| "Salvador", | | "Ceza İşleri Genel Müdürlüğü", |
| "Endonezya", | | "Ceza ve Tevkifevleri Genel Müdürlüğü", |
| "Ermenistan", | | "Coğrafi Bilgi Sistemleri Genel Müdürlüğü", |
| "Estonya", | | "Cumhurbaşkanlığı", |
| "Etiyopya", | | "Çalışma Genel Müdürlüğü", |
| "Fas", | | "Çalışma ve Sosyal Güvenlik Bakanlığı", |
| "Fiji", | | "Çalışma ve Sosyal Güvenlik Eğitim ve Araştırma Merkezi", |
| "Fildişi Sahilleri", | | "ÇASGEM", |
| "Filipinler", | | "Çanakkale Onsekiz Mart Üniversitesi", |
| "Filistin", | | "Çay İşletmeleri Genel Müdürlüğü", |
| "Finlandiya", | | "Çevresel Etki Değerlendirmesi, İzin ve Denetim Genel Müdürlüğü", |
| "Fransa", | | "Çevre ve Şehircilik Bakanlığı", |
| "Gabon", | | "Çevre Yönetimi Genel Müdürlüğü", |
| "Gambia", | | "Çocuk Hizmetleri Genel Müdürlüğü", |
| "Gana", | | "Çölleşme ve Erozyonla Mücadele Genel Müdürlüğü", |
| "Gine-Bissau", | | "Danıştay", |
| "Granada", | | "Dernekler Dairesi Başkanlığı", |
| "Grönland", | | "Devlet Arşivleri Genel Müdürlüğü", |
| "Guadeloupe", | | "Devlet Hava Meydanları İşletmesi Genel Müdürlüğü", |
| "Guam", | | "Devlet Malzeme Ofisi Genel Müdürlüğü", |
| "Guatemala", | | "Devlet Opera ve Balesi Genel Müdürlüğü", |
| "Guernsey", | | "Devlet Personel Başkanlığı", |
| "Guyana", | | "Devlet Su İşleri Genel Müdürlüğü", |
| "Güney Afrika", | | "Devlet Tiyatroları Genel Müdürlüğü", |
| "Güney Kore", | | "Dış İlişkiler ve Yurt Dışı İşçi Hizmetleri Genel Müdürlüğü", |
| "Gürcistan", | | "Dışişleri Bakanlığı", |
| "Haiti", | | "Din Öğretimi Genel Müdürlüğü", |
| "Hindistan", | | "Diyanet İşleri Başkanlığı", |
| "Hollanda", | | "DLH İnşaatı Genel Müdürlüğü", |
| "Honduras", | | "Doğa Koruma ve Millî Parklar Genel Müdürlüğü", |
| "Hong", | | "Doğal Afet Sigortaları Kurumu", |
| "Hırvatistan", | | "Ege Üniversitesi", |
| "Irak", | | "Ekonomi Bakanlığı", |
| "İngiltere", | | "Elektrik İşleri Etüt İdaresi Genel Müdürlüğü", |
| "İran", | | "Elektrik Üretim A.Ş. Genel Müdürlüğü", |
| "İrlanda", | | "Emeklilik Gözetim Merkezi", |
| "İspanya", | | "Emniyet Genel Müdürlüğü", |
| "İsrail", | | "Enerji Piyasaları İşletme A.Ş.", |
| "İsveç", | | "Enerji Piyasası Düzenleme Kurumu Başkanlığı", |
| "İsviçre", | | "Enerji ve Tabii Kaynaklar Bakanlığı", |
| "İtalya", | | "Engelli ve Yaşlı Hizmetleri Genel Müdürlüğü", |
| "İzlanda", | | "Erişim Sağlayıcıları Birliği", |
| "Jamaika", | | "Esnaf ve Sanatkarlar Genel Müdürlüğü", |
| "Japonya", | | "Eti Maden İşletmeleri Genel Müdürlüğü", |
| "Jersey", | | "Et ve Süt Kurumu Genel Müdürlüğü", |
| "Kamboçya", | | "Fatih Sultan Mehmet Üniversitesi", |
| "Kamerun", | | "Fırat Üniversitesi", |
| "Kanada", | | "Finansal Kurumlar Birliği", |
| "Karadağ", | | "Futbol Federasyonu Başkanlığı", |
| "Katar", | | "Gaziosmanpaşa Üniversitesi", |
| "Kayman", | | "Gelir İdaresi Başkanlığı", |
| "Kazakistan", | | "Gelir Politikaları Genel Müdürlüğü", |
| "Kenya", | | "Gençlik ve Spor Bakanlığı", |
| "Kiribati", | | "Genelkurmay Başkanlığı", |
| "Kolombiya", | | "Gıda, Tarım ve Hayvancılık Bakanlığı", |
| "Komorlar", | | "Gıda, Tarım ve Hayvancılık Bakanlığı Personel Genel Müdürlüğü", |
| "Kongo", | | "Gıda ve Kontrol Genel Müdürlüğü", |
| "Kosta Rika", | | "Giresun Üniversitesi", |
| "Kuveyt", | | "Göç İdaresi Genel Müdürlüğü", |
| "Kuzey Kore", | | "Gümrük ve Ticaret Bakanlığı", |
| "Küba", | | "Gümüşhane Üniversitesi", |
| "Kırgızistan", | | "Güneydoğu Anadolu Projesi Bölge Kalkınma İdaresi Başkanlığı", |
| "Laos", | | "Güzel Sanatlar Genel Müdürlüğü", |
| "Lesotho", | | "Haberleşme Genel Müdürlüğü", |
| "Letonya", | | "Harita Genel Komutanlığı", |
| "Liberya", | | "Harran Üniversitesi", |
| "Libya", | | "Hayat Boyu Öğrenme Genel Müdürlüğü", |
| "Liechtenstein", | | "Hayvancılık Genel Müdürlüğü", |
| "Litvanya", | | "Hazine Müsteşarlığı", |
| "Lübnan", | | "Hitit Üniversitesi", |
| "Lüksemburg", | | "İçişleri Bakanlığı", |
| "Macaristan", | | "İç Ticaret Genel Müdürlüğü", |
| "Madagaskar", | | "İlaç ve Eczacılık Genel Müdürlüğü", |
| "Çin", | | "İlksan", |
| "Makedonya", | | "İller Bankası A.Ş.", |
| "Malavi", | | "İller İdaresi Genel Müdürlüğü", |
| "Maldivler", | | "İnsan Hakları Başkanlığı", |
| "Malezya", | | "İnşaat Emlâk ve Nato Enfrastrüktür Dairesi Başkanlığı", |
| "Mali", | | "İskenderun Teknik Üniversitesi", |
| "Malta", | | "İstanbul Gedik Üniversitesi", |
| "Meksika", | | "İstanbul Teknik Üniversitesi", |
| "Mikronezya", | | "İŞKUR", |
| "Moldovya", | | "İş Sağlığı ve Güvenliği Genel Müdürlüğü", |
| "Monako", | | "Jandarma Genel Komutanlığı", |
| "Montserrat", | | "Jandarma ve Sahil Güvenlik Akademisi", |
| "Moritanya", | | "Kadının Statüsü Genel Müdürlüğü", |
| "Mozambik", | | "Kalkınma Bakanlığı", |
| "Moğolistan", | | "Kamu Denetçiliği Kurumu", |
| "Myanmar", | | "Kamu Düzeni ve Güvenliği Müsteşarlığı", |
| "Mısır", | | "Kamu Gözetimi Muhasebe ve Denetim Standartları Kurumu", |
| "Namibya", | | "Kamu İhale Kurumu", |
| "Nauru", | | "Kanserle Savaş Dairesi Başkanlığı", |
| "Nepal", | | "Kanunlar Genel Müdürlüğü ", |
| "Nijer", | | "Kara Kuvvetleri Komutanlığı", |
| "Nijerya", | | "Karamanoğlu Mehmetbey Üniversitesi", |
| "Nikaragua", | | "Karayolları Genel Müdürlüğü", |
| "Niue", | | "Karayolu Düzenleme Genel Müdürlüğü", |
| "Norveç", | | "Kırklareli Üniversitesi", |
| "Afrika", | | "Kıyı Emniyeti Genel Müdürlüğü", |
| "Özbekistan", | | "Kilis 7 Aralık Üniversitesi", |
| "Pakistan", | | "Konya Gıda ve Tarım Üniversitesi", |
| "Palau", | | "KTO Karatay Üniversitesi", |
| "Panama", | | "Küçük ve Orta Ölçekli İşletmeleri Geliştirme ve Destekleme İdaresi", |
| "Paraguay", | | "KOSGEB", |
| "Peru", | | "Kültür Varlıkları ve Müzeler Genel Müdürlüğü", |
| "Pitcairn", | | "Kültür ve Turizm Bakanlığı", |
| "Polonya", | | "Kültür ve Turizm Bakanlığı Araştırma ve Eğitim Genel Müdürlüğü", |
| "Portekiz", | | "Kültür ve Turizm Bakanlığı Döner Sermaye İşletmesi Merkez Müdürlüğü", |
| "Porto Riko", | | "Kültür ve Turizm Bakanlığı Yatırım ve İşletmeler Genel Müdürlüğü", |
| "Reunion", | | "Kütüphaneler ve Yayımlar Genel Müdürlüğü", |
| "Romanya", | | "Maden İşleri Genel Müdürlüğü", |
| "Ruanda", | | "Maden Tetkik ve Arama Genel Müdürlüğü", |
| "Rusya", | | "Mahalli İdareler Genel Müdürlüğü", |
| "Samoa", | | "Makina ve Kimya Endüstrisi Kurumu", |
| "Senegal", | | "Mali Suçları Araştırma Kurulu Başkanlığı", |
| "Seyşeller", | | "Maliye Bakanlığı", |
| "Sierra Leone", | | "Mardin Artuklu Üniversitesi", |
| "Singapur", | | "Mekansal Planlama Genel Müdürlüğü", |
| "Slovakya", | | "Merkezi Kayıt Kuruluşu A.Ş.", |
| "Slovenya", | | "Mesleki Hizmetler Genel Müdürlüğü", |
| "Somali", | | "Meslekî ve Teknik Eğitim Genel Müdürlüğü", |
| "Sri Lanka", | | "Mesleki Yeterlilik Kurumu", |
| "Sudan", | | "Meteoroloji Genel Müdürlüğü", |
| "Surinam", | | "Mevzuatı Geliştirme ve Yayın Genel Müdürlüğü", |
| "Suriye", | | "Millenicom", |
| "Arabistan", | | "Milli Eğitim Bakanlığı", |
| "Svaziland", | | "Milli Emlak Genel Müdürlüğü", |
| "Sırbistan", | | "Milli Güvenlik Kurulu Genel Sekreterliği", |
| "Sırbistan-Karadağ", | | "Milli İstihbarat Teşkilatı Müsteşarlığı", |
| "Şili", | | "Milli Kütüphane Başkanlığı", |
| "Tacikistan", | | "Milli Piyango İdaresi Genel Müdürlüğü", |
| "Tanzanya", | | "Millî Savunma Bakanlığı", |
| "Tayland", | | "Milli Savunma Bakanlığı Seferberlik Dairesi Başkanlığı", |
| "Tayvan", | | "Muhasebat Genel Müdürlüğü", |
| "Togo", | | "Necmettin Erbakan Üniversitesi", |
| "Tokelau", | | "Nüfus ve Vatandaşlık İşleri Genel Müdürlüğü", |
| "Tonga", | | "Okul İçi Beden Eğitimi Spor ve İzcilik Daire Başkanlığı", |
| "Tunus", | | "Orman Genel Müdürlüğü", |
| "Tuvalu", | | "Orman ve Su İşleri Bakanlığı", |
| "Türkmenistan", | | "Orta Anadolu İhracatçı Birlikleri Genel Sekreterliği", |
| "Uganda", | | "Ortaöğretim Genel Müdürlüğü", |
| "Ukrayna", | | "Osmaniye Korkut Ata Üniversitesi", |
| "Umman", | | "Öğretmen Yetiştirme ve Geliştirme Genel Müdürlüğü", |
| "Uruguay", | | "Ölçme, Seçme ve Yerleştirme Merkezi", |
| "Ürdün", | | "ÖSYM", |
| "Vanuatu", | | "Ölçüler ve Standartlar Genel Müdürlüğü ", |
| "Vatikan", | | "Ömer Halisdemir Üniversitesi", |
| "Venezuela", | | "Özel Eğitim ve Rehberlik Hizmetleri Genel Müdürlüğü", |
| "Vietnam", | | "Özelleştirme İdaresi Başkanlığı", |
| "Yemen", | | "Özel Öğretim Kurumları Genel Müdürlüğü", |
| "Yunanistan", | | "Petrol İşleri Genel Müdürlüğü", |
| "Zambiya", | | "Polis Akademisi Başkanlığı", |
| "Zimbabve" | | "PTT", |
|  | | "Radyo ve Televizyon Üst Kurulu", |
|  | | "Refik Saydam Hıfzısıhha Merkezi Başkanlığı", |
|  | | "Rekabet Kurumu", |
|  | | "Sağlık Bakanlığı", |
|  | | "Sağlık Bakanlığı Hudut ve Sahiller Sağlık Genel Müdürlüğü", |
|  | | "Sağlık Bakanlığı Personel Genel Müdürlüğü", |
|  | | "Sahil Güvenlik Komutanlığı", |
|  | | "Sanayi Bölgeleri Genel Müdürlüğü", |
|  | | "Sanayi Genel Müdürlüğü", |
|  | | "Savunma Sanayii ve Teknoloji Eğitim Merkezi", |
|  | | "SATEM" |
|  | | "Savunma Sanayi Müsteşarlığı", |
|  | | "STM", |
|  | | "Sayıştay", |
|  | | "Sermaye Piyasası Kurulu", |
|  | | "Sıtma Savaş Dairesi Başkanlığı", |
|  | | "Sigorta Bilgi ve Gözetim Merkezi", |
|  | | "Sinema Genel Müdürlüğü", |
|  | | "Sinop Üniversitesi", |
|  | | "Sivil Havacılık Genel Müdürlüğü", |
|  | | "Sosyal Güvenlik Kurumu", |
|  | | "Sosyal Yardımlar Genel Müdürlüğü", |
|  | | "Spor Genel Müdürlüğü", |
|  | | "Su Yönetimi Genel Müdürlüğü", |
|  | | "Sümer Halıcılık ve El Sanatları Sanayi ve Ticaret A.Ş.", |
|  | | "Şeker Kurumu", |
|  | | "Tabiat Varlıklarını Koruma Genel Müdürlüğü", |
|  | | "Takasbank", |
|  | | "Talim ve Terbiye Kurulu Başkanlığı", |
|  | | "Tanıtma Fonu Kurulu", |
|  | | "Tanıtma Genel Müdürlüğü", |
|  | | "Tapu ve Kadastro Genel Müdürlüğü", |
|  | | "Tarım İşletmeleri Genel Müdürlüğü", |
|  | | "Tarım Reformu Genel Müdürlüğü", |
|  | | "Tarımsal Araştırmalar ve Politikalar Genel Müdürlüğü", |
|  | | "Tarımsal Ekonomi ve Politika Geliştirme Enstitüsü", |
|  | | "Tarım Sigortaları Havuzu", |
|  | | "Tarım ve Kırsal Kalkınmayı Destekleme Kurumu Başkanlığı", |
|  | | "Tasarruf Mevduatı Sigorta Fonu Başkanlığı", |
|  | | "TBMM", |
|  | | "TCDD", |
|  | | "T.C. Mehmet Akif Ersoy Üniversitesi", |
|  | | "Telif Hakları Genel Müdürlüğü", |
|  | | "Temel Eğitim Genel Müdürlüğü", |
|  | | "Toplu Konut İdaresi Başkanlığı", |
|  | | "TOKİ", |
|  | | "Toprak Mahsulleri Ofisi", |
|  | | "TRT", |
|  | | "Tüketicinin Korunması ve Piyasa Gözetimi Genel Müdürlüğü", |
|  | | "Türk Akreditasyon Kurumu", |
|  | | "Türk-Alman Üniversitesi", |
|  | | "Türk Dil Kurumu", |
|  | | "Türk İşbirliği ve Koordinasyon Ajansı Başkanlığı", |
|  | | "Türkiye Adalet Akademisi Başkanlığı", |
|  | | "Türkiye Atom Enerjisi Kurumu Başkanlığı", |
|  | | "Türkiye Bankalar Birliği", |
|  | | "Türkiye Bilimler Akademisi", |
|  | | "TÜBA", |
|  | | "Türkiye Bilimsel ve Teknolojik Araştırma Kurumu", |
|  | | "TÜBİTAK", |
|  | | "Türkiye Cumhuriyet Merkez Bankası", |
|  | | "Türkiye Demiryolu Makinaları Sanayii", |
|  | | "TÜDEMSAŞ" |
|  | | "Türkiye Elektrik Dağıtım A.Ş.", |
|  | | "Türkiye Elektrik İletim A.Ş.", |
|  | | "TEİAŞ", |
|  | | "Türkiye Elektrik Ticaret ve Taahhüt A.Ş.", |
|  | | "TETAŞ", |
|  | | "Türkiye Elektromekanik Sanayii A.Ş. Genel Müdürlüğü", |
|  | | "Eximbank", |
|  | | "Türkiye İstatistik Kurumu Başkanlığı", |
|  | | "TÜİK", |
|  | | "Türkiye Kalkınma Bankası A.Ş. Genel Müdürlüğü", |
|  | | "Türkiye Kömür İşletmeleri Kurumu Genel Müdürlüğü", |
|  | | "Türkiye Lokomotif ve Motor Sanayii A.Ş.", |
|  | | "Türkiye Muhasebe Standartları Kurulu", |
|  | | "Türkiye Odalar ve Borsalar Birliği", |
|  | | "TOBB", |
|  | | "Türkiye Petrolleri Anonim Ortaklığı Genel Müdürlüğü", |
|  | | "Türkiye Şeker Fabrikaları A. Ş.", |
|  | | "Türkiye Tarım Kredi Kooperatifleri Birliği", |
|  | | "Türkiye Taşkömürü Kurumu Genel Müdürlüğü", |
|  | | "Türkiye Ulusal Ajansı", |
|  | | "Türkiye Vagon Sanayii A.Ş.", |
|  | | "Türkiye ve Orta Doğu Amme İdaresi Enstitüsü ", |
|  | | "Türkiye Yatırım Destek ve Tanıtım Ajansı Başkanlığı", |
|  | | "Türk Patent ve Marka Kurumu", |
|  | | "Türk Standardları Enstitüsü", |
|  | | "Türk Tarih Kurumu", |
|  | | "Tütün ve Alkol Piyasası Düzenleme Kurumu", |
|  | | "Ulaştırma, Denizcilik ve Haberleşme Bakanlığı", |
|  | | "Ulusal Bor Araştırma Enstitüsü Başkanlığı", |
|  | | "Uşak Üniversitesi", |
|  | | "Üsküdar Üniversitesi", |
|  | | "Vakıflar Genel Müdürlüğü", |
|  | | "Vergi Denetim Kurulu Başkanlığı", |
|  | | "Verimlilik Genel Müdürlüğü", |
|  | | "Yapı İşleri Genel Müdürlüğü", |
|  | | "Yargıtay", |
|  | | "Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü", |
|  | | "Yıldırım Beyazıt Üniversitesi", |
|  | | "Yurt Dışı Türkler ve Akraba Topluluklar Başkanlığı", |
|  | | "Yüksek Öğrenim Kredi ve Yurtlar Kurumu Genel Müdürlüğü", |
|  | | "Yükseköğretim Kurulu Başkanlığı", |
|  | | "YÖK", |
|  | | "Yüksek Seçim Kurulu Başkanlığı" |
|  | | "YSK", |

The software has a logger mechanism. Because the logger related codes are also configurable, the logger codes are stored in “config.py” file too.

### hurriyetApiToDb.py File

One of the processes fetches the news from Hurriyet Newspaper API and store them in the database. It uses an open source Hurriyet API wrapper coded for python language. It imports the following python libraries to use.

* requests
* json
* time
* sched

The code does the following explaining in pseudo-code.

for each keyword in cybersecurity keyword list:

data = api.search(keyword)

writeDatabase(data)

### ituNlpPipeline.py File

It has the process code regularly checks the database for finding new entries. If it can find, the process sends the raw data to normalize it with using the ITU NLP API. After that, it updates the related field of the database with the respond texts of the API request.

High-level pseudo-code of that process is as the following.

1. Create a database connection
2. Select the rows of the database which has Status column = ‘0’
3. Send the selected rows one by one to ITU NLP API to normalize their texts.
4. Wait 10 seconds between each request
5. Update the database with the responded text of ITU NLP API request and set their Status column with ‘1’
6. Wait one minute before next control of the database for new entries.

It imports the following python libraries to use.

* time
* re
* urllib

### securityEventsWebPortal.py File

It has the process code which regularly checks the database for finding new entries with Status ‘1’ column. Status=’1’ means that the row is normalized, filtered its alphanumeric characters, and omit the rest of the characters and ready for analyzing to find cybersecurity event. If it can find, the process dynamically creates or updates the user interface HTML page. Then set the processed row’s Status column with ‘2’. ‘2’ means that that row is used before and the user interface is updated according to the processed row’s results.

It imports the following python libraries to use.

* re
* string
* datetime
* BeautifulSoup
* import time

According to the crummy website “Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree.”(Richardson, 2015)

### sqliteOperations.py File

We develop a library for commonly used SQLite operations in our software system. It contains the following methods.

* Create SQLite Table
* Create Connection
* Select Task by Status
* Update Text Column of the Row by Status

It imports the following python libraries to use.

* sqlite3
* json
* datetime
* BeautifulSoup

### twitterStreamToDb.py File

One of the processes listen to Twitter streams and store the streaming data in the database. The process codes are in this code file. We use the Tweepy library for this implementation. According to Tweepy website, Tweepy library is “ An easy-to-use Python library for accessing the Twitter API.“ (“Tweepy,” n.d.) It imports the following Python libraries to use.

* tweepy
* string
* sqlite3
* time

The code does the following explaining in pseudo-code.

1. Connect the Twitter API with using Tweepy library
2. Listen to Twitter Streams and filter Turkish Tweets, which includes the cybersecurity keywords in Tweets.
3. Write the filtered data to the database and set their status with 0.

### userInterface.html File

At first, it includes a template HTML file. “securityEventsWebPortal.py” python code update and modify this HTML file and populates with cybersecurity-related events. The HTML refreshes itself regularly in ten seconds periods.

**CHAPTER 6**

# RESULTS

In this section, we discuss the results of the cybersecurity events which are discovered by our software solution. We focus on what our software system succeeded and what it did not achieve. We share successful cybersecurity event detection samples and share the not successful cybersecurity event detection samples.

In Figure 14 below, the user interface of our cybersecurity detection software can be seen. As described in the previous subsection, it is a dynamically created HTML page. We divide the events by their dates. As cybersecurity event information, we represent an entity, a representative news title or tweet and a count which shows how many times the entity is seen in the data on the same day.

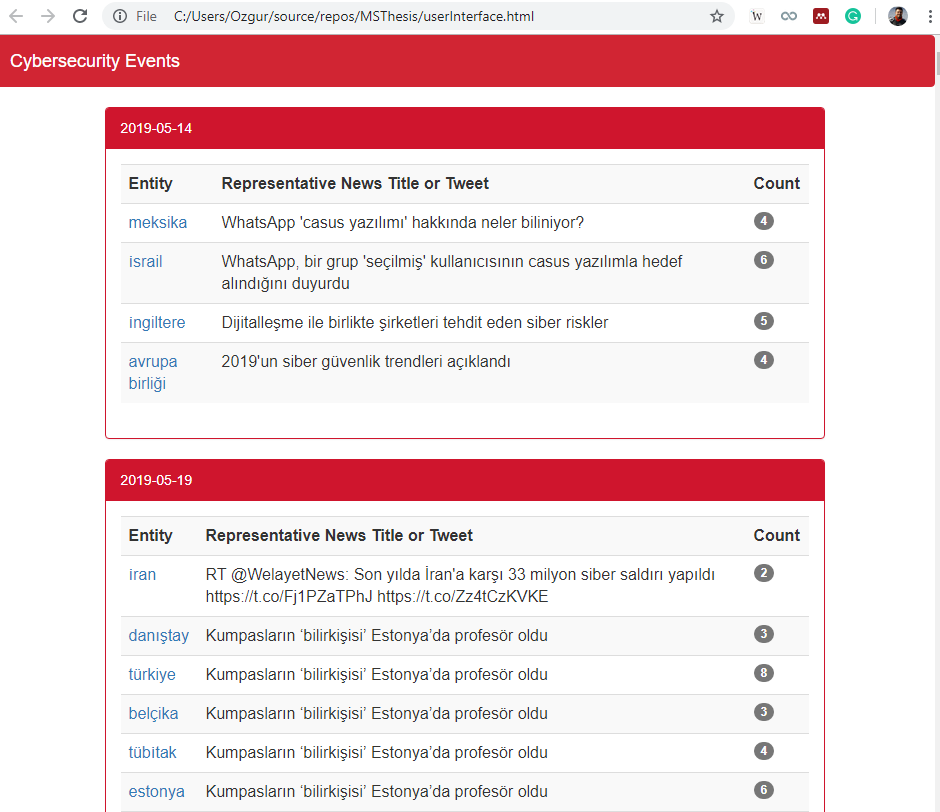


Figure 14 User Interface of the Cybersecurity Detection Software

## Successful Cybersecurity Event Detection Samples

In the following subsections, we share successful cybersecurity event detection samples and briefly try to explain how a security analyst can use this information.

### WhatsApp Spyware Attack

As can be seen in the figure below, our software system can detect this event on 5 May 2019. However, there are two different entities about the same event.



Figure 15 WhatsApp Spyware Attack Detection

Assume that a security analyst wants to track security events related to countries. When the security analyst sees the “WhatsApp Spyware Attack” event in the user interface page with a country name entity, he should check the news or tweets to control whether it is a positive or false positive event detection. If it is a positive and useful cybersecurity event detection, the security analyst takes the required actions.

There are two entities as “meksika” which is the Turkish synonym of Mexico, and “israil” which is the Turkish synonym of Israel.

When we control the related news and tweets, we can see that an Israel firm named NSO Group performs the cyber-attack. Therefore “israil” is passing six times in the detected news and tweets.

A Mexican journalist is affected by the cyber attack. That is why we capture the “meksika” entity.

The security analyst can notice such attack with following our software solutions user interface and can learn what the new WhatsApp cyber-attack is, how one can protect from such attacks and so on from the related news and tweets.

### Vulnerabilities in Remote Patient Tracking System Applications

STM is a Turkish software company which does researches about cybersecurity domain. They find a vulnerability about Remote Patient Tracking System Applications and share this information from Twitter and with using newspapers.



Figure 16 STM Warns about Remote Patient Tracking System Applications

If our software solution were to have used English texts as a data source, we could not detect such a cybersecurity event published in Turkish. Because of our software solution can analyze Turkish texts, we can detect such a cybersecurity event. This is an excellent example to show what our solution can do while the other solutions in the literature cannot do.

### Other Successful Detection Examples

We share the following figures to exemplify the success of our software event detection solution.



Figure 17 Supply Chain Attack Targets ASUS Computers Through Backdoored Update



Figure 18 MuddyWater Attack



Figure 19 MegaCortex Ransomware



Figure 20 Using Popular TV Shows to Spread Malware



Figure 21 RTM Banking Trojan



Figure 22 Mother's Day Cyberattack Alerts



Figure 23 Valentine’s Day Phishing Attacks



Figure 24 Pirate Matryoshka Virus



Figure 25 Cyber Attacks Using Smart TVs



Figure 26 Smart Home Cyberattacks



Figure 27 Valletta Bank Cyberattack



Figure 28 Business Email Compromise Attack



Figure 29 Cyberattacks Which Use Netflix Brand



Figure 30 Cyberattacks Targeting Job Seekers



Figure 31 Phishing Attacks Targeting Social Media Users



Figure 32 Zombie Cookies



Figure 33 Millions of Email Addresses Infiltrated the Internet



Figure 34 The USA and China Cyberwar Started Tweet



Figure 35 Angela Merkel and Hundreds of German Politicians Hacked

## Unsuccessful Cybersecurity Event Detection Samples

Sometimes our software solution can detect false positive events, or even it is a cybersecurity event, the detection may not be a useful event for security analysts. The following subsections examine such scenarios.

### Sample False Positive Cybersecurity Event Detection

In the figure below, you can see an event from our user interface. Even the tweet has “hacklendi” word, which is one of our keywords from our keyword vector; the event is not a real cybersecurity event. Analyzing such tweets to realize that it is not a real security event is hard for an automated system.

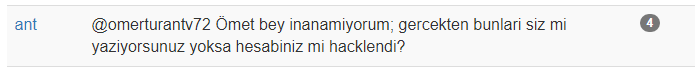


Figure 36 Sample False Positive Cybersecurity Event Detection

### Sample not Useful Cybersecurity Event Detection

Sometimes, even the detected event is a cybersecurity event; it may be a personal status primarily if it is published on Twitter. Security analysts should read the detected event from the user interface and decide that it is useful or not for him.

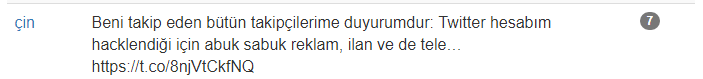


Figure 37 Sample, not Useful Cybersecurity Event Detection

**CHAPTER 7**

# CONCLUSION AND FUTURE WORK

## Conclusion

In the last few decades, automation has been increasingly used in various field of people’s life due to its benefits like cost reduction, productivity, availability, reliability, and performance. Cybersecurity is one of the fields which automation is often used. However, every automation software system has unique requirements to achieve its purposes. It leads to lots of research areas and unique automation systems. Automatic event detection is one of these research fields. Social media is one of the fastest ways to detect cybersecurity events because people and bots share such events in there. Newspapers are also shared such cybersecurity events and processing the newspaper data is relatively more straightforward because false-positive cybersecurity events are rarely shared in the newspaper websites.

In this thesis, we investigated automatic event detection of cyber security events from Turkish Twitter Stream and Turkish newspaper data. We work on real-time data to achieve that our research can be used by security analysts. Existing publications about real-time cybersecurity event detection system generally use English texts to analyze and detect the events. We cannot find any research which use Turkish data sources to detect cybersecurity events. Using Turkish data sources for cybersecurity event detection is a new topic for literature. We believe that this research contributes to the literature by filling an uninvestigated field. We proposed an automated software system which works using different data sources, named entities, text mining methods, and "state of art" software techniques. Then we analyze the results of our software system. Even if our software system detects few false positive cybersecurity events, it was often able to detect a useful cybersecurity event. For example, our software system can detect cybersecurity events such as WhatsApp Spyware, MuddyWater Attack, the Remote Patient Tracking System Applications vulnerability, Pirate Matryoshka Virus, Zombie Cookies threat.

We concluded that event detection with using Turkish texts is applicable, and security analysts can use such a system like our software system as a helper tool.

## Future Work

Currently, our software system works on a local computer. We will plan to move our project to a server, obtain a website. After this, security analysts do not need to work on software codes. They can easily follow current Turkish security events from a website. Even journalists or regular people can visit our software to follow security events. When we move the software to a server(i.e., AWS), our software can work 7x24, which will be useful for detection success. If our software can work with bigger data, it will detect more events with more accurate event detection. To increase the streaming data, we are planning to add new Turkish data sources from other websites like Eksisozluk, Linkedin, Facebook, and so on. This improvement will make our datasets an excellent resource for future work. After these improvements, our datasets can be useful not only for us but also the other researchers work on cybersecurity, cognitive science or computer science field.

Moreover, we may try to detect cybersecurity events that have not yet taken place by processing streaming data using Turkish. Even there are researches about it; these researches do not use Turkish texts as a data source. We may research that topic by using Turkish texts using our infrastructure done in the scope of this thesis.

Lastly, we can add malicious URL detection to our software system, as we mentioned in the literature review section. Sometimes people share malicious URLs from social media. We may try to detect such an attack using Turkish Tweets.

# REFERENCES

Aslan, Ç. B., Sağlam, R. B., & Li, S. (2018). *Automatic Detection of Cyber Security Related Accounts on Online Social Networks*. 236–240. https://doi.org/10.1145/3217804.3217919

Borrett, M. (2017). *Security in the Cognitive Era BRINGING THE POWER OF COGNITIVE SECURITY TO THE SECURITY ANALYST Motivations for Change*. Retrieved from http://www.crestcon.co.uk/wp-content/uploads/2017/04/MartinBorrett.pdf

Businesses Can Now Use The Same Stats Language As Universities, Thanks. (n.d.). Retrieved April 19, 2019, from https://www.fastcompany.com/3030877/businesses-can-now-use-the-same-stats-language-as-universities-thanks-to-pandas

Countries with most Twitter users 2019 | Statistic. (2019). Retrieved April 15, 2019, from Statista website: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Eryiğit, G. (2015). *ITU Turkish NLP Web Service*. 1–4. https://doi.org/10.3115/v1/e14-2001

Eryiǧit, G., & Torunoǧlu-Selamet, D. (2017). Social media text normalization for Turkish. *Natural Language Engineering*, *23*(6), 835–875. https://doi.org/10.1017/S1351324917000134

Global Digital Report 2019 | Free Download | We Are Social UK. (n.d.). Retrieved April 26, 2019, from https://wearesocial.com/uk/digital-2019

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social Networks that Matter: Twitter Under the Microscope. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1313405

Hurriyet Developers API v1.0 Docs — Hürriyet Public API. (n.d.). Retrieved April 11, 2019, from https://developers.hurriyet.com.tr/docs/versions/1.0

*International Journals of Management, IT et Engineering IJMIE.* (n.d.). Retrieved from https://www.academia.edu/35925695/TEACHBOT\_TEACHING\_ROBOT\_USING\_ARTIFICIAL\_INTELLIGENCE\_AND\_NATURAL\_LANGUAGE\_PROCESSING

ITU Turkish Natural Language Processing Web Interface. (n.d.). Retrieved April 18, 2019, from http://tools.nlp.itu.edu.tr/index.jsp

Javed, A., Burnap, P., & Rana, O. (2019). Prediction of drive-by download attacks on Twitter. *Information Processing and Management*, *56*(3), 1133–1145. https://doi.org/10.1016/j.ipm.2018.02.003

Khandpur, R. P., Ji, T., Jan, S., Wang, G., Lu, C.-T., & Ramakrishnan, N. (2017). *Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media*. https://doi.org/10.1145/3132847.3132866

Kwak, H., Lee, C., Park, H., & Moon, S. (n.d.). *What is Twitter, a Social Network or a News Media?* Retrieved from http://bit.ly

Langley, A. (2013). Enhancing digital certificate security. Retrieved April 15, 2019, from https://security.googleblog.com/2013/01/enhancing-digital-certificate-security.html

Overview of Artificial Intelligence &amp; Role of NLP in Big Data - XenonStack Blog. (n.d.). Retrieved April 18, 2019, from https://www.xenonstack.com/blog/ai-nlp-big-deep-learning/

Petersen, J. (2017). Sonar. *Handbook of Surveillance Technologies, Third Edition*, (August), 223–291. https://doi.org/10.1201/b11594-7

Python. (2017). What is Python? Executive Summary | Python.org. *Python Software Foundation*. Retrieved from https://www.python.org/doc/essays/blurb/

Richardson, L. (2015). Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. Retrieved May 15, 2019, from crummy.com website: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Ritter, A., Wright, E., Casey, W., & Mitchell, T. (2016). *Weakly Supervised Extraction of Computer Security Events from Twitter*. 896–905. https://doi.org/10.1145/2736277.2741083

Sqlite.org. (2013). About SQLite. Retrieved April 19, 2019, from https://www.sqlite.org/about.html

Stephanie Prato. (2013). What is Text Mining? - Information Space. Retrieved April 18, 2019, from https://ischool.syr.edu/infospace/2013/04/23/what-is-text-mining/

Text Mining in Data Mining - Concepts, Process & Applications - DataFlair. (n.d.). Retrieved April 29, 2019, from https://data-flair.training/blogs/text-mining/

Text mining vs data mining: discover the differences -. (n.d.). Retrieved April 18, 2019, from https://www.expertsystem.com/text-mining-vs-data-mining-differences/

Tiraj | MedyaTava - Yazmadıysa Doğru Değildir. (n.d.). Retrieved April 11, 2019, from http://www.medyatava.com/tiraj/2018-01-08

Tweepy. (n.d.). Retrieved May 15, 2019, from https://www.tweepy.org/

Twitter. (2019). Twitter: Number of active users 2010-2017. Retrieved April 15, 2019, from statista.com website: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Understanding normalization - Natural Language Processing with Java. (n.d.). Retrieved April 29, 2019, from https://subscription.packtpub.com/book/application\_development/9781784391799/2/ch02lvl1sec20/understanding-normalization

Wang, Z., & Zhang, Y. (2017). DDoS event forecasting using twitter data. *IJCAI International Joint Conference on Artificial Intelligence*, 4151–4157. Retrieved from https://code.google.com/p/word2vec/

What does an information security analyst do? ‐ CareerExplorer. (n.d.). Retrieved April 18, 2019, from https://www.careerexplorer.com/careers/information-security-analyst/

What is a Security Analyst? Responsibilities, Qualifications, and More | Digital Guardian. (n.d.). Retrieved April 29, 2019, from https://digitalguardian.com/blog/what-security-analyst-responsibilities-qualifications-and-more

What is Named-entity recognition (NER)? - WordLift. (n.d.). Retrieved April 30, 2019, from https://wordlift.io/blog/en/entity/named-entity-recognition/

Yanlış sertifika Google’dan döndü - Teknoloji Haberleri. (n.d.). Retrieved April 16, 2019, from http://www.hurriyet.com.tr/teknoloji/yanlis-sertifika-googledan-dondu-22290509

# APPENDICES

# APPENDIX A

**TITLE**

xxx

# APPENDIX B

**TITLE**

xxx

1. “Normalization is a process that converts a list of words to a more uniform sequence. This is useful in preparing text for later processing” (“Understanding normalization - Natural Language Processing with Java,” n.d.) [↑](#footnote-ref-1)
2. Named-entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, places, expressions of times, quantities, monetary values, percentages and more.(“What is Named-entity recognition (NER)? - WordLift,” n.d.) [↑](#footnote-ref-2)