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 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 by means of 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 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 the 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 is able to 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 farklı ve zamana bağlı bilgiler içerir. Güvenlik uzmanları her gün interneti tarıyor ve insanları, kurumları ve hükümetleri etkileyen güvenlik olaylarıyla karşı karşıya kalıyorlar. Bir bilgi analisti gereken güvenlik olayları hakkında güncel kalmak için kaynakları düzenli olarak inceler. Bu durum, uygun araçlara sahip olmadıkları takdirde bilgi 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 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 gerekir. Ancak, çok sayıda olay nedeniyle, güvenlik analistlerinin ve araştırmacıların bu bilgi akışını yeterli şekilde ele alması zor olabilir. Bu sorunu çözmek için akademik ve ticari başarılı araştırmalar var. Bilgi analistleri bu tür teknik çözümleri günlük rutinlerinde aktif olarak kullanırlar.

Öte yandan, internet bilgi paylaşmak için farklı dilleri kullanır. 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 araştırıyor. 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](#_Toc6798692)

[ÖZ\_\_ vi](#_Toc6798693)

[DEDICATION viii](#_Toc6798694)

[ACKNOWLEDGMENTS ix](#_Toc6798695)

[TABLE OF CONTENTS x](#_Toc6798696)

[LIST OF TABLES xii](#_Toc6798697)

[LIST OF FIGURES xiii](#_Toc6798698)

[LIST OF ABBREVIATIONS xiv](#_Toc6798699)

[1 INTRODUCTION 15](#_Toc6798700)

[1.1 Motivation 15](#_Toc6798701)

[1.2 Objectives 19](#_Toc6798702)

[1.3 Use Cases 19](#_Toc6798703)

[1.4 Outline 20](#_Toc6798704)

[2 BACKGROUND INFORMATION 21](#_Toc6798705)

[2.1 What is an Information Security Analyst? 21](#_Toc6798706)

[2.2 What is Natural Language Processing? 22](#_Toc6798707)

[2.3 What is Text Mining? 23](#_Toc6798708)

[2.4 Twitter Social Network 24](#_Toc6798709)

[2.4.1 What is Twitter? 24](#_Toc6798710)

[2.4.2 Twitter API 25](#_Toc6798711)

[2.5 Hürriyet Turkish Newspaper as a Data Source 26](#_Toc6798712)

[2.5.1 What is Hurriyet Newspaper? 26](#_Toc6798713)

[2.5.2 Hurriyet Newspaper API 26](#_Toc6798714)

[2.5.3 What is OData 26](#_Toc6798715)

[2.5.4 Hürriyet API - OData Usage 27](#_Toc6798716)

[2.6 Python Programming Language 28](#_Toc6798717)

[2.7 Istanbul Technical University NLP API 28](#_Toc6798718)

[2.8 SQLite Database 29](#_Toc6798719)

[2.9 Other Technologies Used in the Thesis Project 29](#_Toc6798720)

[3 LITERATURE REVIEW 30](#_Toc6798721)

[3.1 Weakly Supervised Extraction of Computer Security Events from Twitter 30](#_Toc6798722)

[3.2 Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an Example 32](#_Toc6798723)

[3.3 DDoS Event Forecasting using Twitter Data 32](#_Toc6798724)

[3.4 Prediction of drive-by download attacks on Twitter 33](#_Toc6798725)

[3.5 SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream 33](#_Toc6798726)

[3.6 Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media 35](#_Toc6798727)

[4 SYSTEM ARCHITECTURE AND DESIGN 36](#_Toc6798728)

[4.1 Approach 36](#_Toc6798729)

[4.2 Taxonomy 38](#_Toc6798730)

[4.3 Data Collection 38](#_Toc6798731)

[4.4 Data Preprocessing 38](#_Toc6798732)

[4.5 Cybersecuirty Related Event Detection 38](#_Toc6798733)

[5 IMPLEMENTATION 38](#_Toc6798734)

[6 RESULTS 38](#_Toc6798735)

[7 CONCLUSION AND FUTURE WORK 39](#_Toc6798736)

[7.1 Conclusion 39](#_Toc6798737)

[7.2 Future Work 39](#_Toc6798738)

[REFERENCES 39](#_Toc6798739)

[APPENDICES 42](#_Toc6798740)

[APPENDIX A 42](#_Toc6798741)

[APPENDIX B 43](#_Toc6798742)

# 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#_Toc7028107)

[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#_Toc7028108)

[Table 3 Seed Instances for DDoS Attacks. 32](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028109)

[Table 4 Tweet Examples with Attack Targets. 33](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028110)

# 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#_Toc7028062)

[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#_Toc7028063)

[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#_Toc7028064)

[Figure 4: A Simple diagram to explain what NLP does. 24](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028065)

[Figure 5 Sample Turkish Tweets Related with a Security Incident 26](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028066)

[Figure 6: Architecture of the Keyword Finder Component. 35](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028067)

[Figure 7: Technical Overview of Sonar. 36](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028068)

[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#_Toc7028069)

[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. 37](file:///C:\Users\Ozgur\source\repos\MSThesis\thesis\ozgur_ural_cybersecurity_thesis.docx#_Toc7028070)

# 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 at 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 a 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? This 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 and so on. 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 infortmation 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 Ekşi Sözlük 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 has to 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 presented in this chapter.
* In Chapter 3, we present relavent literature.
* In Chapter 4, we introduce the software system in 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 analyst’s 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 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.

In order to develop automated systems, NLP is one of the actively used concepts in text mining. According to 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.

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

## 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.) and 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 are 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 with 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 which 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 system, a software implementation is required. We used Python Programming Language to implement the stystem. Python Programming Language and its selection criterias 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 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 I 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 the 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**

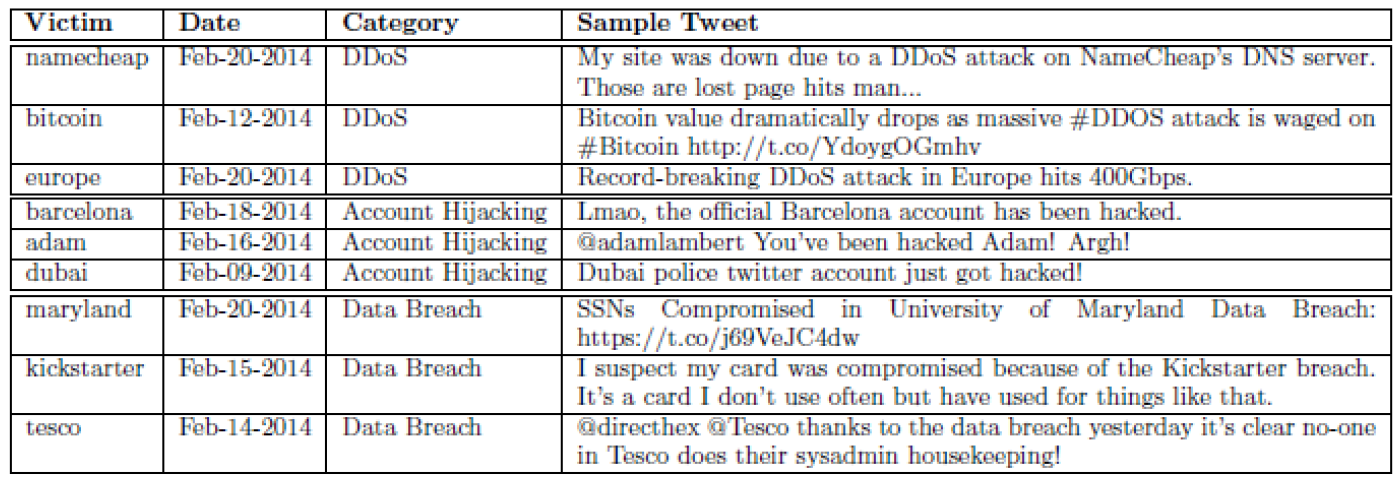
# LITERATURE REVIEW

In this section, we share the results of the literature review. We introduce the most relevant researches with my research and explain how they give shape to my 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.



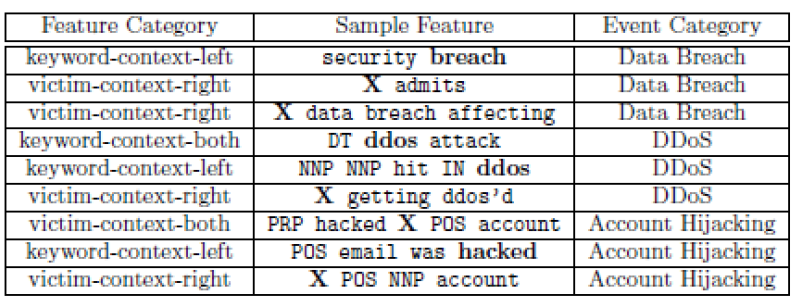
They determine candidate events as in Table 2.

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.

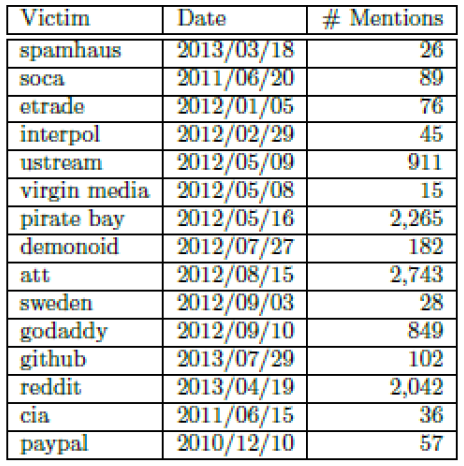


Table 3 Seed Instances for DDoS Attacks.

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.

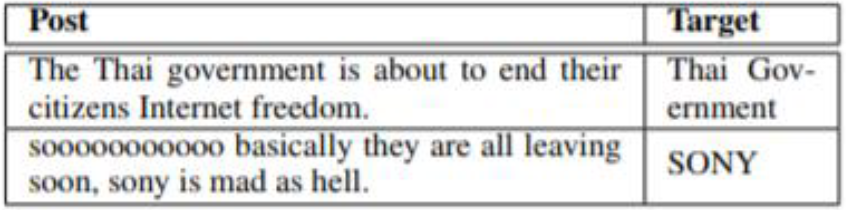


Table 4 Tweet Examples with Attack Targets.

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 with 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)

My thesis research may be useful to inform 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 a future work. However, we will try to detect such an attack with 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.

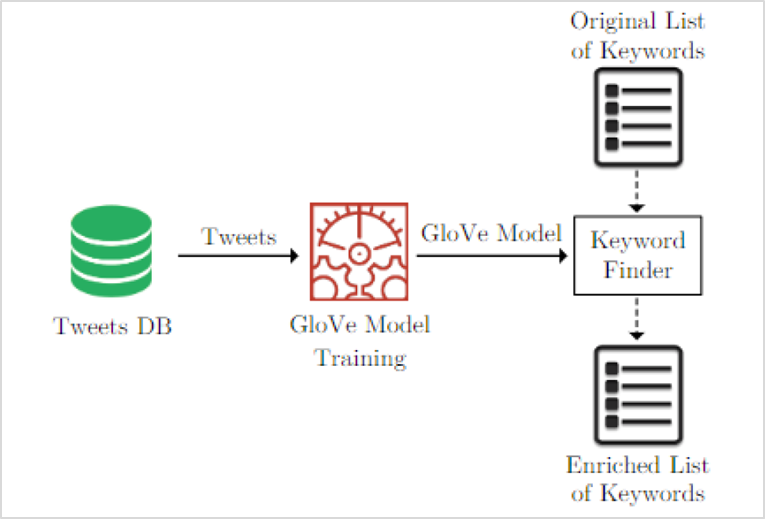


Figure 6: Architecture of the Keyword Finder Component.

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

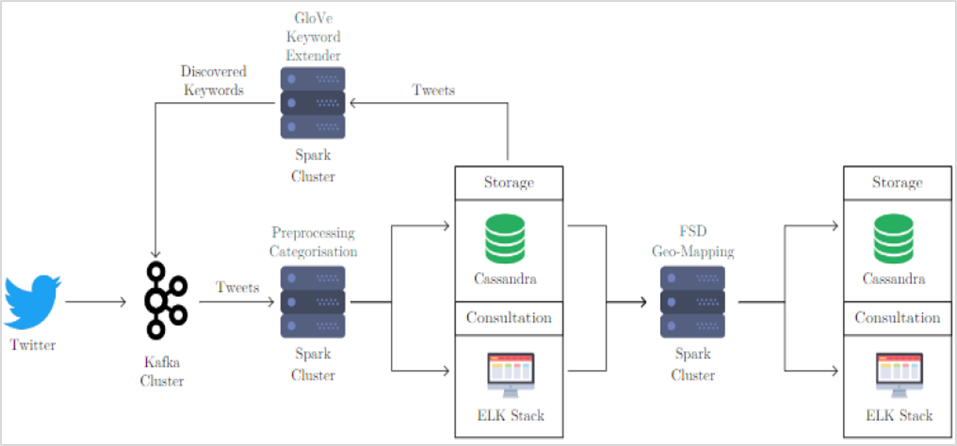


Figure 7: Technical Overview of Sonar.

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

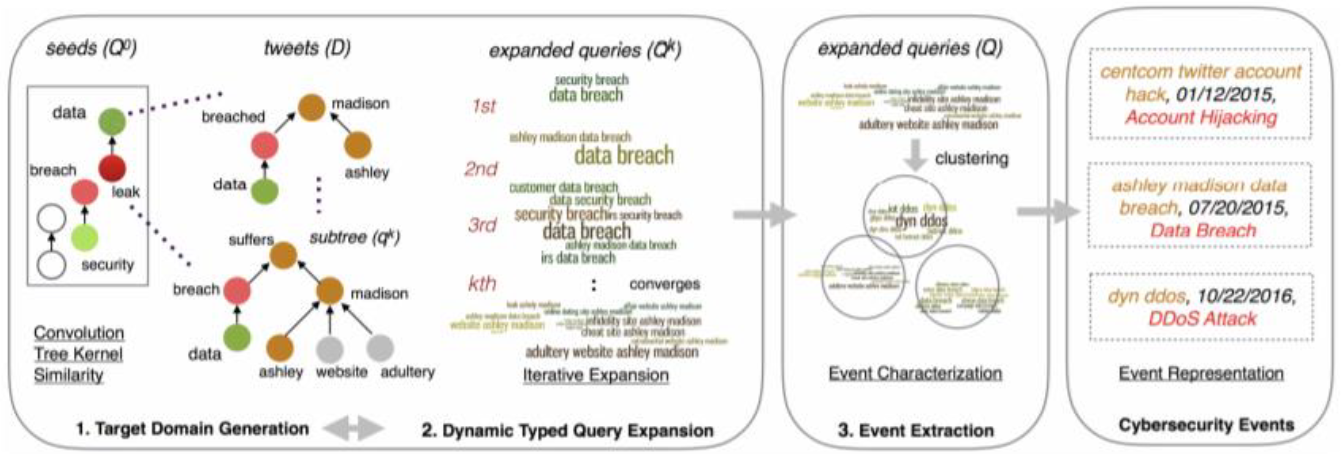
It is another study on detecting cybersecurity attacks by processing Twitter data. They acknowledge that their work is similar to that of previous studies, but they claim to have more successful results. (Khandpur et al., 2017)

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

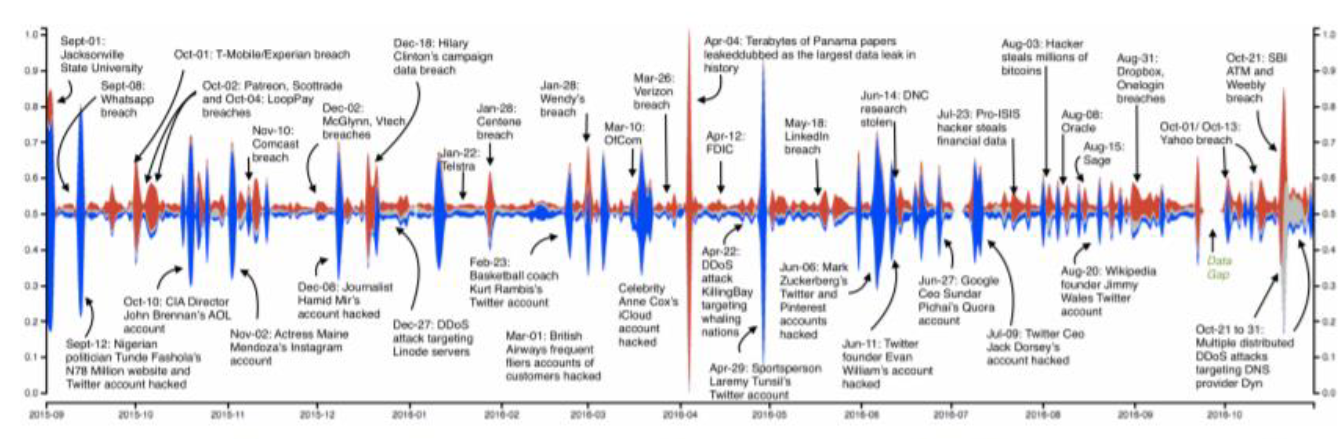


Figure 9: Streamgraph Showing Normalized Volume of Tweets (September 2015 through October 2016) Tagged with Data Breach (red), DDoS Activity (grey) and Account Hijacking (blue) Types of Cybersecurity Events.

This research one of the “state-of-art project” in cyber attack detection field. we also use this research to detect our boundry 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 present thesis.

**CHAPTER 4**

# SYSTEM ARCHITECTURE AND DESIGN

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

## Approach

Our software system has a configuration file. The configuration file includes constant values such as Twitter API and İTU 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 Twitter stream connection, the software uses configuration file values which are statically defined in configuration file. We use a cybersecurity related Turkish keyword vector to gather useful Twitter stream for our research. Moreover, we use language filter feature of the Twitter API in order to fetch only the 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.



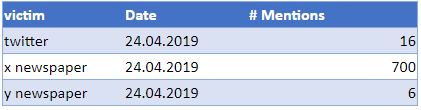
Figure 10: Normalizer Sample from ITU NLP API

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

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

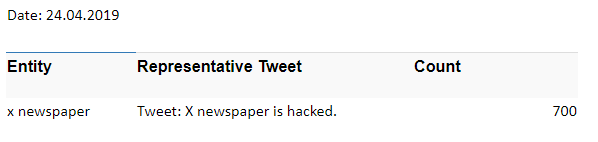
With this analyze we will get a table like as shown in the following. The period of the table is day. Every new day, number of mentions per victim is set and start with zero.

Table 5: Sample Table after the Analyze



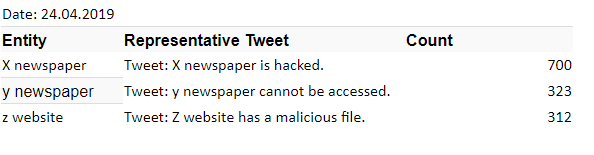
We add daily thresholds. (i.e. 10 per day) If the number of mentions is more than the thresholds value, we will share this information within a user interface like the following Figure.

Table 6 Sample Information Sharing Plan Table

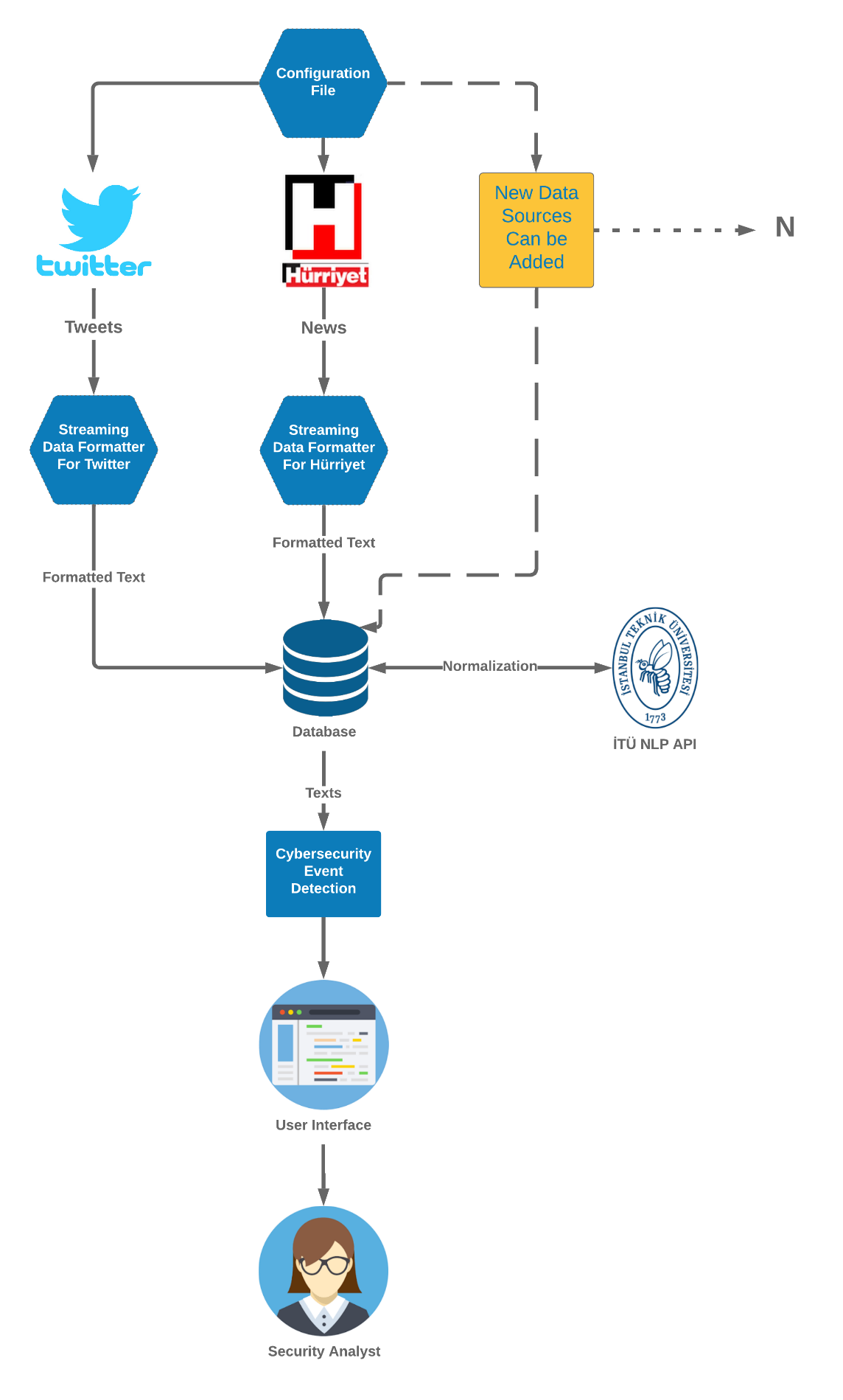


Permanently, the software checks the database and analyze new texts for detecting new cybersecurity events. If one of the possible victim’s number of mentions in the cybersecurity related database texts exceeds the threshold limit per day, the software system adds them to the table too like the following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



## 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 with a problem at that step. More than a half cybersecurity related terms have no a 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 important 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 returns 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 not useful data in the json too, so we filter them and do not write in our database. We filter 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 firstly 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 a 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.

## Cybersecuirty Related Event Detection

From the previous steps of the software system, we get the posssible 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 that events we prapered 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 configuration file to change tracked entities.

**CHAPTER 5**

# IMPLEMENTATION

**CHAPTER 6**

# RESULTS

**CHAPTER 7**

# CONCLUSION AND FUTURE WORK

## Conclusion

## Future Work

# 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/

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

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)