

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

SEPTEMBER 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**

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, **Cognitive Science Dept.**

Examining Committee Members:

Asst. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science Dept., METU

Assoc. Prof. Dr. Hacer Karacan
Computer Engineering Dept., Gazi University

Date: **10.09.2019**

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

September 2019, 99 pages

Cybersecurity experts scan the internet and face security events that influence users, institutions, and governments. An information security analyst regularly examines sources to stay up to date on security events in her/his domain of expertise. This may lead to a heavy workload for the information analysts if they do not have proper tools for security event investigation. For example, an information analyst may want to stay aware of cybersecurity events, such as a DDoS (Distributed Denial of Service) 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 event. Therefore, information security analysts need to establish and keep situational awareness active about 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 timely. There have been attempts to solve this problem both from an academic perspective and engineering purposes.

A recent challenge in this domain is that the internet community use different languages to share information. For instance, information about security events in Turkey is mostly shared on the internet 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: Doç. Dr. Cengiz Acartürk

Eylül 2019, 99 sayfa

Siber güvenlik uzmanları interneti taramakta ve kullanıcıları, kurumları ve hükümetleri etkileyen güvenlik olaylarıyla karşı karşıya kalmaktalar. Bir bilgi güvenliği analisti, kendi uzmanlık alanındaki güvenlik olaylarından haberdar olmak için kaynakları düzenli olarak inceler. Bu incelemeler, güvenlik olayı incelemesi için uygun araçları yoksa, bilgi analistleri için ağır bir iş yüküne yol açabilir. Örneğin, bir bilgi analisti, bir devlet kurumu web sitesine yapılan DDoS (Dağıtılmış Hizmet Reddi) saldırısı gibi siber güvenlik olaylarından haberdar kalmak isteyebilir. Tehditleri ne kadar erken saptar ve anlarsa, engeli azaltmak ve gerçekleşen olayı araştırmak için o kadar uzun süresi kalır. Bu nedenle, bilgi güvenliği analistlerinin, güvenlik olayları ve muhtemel etkileri hakkında durumsal farkındalıklarını oluşturmaları ve aktif tutmaları gereklidir. Bununla birlikte, büyük miktarda bilgi akışı nedeniyle, güvenlik analistlerinin ve araştırmacıların güvenlik olaylarını zamanında tespit etmesi ve analiz etmesi zor olabilir. Bu sorunu hem akademik açıdan hem de mühendislik odaklı çözme girişimleri bulunmaktadır.

Bu alandaki son zorluk, internet camiasının bilgi paylaşmak için farklı diller kullanmasıdır. Örneğin, Türkiye ile ilgisi bulunan güvenlik olaylarına ait bilgiler

internette çoğunlukla Türkçe olarak paylaşılmaktadır. Bu tez, Twitter ve haber medya kaynaklarındaki Türkçe metinleri işleyerek güvenlik olaylarının otomatik olarak tespit edilmesini araştırmaktadır. Siber güvenlik olaylarını gerçek zamanlı olarak tespit edebilen otomatik, Türk diline özgü bir yazılım sistemi önermektedir.

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

To My Family

ACKNOWLEDGMENTS

I would like to wholeheartedly thank my advisor Assoc. Prof. Dr. Cengiz Acartürk for his constant support, endless patience, valuable advice, guidance, continuous encouragement, constructive criticisms, and friendship. It was a great honor to work with him for the last two years and our cooperation influenced my academical life highly.

I would also like to sincerely thank Cyber Security Department's faculty members and staff for their help. It's been a great experience. I never felt alone and whenever I needed help or advice I always found it. I also want to thank my colleagues supported and encouraged me to accomplish my master's degree successfully. I especially want to thank one of my colleagues, Burak Çelik. He shared his knowledge anytime I need during my master research and be a good fellow for me. I am truly grateful for his friendship.

Finally, and most importantly, I would like to thank my family, my dear father Mustafa Ural, my dear mother Ayşe Ural and my dear sister Özlem Ural Fatihoglu for their unconditional love. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

ABSTRACT.....	v
ACKNOWLEDGMENTS	x
TABLE OF CONTENTS.....	xi
LIST OF TABLES.....	xiv
LIST OF FIGURES.....	xv
LIST OF ABBREVIATIONS	xvii
1 INTRODUCTION	18
1.1 Motivation	18
1.2 Research Question and Objectives	21
1.3 Use Cases	22
1.4 Routine Tasks of an Information Security Analyst	23
1.5 Outline.....	24
2 LITERATURE REVIEW	26
2.1 Researches on Identifying Victims Affected by Cybersecurity Attacks.	26
2.2 Cybersecurity Event Forecasting	28
2.3 Drive-by Download Attack Prediction.....	29
2.4 Cyberattack Detection using Social Media.....	29
3 SYSTEM ARCHITECTURE AND DESIGN AND METHODOLOGY	34
3.1 Approach	34
3.2 Determination of Cybersecurity Related Keywords Vector.....	38
3.3 Data Collection	43
3.3.1 Twitter Social Network as a Data Source.....	44
3.3.2 Hürriyet Turkish Newspaper as a Data Source	47

3.4	Data Preprocessing.....	48
3.5	Data Processing.....	49
3.5.1	Natural Language Processing.....	49
3.5.2	Istanbul Technical University NLP API	50
3.5.3	Text Mining.....	52
3.6	Cybersecurity Related Event Detection.....	52
4	IMPLEMENTATION.....	54
4.1	Multi-Process Architecture	54
4.1.1	Twitter API Stream to Database Process	54
4.1.2	Hurriyet API Stream to Database Process.....	56
4.1.3	ITU NLP API Normalization Process.....	58
4.1.4	Security Events Web Portal Process.....	58
4.2	Microservice Architecture	58
4.3	Database Architecture of the System	59
4.3.1	Source Column of the Database	59
4.3.2	Date Column of the Database.....	59
4.3.3	UserName Column of the Database	59
4.3.4	Title Column of the Database.....	60
4.3.5	Text Column of the Database.....	60
4.3.6	Status Column of the Database	60
4.4	User Interface of the System.....	60
4.5	Other Technologies Used in the Thesis Study	61
4.6	Summary of the Implementation Chapter	62
5	RESULTS.....	64
5.1	Historical Cybersecurity Event Detection Test with an Independent Dataset - Nic.tr DDOS Attack.....	65

5.2	Successful Cybersecurity Event Detection Samples.....	68
5.2.1	WhatsApp Spyware Attack.....	68
5.2.2	Vulnerabilities in Remote Patient Tracking System Applications..	69
5.2.3	Other Successful Detection Examples.....	70
5.3	Unsuccessful Cybersecurity Event Detection Samples.....	75
5.3.1	Sample False Positive Cybersecurity Event Detection	75
5.3.2	Sample not Useful Cybersecurity Event Detection.....	76
5.4	Evaluation of the Results.....	76
6	CONCLUSION AND FUTURE WORK	78
6.1	Conclusion.....	78
6.2	Future Work.....	79
	REFERENCES	80
	APPENDICES	84
	APPENDIX A.....	84
	manager.py File	84
	config.py File	84
	hurriyetApiToDb.py File	85
	ituNlpPipeline.py File	86
	securityEventsWebPortal.py File.....	86
	sqliteOperations.py File	87
	twitterStreamToDb.py File.....	88
	userInterface.html File	88
	APPENDIX B.....	89

LIST OF TABLES

Table 1: Example high-confidence events extracted using the system published within this paper.....	26
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.....	27
Table 3: Seed Instances for DDoS Attacks.	27
Table 4: Tweet Examples with Attack Targets.	28
Table 5: Sample Table after the Analyze.....	37
Table 6: Sample Information Sharing Plan Table.....	37
Table 7: Sample Detected Events Table.....	37

LIST OF FIGURES

Figure 1 Tweets in Turkish After the Turktrust Vulnerability Announcement on 3 January 2013. Retrieved June 28, 2019, from https://twitter.com/	18
Figure 2 Hürriyet Newspaper News after the Turktrust SSL Vulnerability is Detected. Retrieved June 28, 2019, from http://www.hurriyet.com.tr/teknoloji/yanlis-sertifika-googledan-dondu-22290509 . Copyright 2013 by Hürriyet Gazetecilik ve Matbaacılık A.Ş.	20
Figure 3 Research results of IBM Security Lab about Cyber Security Analysts	24
Figure 4 Architecture of the Keyword Finder Component.....	30
Figure 5 Technical Overview of Sonar.	30
Figure 6 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.....	31
Figure 7 A Schematic Overview of Cybersecurity Event Detection System from The Publication.....	31
Figure 8 The General Overview of the System.....	35
Figure 9 Normalizer Sample from ITU NLP API.....	36
Figure 10 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends Results before Nic.tr Attack Start Day. Retrieved September 7, 2019, from https://voyant-tools.org	39
Figure 11 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends Results at Nic.tr Attack Start Day. Retrieved September 7, 2019, from https://voyant-tools.org	40
Figure 12 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends after Nic.tr Attack Start Day (2 weeks period). Retrieved September 7, 2019, from https://voyant-tools.org	41
Figure 13 Sinclair, S. G. Rockwell. (2019). Cirrus after Nic.tr Attack Start Day (2 weeks period). Retrieved September 7, 2019, from https://voyant-tools.org . 42	42
Figure 14 Sample Turkish Tweets Related with a Security Incident. Retrieved June 28, 2019, from https://twitter.com/	45
Figure 15 A Simple diagram to explain what NLP does.	49

Figure 16 Normalization Architecture	51
Figure 17 Normalization Samples for Complex Error Types.....	51
Figure 18 Sample Stored Twitter Data in the Database File.....	55
Figure 19 Sample Stored Hurriyet Newspaper Data in the Database File.....	57
Figure 20 User Interface of the Cybersecurity Detection Software	64
Figure 21 Detected Security Events Before the Nic.tr Attack (2014-2015).....	65
Figure 22 Nic.tr Attack Start Day Detected Security Events Samples.....	66
Figure 23 Detected Security Events Samples between 14 and 28 December 2015	67
Figure 24 WhatsApp Spyware Attack Detection.....	69
Figure 25 STM Warns about Remote Patient Tracking System Applications...	70
Figure 26 Supply Chain Attack Targets ASUS Computers Through Backdoored Update.....	70
Figure 27 MuddyWater Attack.....	71
Figure 28 MegaCortex Ransomware	71
Figure 29 Using Popular TV Shows to Spread Malware.....	71
Figure 30 RTM Banking Trojan	72
Figure 31 Mother's Day Cyberattack Alerts.....	72
Figure 32 Valentine's Day Phishing Attacks	72
Figure 33 Pirate Matryoshka Virus	72
Figure 34 Cyber Attacks Using Smart TVs.....	72
Figure 35 Smart Home Cyberattacks.....	73
Figure 36 Valletta Bank Cyberattack	73
Figure 37 Business Email Compromise Attack	73
Figure 38 Cyberattacks Which Use Netflix Brand.....	73
Figure 39 Cyberattacks Targeting Job Seekers.....	74
Figure 40 Phishing Attacks Targeting Social Media Users	74
Figure 41 Zombie Cookies.....	74
Figure 42 Millions of Email Addresses Infiltrated the Internet.....	75
Figure 43 The USA and China Cyberwar Started Tweet	75
Figure 44 Angela Merkel and Hundreds of German Politicians Hacked	75
Figure 45 Sample False Positive Cybersecurity Event Detection.....	76
Figure 46 Sample, not Useful Cybersecurity Event Detection	76

LIST OF ABBREVIATIONS

DDoS	Distributed Denial of Service
DoS	Denial of Service
TF-IDF	Term Frequency-Inverse Document Frequency
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
NER	Named Entity Recognition
NLP	Natural Language Processing

CHAPTER 1

INTRODUCTION

1.1 Motivation

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.



Figure 1 Tweets in Turkish After the Turktrust Vulnerability Announcement on 3 January 2013. Retrieved June 28, 2019, from <https://twitter.com>.

Since Turktrust certificates were a significant part of certificate use market in Turkey, numerous Tweets circulated in Turkish related to the vulnerability.

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.

Twitter is a major source of up to date information. 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 reveal that most of the research conducted on security event¹ detection has been developed for analyzing the text in English. As of our knowledge, research is lacking on real-time security event detection in Turkish language streams.

Given the significant share of the use of the Turkish language on the Internet, it is necessary to develop security event detection tools that process Turkish data. 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," 2019) With emerging internet adoption in Turkey, there are much timely information shared in Turkish. Event detection systems

¹ According to advisera.com website information security event refers to "something that can affect risk levels, without necessarily impacting the business or information." On the other hand, information security incident refers to "something that in fact negatively affected the business or information which should be protected." ("Differences between a security event vs incident vs non-compliance," 2018) The present thesis, we use event term, because it is more suitable for our study.

which developed for English texts are not useful for Turkish texts mining. Therefore, in order to use Turkish texts at detection of cybersecurity 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,” 2013)



Figure 2 Hürriyet Newspaper News after the Turktrust SSL Vulnerability is Detected.
Retrieved June 28, 2019, from <http://www.hurriyet.com.tr/teknoloji/yanlis-sertifika-google-dan-dondu-22290509>. Copyright 2013 by Hürriyet Gazetecilik ve Matbaacılık A.Ş.

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 LinkedIn², Facebook³, Eksisozluk⁴ website streams to gain more accurate results.

1.2 Research Question and Objectives

The objective of this thesis is to develop a security event detection tool for processing Turkish data. Current cybersecurity event detection tools are developed for extracting 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.

For this, we reviewed the state of the art studies and software systems in real-time event detection, as reported in the literature review chapter. We then investigated potential data sources to determine the most suitable ones to use it for real-time event detection with Turkish-text. We investigated methodologies and API's related to NLP (Natural Language Processing) to use it for normalization⁵ of Turkish texts.

² www.linkedin.com

³ www.facebook.com

⁴ www.eksisozluk.com

⁵ “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,” 2015)

We designed and developed a software system for real-time cybersecurity event detection using Turkish texts. We designed 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 content of information to be extracted from online sources.

We validated the proposed approach using several detected events already shared in Turkish-in online platforms.

1.3 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 Eksisozluk 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.

1.4 Routine Tasks of an Information Security Analyst

According to the careerexplorer.com website (“What does an information security analyst do? - CareerExplorer,” 2019), an information security analyst 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 systems⁶. One of their responsibility is to analyze data and to recommend changes to managers. 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 and 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,” 2019) 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.

⁶ <https://www.careerexplorer.com/careers/information-security-analyst/>

In Figure 3 (Borrett, 2017), 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

A manual investigation of security events is not sustainable without automation. To make it sustainable, automated NLP analysis tools and Text mining methods need to be used.

1.5 Outline

Chapter 2 presents the relevant literature. In Chapter 3, we introduce the software system in terms of its architectural and design perspectives. In Chapter 4, we present the software system in terms of its implementation and evaluation perspectives. In Chapter 5, we discuss thesis results. Finally, in Chapter 6, we present the conclusion and propose possible future work.

CHAPTER 2

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.

2.1 Researches on Identifying Victims Affected by Cybersecurity Attacks

On this research field, one of paper is “Weakly Supervised Extraction of Computer Security Events from Twitter”. It is a 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.

Victim	Date	Category	Sample Tweet
namecheap	Feb-20-2014	DDoS	My site was down due to a DDoS attack on NameCheap's DNS server. Those are lost page hits man...
bitcoin	Feb-12-2014	DDoS	Bitcoin value dramatically drops as massive #DDOS attack is waged on #Bitcoin http://t.co/YdoygOGmhv
europe	Feb-20-2014	DDoS	Record-breaking DDoS attack in Europe hits 400Gbps.
barcelona	Feb-18-2014	Account Hijacking	Lmao, the official Barcelona account has been hacked.
adam	Feb-16-2014	Account Hijacking	@adamlambert You've been hacked Adam! Argh!
dubai	Feb-09-2014	Account Hijacking	Dubai police twitter account just got hacked!
maryland	Feb-20-2014	Data Breach	SSNs Compromised in University of Maryland Data Breach: https://t.co/j69VeJC4dw
kickstarter	Feb-15-2014	Data Breach	I suspect my card was compromised because of the Kickstarter breach. It's a card I don't use often but have used for things like that.
tesco	Feb-14-2014	Data Breach	@directhex @Tesco thanks to the data breach yesterday it's clear no-one in Tesco does their sysadmin housekeeping!

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.

Feature Category	Sample Feature	Event Category
keyword-context-left	security breach	Data Breach
victim-context-right	X admits	Data Breach
victim-context-right	X data breach affecting	Data Breach
keyword-context-both	DT ddos attack	DDoS
keyword-context-left	NNP NNP hit IN ddos	DDoS
victim-context-right	X getting ddos'd	DDoS
victim-context-both	PRP hacked X POS account	Account Hijacking
keyword-context-left	POS email was hacked	Account Hijacking
victim-context-right	X POS NNP account	Account Hijacking

Then they are aiming at finding the victim, institution, or program affected by these events.

Table 3: Seed Instances for DDoS Attacks.

Victim	Date	# Mentions
spamhaus	2013/03/18	26
soca	2011/06/20	89
etrade	2012/01/05	76
interpol	2012/02/29	45
ustream	2012/05/09	911
virgin media	2012/05/08	15
pirate bay	2012/05/16	2,265
demonoid	2012/07/27	182
att	2012/08/15	2,743
sweden	2012/09/03	28
godaddy	2012/09/10	849
github	2013/07/29	102
reddit	2013/04/19	2,042
cia	2011/06/15	36
paypal	2010/12/10	57

In this study, 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.

Another example research is “Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter as an Example”. In that paper(Aslan, Sağlam, & Li, 2018), they use 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.

2.2 Cybersecurity Event Forecasting

One of example research in this research filed is “DDoS Event Forecasting Using Twitter Data”. (Wang & Zhang, 2017) It is a publication to estimate the DDoS attacks that have not yet taken place by processing Twitter data.

Table 4: Tweet Examples with Attack Targets.

Post	Target
The Thai government is about to end their citizens Internet freedom.	Thai Government
soooooooooo basically they are all leaving soon, sony is mad as hell.	SONY

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, 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.

2.3 Drive-by Download Attack Prediction

Cyber attackers may use the URL abbreviation method to show malicious websites as if a harmless website and share them on twitter as an abbreviated URL. Twitter users may believe in this deception and click on such website abbreviations, and these links can harm the users. “Prediction of Drive-by Download Attacks on Twitter” is an example which researches this field. (Javed, Burnap, & Rana, 2019) 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.

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.

2.4 Cyberattack Detection using Social Media

A sample study on this field is “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 follow related to cybersecurity with the help of previously given keywords.

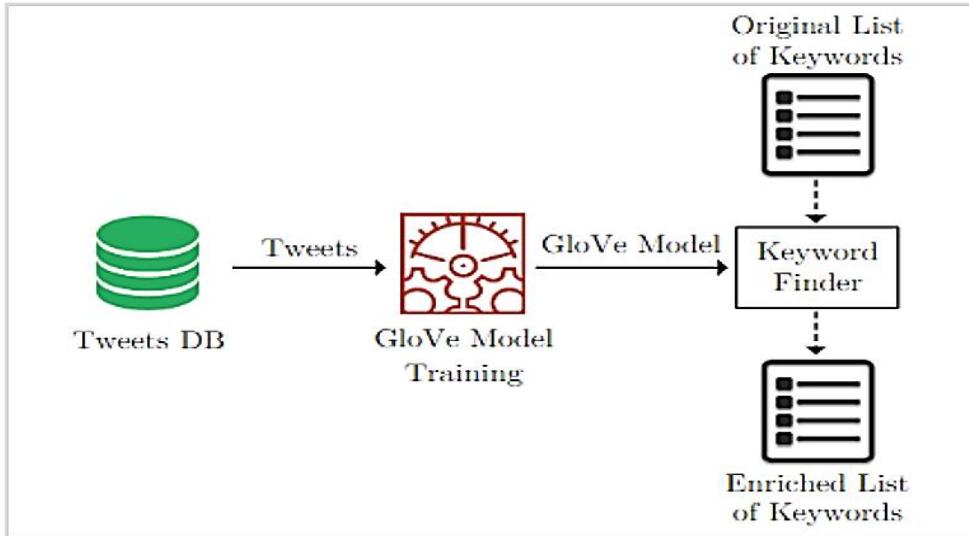


Figure 4 Architecture of the Keyword Finder Component.

They have also benefited from big data technologies.

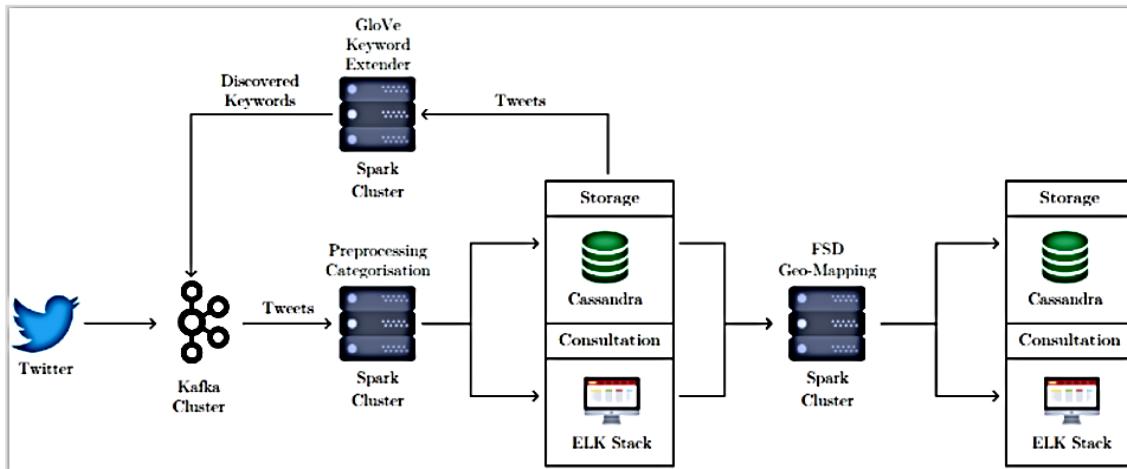


Figure 5 Technical Overview of Sonar.

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

Another example is “Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media”. (Khandpur et al., 2017) It is another study on detecting cybersecurity attacks by processing Twitter data. They acknowledge that their work is like that of previous studies, but they report more successful results.

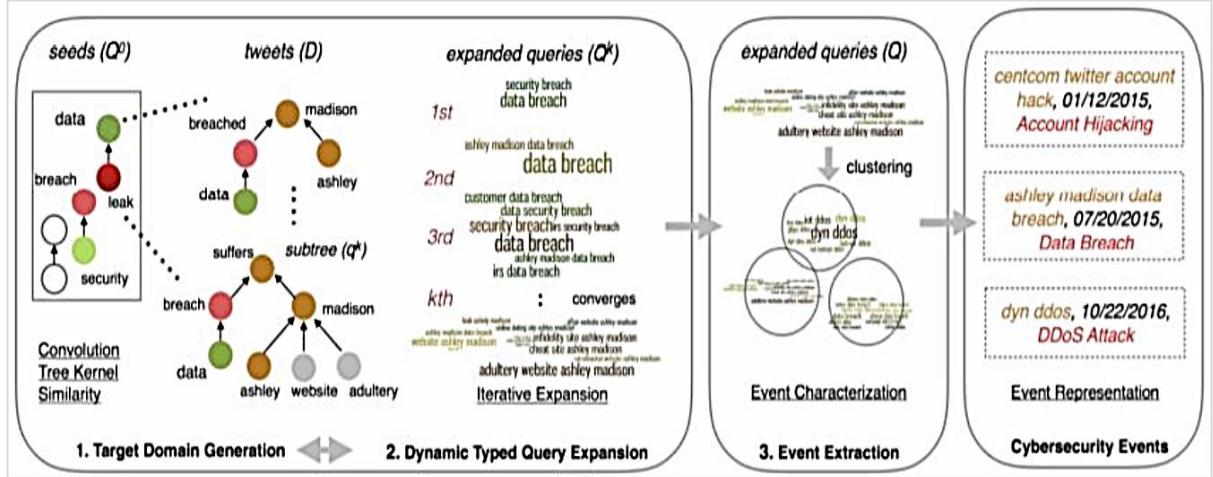


Figure 7 A Schematic Overview of Cybersecurity Event Detection System from The Publication.

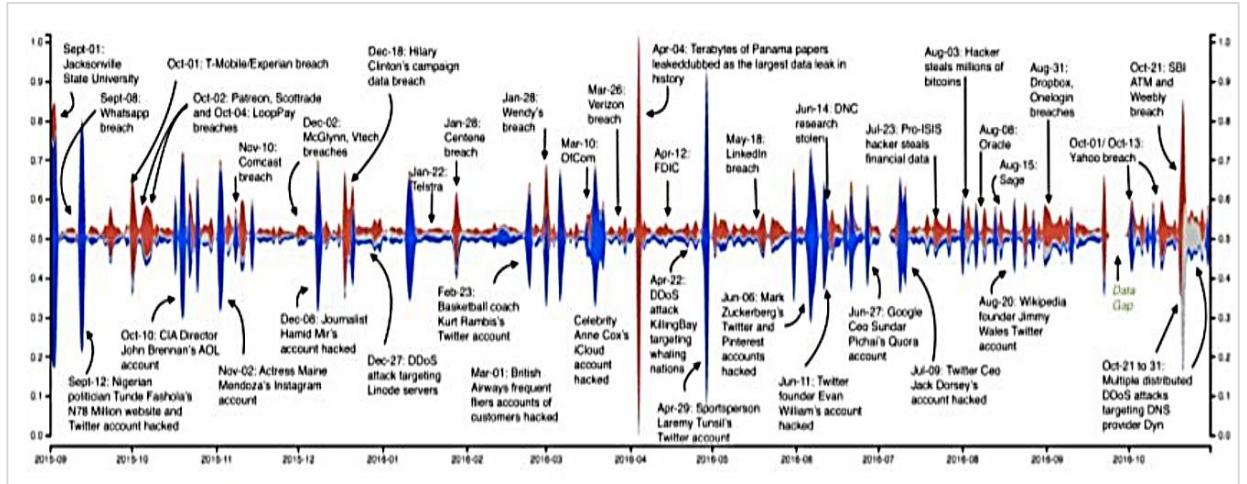


Figure 6 Streamgraph Showing Normalized Volume of Tweets (September 2015 through October 2016) Tagged with Data Breach (red), DDoS Activity (grey) and Account Hijacking (blue) Types of Cybersecurity Events.

This research is one of the state-of-art projects in the cyber attack detection domain. 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 cyberattacks, we investigate detecting cyber events in the present thesis.

CHAPTER 3

SYSTEM ARCHITECTURE AND DESIGN AND METHODOLOGY

In this chapter, we explain the software system's architecture and design and methodology. Firstly, we explain the general approach. Then we present data collection using Standard Twitter API, Twitter Premium API, Hurriyet API, and Selenium. After that we mention how we can preprocess and process the data. Then we present how we detect a cybersecurity event with using anomaly detection which is one of the machine learning techniques.

3.1 Approach

Figure 9 presents a general overview of the architecture and design. First, we need real-time streaming data to process. In order to establish a Twitter stream connection, the software uses statically defined the configuration file values. To gathering the data in real-time, we use Standard Twitter API. We create cybersecurity-related Turkish keyword vector with using Term Frequency - Inverse Term Frequency analysis of past security incidents. (See Appendix B) We explain the details in Section 3.2 determination of cybersecurity-related keywords vector. We use this keyword vector to gather useful Twitter stream and Hurriyet Newspaper stream for our research. We use the language filter feature of the Twitter API in order to fetch only the Turkish Tweets. Hurriyet is a Turkish newspaper, therefore we did not need a language filter for it. To establish the Hurriyet Newspaper stream connection, the software also uses the configuration file.

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.

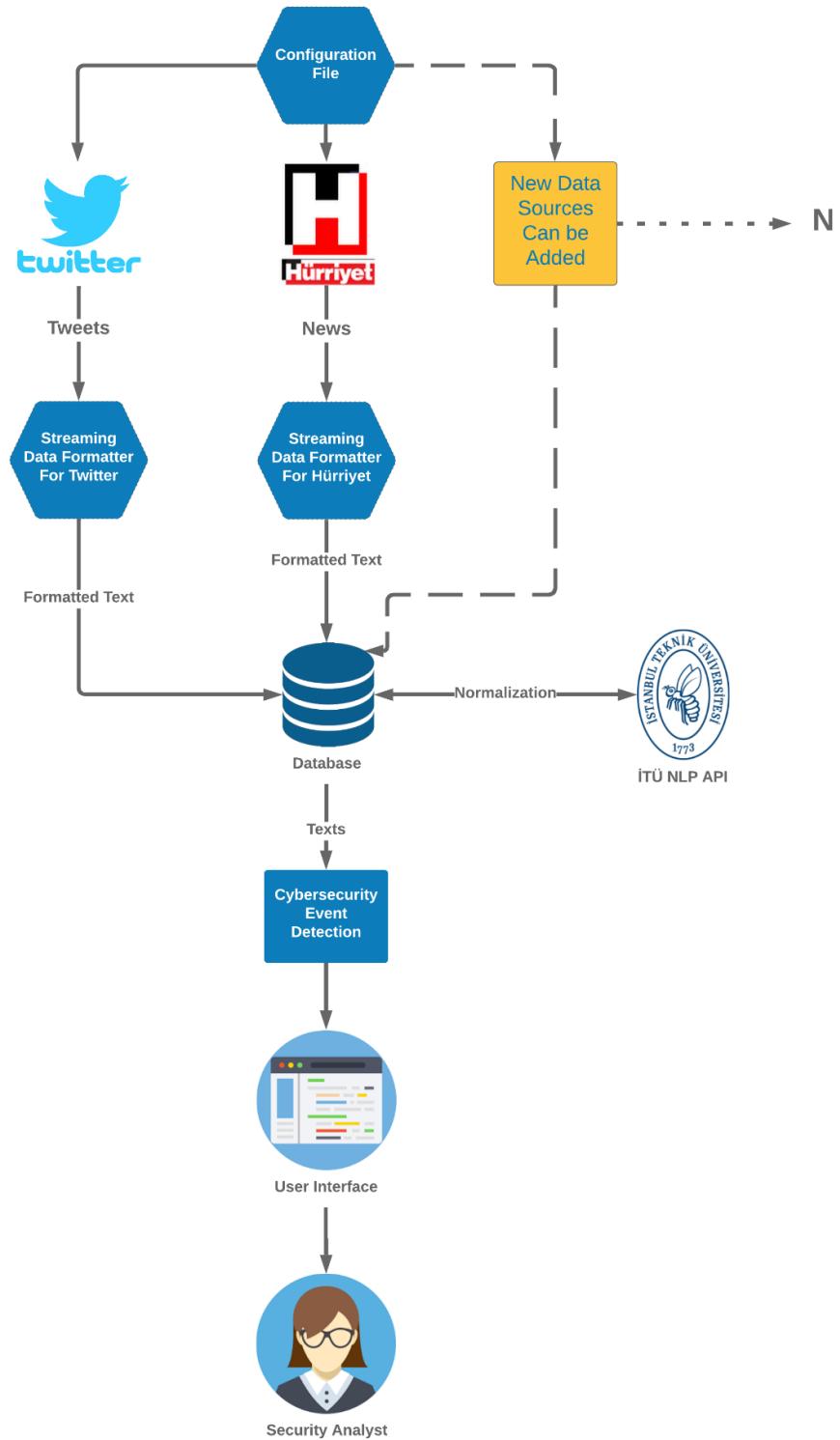


Figure 8 The General Overview of the System

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



Figure 9 Normalizer Sample from ITU NLP API

After the normalization step, we move forward to Named Entity Recognition⁷ 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. (See Appendix B)

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 Table 5. The period of the table is one day. Every new day, the number of mentions per victim is set and start with zero.

⁷ 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”, 2019)

Table 5: Sample Table after the Analyze

victim	Date	# Mentions
twitter	24.04.2019	16
x newspaper	24.04.2019	700
y newspaper	24.04.2019	6

We add daily thresholds. (i.e. 10 per day) If the number of mentions is more than the thresholds value, we will share this detected cybersecurity event within the user interface.

Table 6: Sample Information Sharing Plan Table

Date: 24.04.2019		
Entity	Representative Tweet	Count
x newspaper	Tweet: X newspaper is hacked.	700

The software repeatedly 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 in Table 7.

Table 7: Sample Detected Events Table

Date: 24.04.2019		
Entity	Representative Tweet	Count
X newspaper	Tweet: X newspaper is hacked.	700
y newspaper	Tweet: y newspaper cannot be accessed.	323
z website	Tweet: Z website has a malicious file.	312

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

3.2 Determination of Cybersecurity Related Keywords Vector

To create an optimum version of cybersecurity-related keyword vector, we used term frequency-inverse document frequency(TF-IDF) technique, keyword-based analysis, the statistical technique, and A/B testing.

Even if the Tweets or the news are in the Turkish language, there are widespread English cybersecurity terms used in Turkish texts. Therefore we create the vector using both English and Turkish keywords.

It is a numerical statistic that intends to reflect importance a keyword or phrase is for within a document or a Web page in a corpus or in a collection. TF-IDF is the product of two statistics, frequency of term and inverse frequency of the document. There are several ways to determine the exact values of both statistics. (Rajaraman, A.; Ullman, 2011)

In the case of the term frequency $tf(t, d)$, the simplest choice is to use the raw count of a term in a document or the frequency with which the term t occurs in document d . If we denote the raw number for ft, d , the simplest tf scheme is $tf(t, d) = f_{t, d}$.

The inverse document frequency(IDF) is a measure of the amount of information provided by the word, that is, whether it is common or rare in all documents. This is the inverse fraction of documents containing the word with logarithmic scaling.

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

N: total number of documents in the corpus $N = |D|$

$|\{d \in D : t \in d\}|$: number of documents where the term t appears (i.e., $tf(t, d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

Then Term frequency–Inverse document frequency(tf–idf) is calculated as

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

In order to identify our cybersecurity-related keywords vector, we used the Term frequency–Inverse document frequency technique. Firstly, we find one of the past important cybersecurity events related to Turkey from history. We select “nic.tr DDOS attack” as the cybersecurity event. Then we create three different training database related to this attack with using Twitter Premium API. We select “nic.tr” as keyword and filter only the Turkish Tweets. With using tf-idf technique, we identify the most important words in these databases. Then we select the cybersecurity-related ones from the results of the tf-idf technique. And then add them to the cybersecurity-related keyword vector.

The first query includes tweets containing nic.tr keyword at the dates between 10.12.2014 and 13.12.2015. These dates are the one year period of time before the nic.tr attack. Then we do term frequency-inverse document frequency (TF-IDF) analysis. TF-IDF is a numerical statistic used in information retrieval to represent how important a specific word or phrase is to a given document. For doing such an analysis, we used Voyent Tools⁸. Figure 10 exemplifies the analysis of identifying important words with using the Tweet database which created before the nic.tr attack.

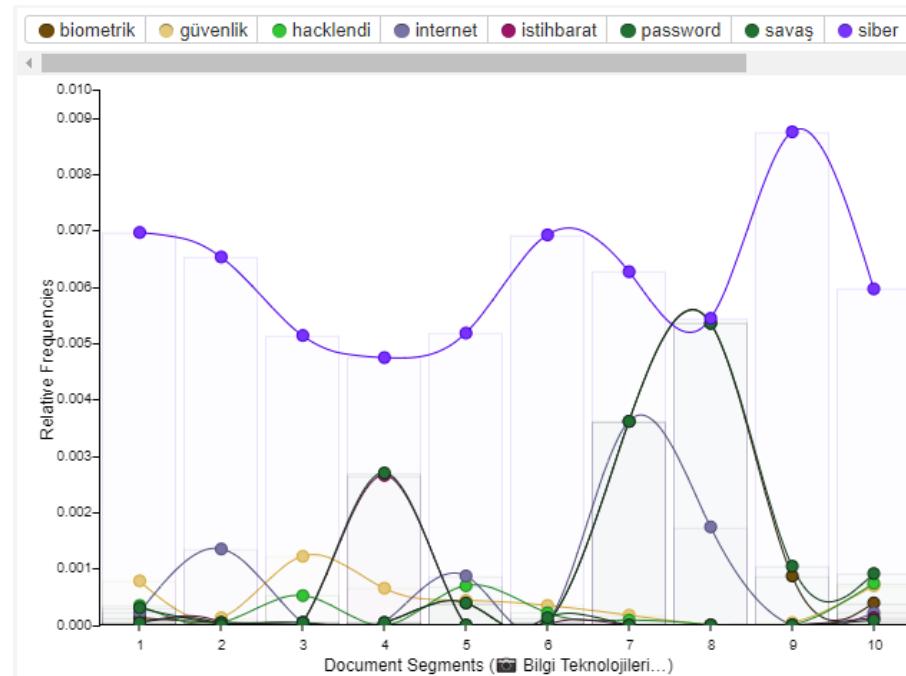


Figure 10 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends Results before Nic.tr Attack Start Day. Retrieved September 7, 2019, from <https://voyant-tools.org>.

⁸ <https://voyant-tools.org/>

Then we create another training database. That time we select the Tweets only at the day of the nic.tr attack on 14 December 2015. We analyze the tweets in the database with TF-IDF frequency analysis and do A/B test to select words from them to our cybersecurity-related keyword vector.

Figure 11 exemplifies the analysis of identifying important words with using the Tweet database which created at the nic.tr attack start day. Some of the important words identified by the analysis are “DDoS”, “fidye”, “hacklendi”, “nic.tr”, “saldırısı”, “siber”, “virüs”, as you can see in Figure below.

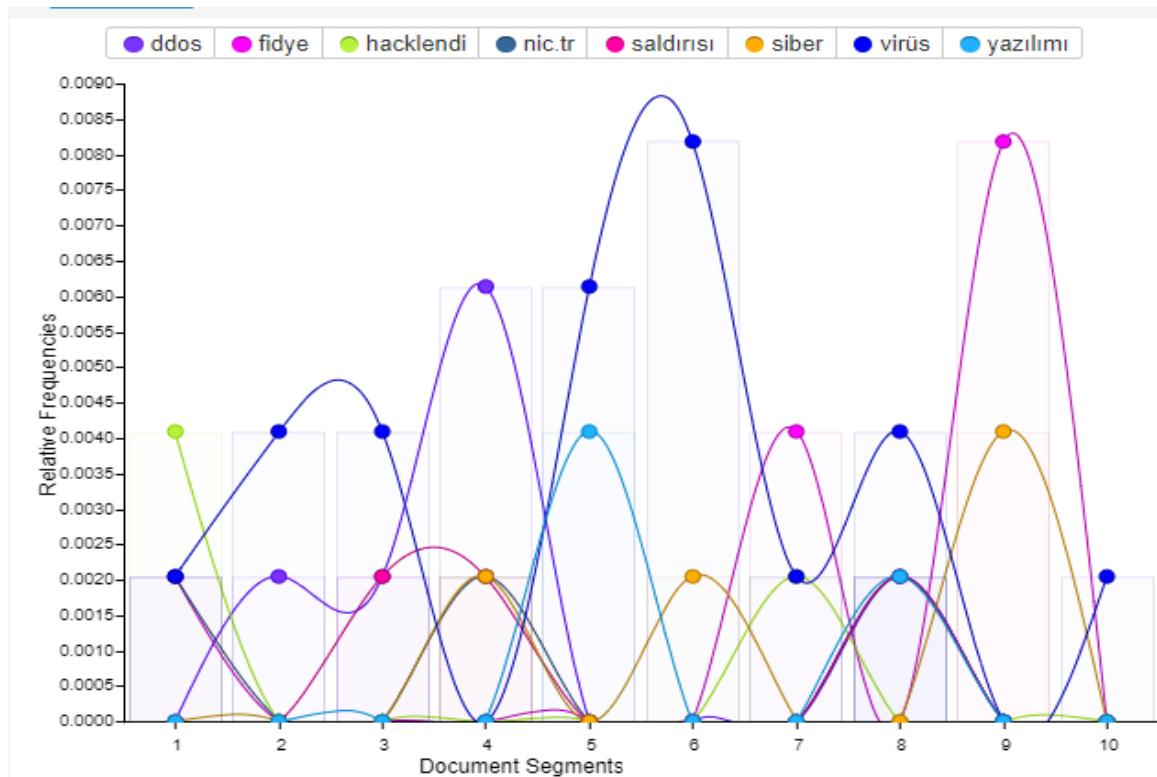


Figure 11 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends Results at Nic.tr Attack Start Day. Retrieved September 7, 2019, from <https://voyant-tools.org>.

Lastly, we create another training database. It includes the Tweets between 14 December 2015 and 28 December 2015. Within two weeks period of time, nearly 1000 Tweet had been tweeted related with "nic.tr".

Figure 12 exemplifies the analysis of identifying important words with using the Tweet database which created after the nic.tr attack start day. Some of the important words identified by the analysis are “saldırı”, “saldırıları”, “saldırısı”, “DDoS”, “hacker”, “hacklendi” as you can see in Figure 12.

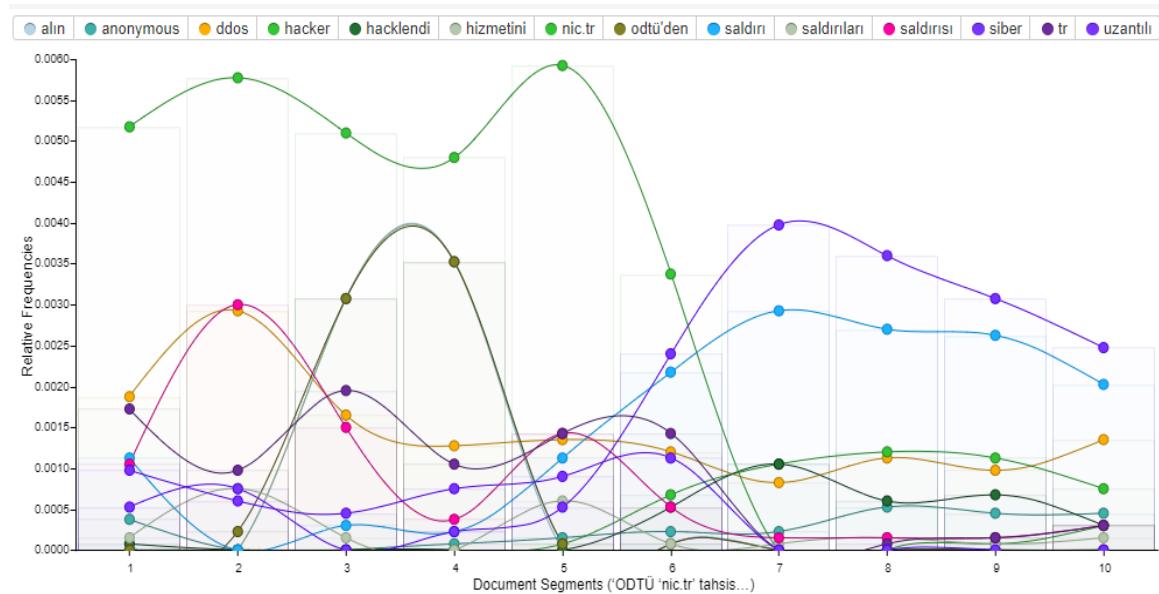


Figure 12 Sinclair, S. G. Rockwell. (2019). TF-IDF Words Trends after Nic.tr Attack Start Day (2 weeks period). Retrieved September 7, 2019, from <https://voyant-tools.org>.

Figure 13 shows Cirrus⁹ after nic.tr DDOS Attack Start Day (2 weeks period).

⁹ Cirrus is a word cloud that visualizes the most common words in a corpus or document. The most common terms are centered and have the largest size. As the algorithm scans the list and continues to try to draw words as close to the center of the visualization as possible, it will also include small words in spaces left by larger words that do not fit tightly. (“Voyant Tools Help”, 2019)



Figure 13 Sinclair, S. G. Rockwell. (2019). Cirrus after Nic.tr Attack Start Day (2 weeks period). Retrieved September 7, 2019, from <https://voyant-tools.org>.

We analyze their results and create cybersecurity-related keyword lists for each one of them. Then we used these keywords lists for A/B testing¹⁰. We compare the results of the A/B test and update the elements in the keyword vector according to their success rate. For A/B test we used the number of false-positive cybersecurity event detection and number of cybersecurity event detection. If a keyword significantly increases the number of false-positive detection, we do not add it to our cybersecurity-related keyword vector. On the other hand, if a keyword does not affect so much the false positive detection but

¹⁰ The A/B test is a randomized experiment with two variants, A and B. It includes the application of the statistical hypothesis test or “two-sample hypothesis test” used in the field of statistics. The A/B test is a method of comparing two versions of the same variable and determining which of the two variants is more effective. (Pardot, 2012)

increases the number of detection we add it to our cybersecurity-related keyword vector list.

3.3 Data Collection

In order to collect data, we use Twitter and Hürriyet newspaper. 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. To train our solution algorithm, we need training data. Twitter's standard search API (free version) allows searches against a sampling of recent Tweets published in the past 7 days. However, we need Turkey related past security events datasets such as nic.tr DDOS attack¹¹ to train our data.

Firstly, we searched online to find the sample datasets with our desired constraints. We could find only Yıldız Teknik University Kemik Natural Language Processing Research Group website¹² which shares sample Twitter datasets in the Turkish language. We sent mail and mentioned about our thesis project and ask sample datasets from them. They accepted our request to send us Turkish Twitter dataset. However, we could not use their datasets to train our solution algorithm because their dataset is created with random keywords with a random time interval, and the dataset was not big enough to use as training data.

We needed Twitter dataset in Turkish which created with cyber-security related keywords and specific time intervals. Twitter's Premium API can be used to create such datasets because it allows searching within the full archive of Twitter. However, the API price is between 99 dollars to 1900 dollars monthly. For educational purposes, Twitter may give

¹¹ Turkey's official domain name servers had been under a Distributed Denial of Service attack in 2015.

¹² <http://www.kemik.yildiz.edu.tr/>

premium API for free to students. We had applied Twitter to request free Premium API usage. After the application one of Twitter staff reached us from e-mail and we share some of the thesis details like abstract and approach section of the thesis. The Twitter staff is convinced to support our research and give us Premium API access for free.

We used Twitter Premium API to create training datasets. However, the given Premium API has 50 request limits monthly. We used Twitter Premium API to create training datasets. However, the given Twitter Premium API has only 50 request limits monthly and we reached the monthly request limit in a short period of time and our research faced with another problem. To solve this problem, we had two option. Option one was upgrading our Premium API request limit with spending thousands of dollars money, and the second option was to try to find an alternative way to collect the past Twitter data. After researches, we realize that there is another option to collect Twitter data. Instead of using Twitter API, we implemented a Python code to parse the Twitter website with using Selenium automation tool and chrome browser web driver and, created our desired training datasets. The selenium solution for fetching Twitter data is not a known method and it is firstly had implemented by us.

In the subsections below, we shared the details of the methods we used to collect data.

3.3.1 Twitter Social Network as a Data Source

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, 2010), 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.

In Figure 14, sample Turkish Tweets related to a security incident is presented.



Figure 14 Sample Turkish Tweets Related with a Security Incident. Retrieved June 28, 2019, from <https://twitter.com/>.

The Twitter API is a set of URLs. The URLs can 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 Standard API allows users to get tweets which includes specific parameters. Moreover, the resulting stream can be filtered according to Tweet languages, geolocation and so on. However, this API cannot retrieve tweets older than seven days. It 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, which is approximately one million Tweets per day due to Standard Twitter API limitation. 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. We use Twitter Standard API for collecting the real-time Twitter stream.

Twitter's Premium API has more capabilities like accessing the Full-archive of Twitter data from as early as 2006. The API is sold with monthly prices according to allowable request limit of the API. For academic purposes, Twitter may give support for free access. After sharing our research details with documents, they gave us Premium API which allows 50 requests monthly request limit. We used Twitter's Premium API for generating training datasets.

At one point, the monthly 50 request limit of Twitter's Premium API blocked our research. Therefore we developed another method to create training datasets with python code with using Selenium Python Framework which automates browsers and chrome web driver. With Selenium, The software opens a chrome browser as if a normal user, writes the query specified in the code in the search box of the twitter, and parses the Html. The Twitter Html file is a dynamically creating web page. To overcome this problem, the software automatically scrolls down the page as it parses. This method gave us unlimited training dataset creation right without using any API.

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. Our second data source, Hurriyet Newspaper, is introduced in the following section.

3.3.2 Hürriyet Turkish Newspaper as a Data Source

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,” 2018)

We can 12,000 request per day in Hurriyet Newspaper API. Therefore, the keyword list is essential to get relevant data in the result streams.

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,” 2019)

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.

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. 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](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](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/'](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, 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 and the writers can be queried and used in applications.

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

3.4 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.

3.5 Data Processing

3.5.1 Natural Language Processing

NLP is “the ability of machines to understand and interpret human language the way it is written or spoken” (“Teachbot(Teaching Robot) Using Artificial Intelligence and Natural Language Processing,” 2017). In Figure 4 (“Overview of Artificial Intelligence; Role of NLP in Big Data - XenonStack Blog,” 2019), can be seen as a simple explanation of What NLP does. In the present thesis, we used a few NLP techniques and Istanbul Technical University’s NLP API (“ITU Turkish Natural Language Processing Web Interface,” 2019) for normalization of the texts.

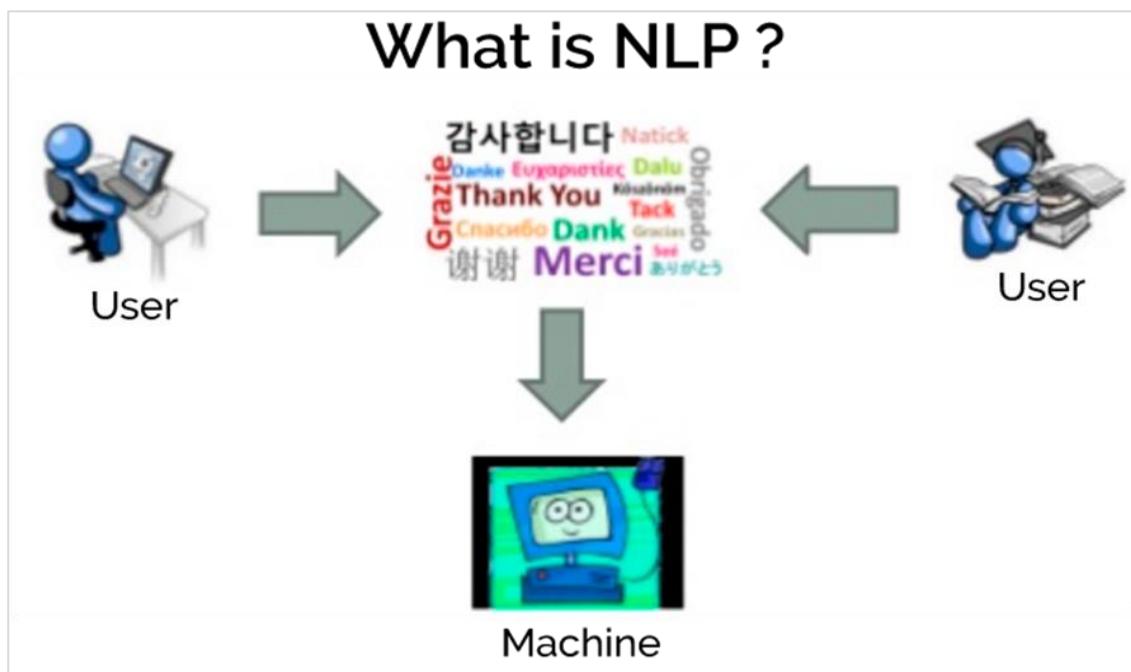


Figure 15 A Simple diagram to explain what NLP does.

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,” 2018)

3.5.2 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”, 2019) 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
- Vowelizers
- 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.

They divide the normalization tasks into two stages named IWD and CG. Input tokens are analyzed in order to select the ones to be normalized in the IWD stage. CG stage generates normalized output for its inputs coming from the prior stage.(Eryiğit & Torunoğlu-Selamet, 2017b)

In Figure 16, you can see the normalization architecture of the ITU NLP API.

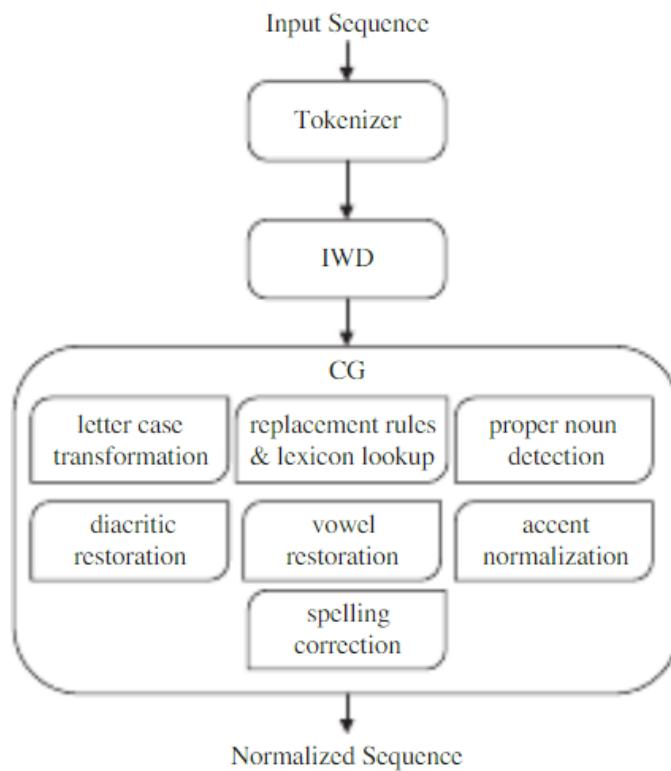


Figure 16 Normalization Architecture

Normalization samples for complex error types and their normalized outputs are as in the figure below.

Input	Normalized output	Modules needed
gdcem	gideceğim (I'll go)	Accent normalization + Vowel restoration
cagladan	Çağla'dan (from Çağla)	Diacritization + Proper noun detection
gitmicem	gitmeyeceğim (I won't go)	Accent normalization + Diacritization
Şekeeeeerim	şekerim (my sweetie)	Replacement rules + Diacritization

Figure 17 Normalization Samples for Complex Error Types

After fetching the data, during the processing and presentation steps, we store information in the database.

3.5.3 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,” 2016) 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,” 2016)

In this thesis, we used keyword-based analysis and statistical techniques. We use two keyword vectors for keyword-based analysis. One of the keyword vectors stores possible victims who are tracked by our software solution. The other keyword vector stores the possible useful cybersecurity-related Turkish terms such as “hacklendi” and “erişilemiyor”. We analyze the results by comparing the past frequency statistics and current results as described in the Approach section.

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.

3.6 Cybersecurity Related Event Detection

From the previous steps of the software system, we get the possible cybersecurity-related texts from different sources. Then preprocess and process 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. Then with using term frequency-inverse document frequency (TF

- IDF) technique, keyword-based analysis, the statistical technique, and A/B testing; we analyze past cybersecurity events and create cybersecurity-related keywords vector.

As we explained in Section 3.1 Approach section, we analyze real-time Turkish text data to detect cybersecurity events. In order to do this, we send requests to Twitter and Hurriyet newspaper with our cybersecurity-related keywords vector and we add Turkish language filter to our request. The possible victim vector of the solution periodically checked in the database in terms of the number of occurrences. If the number of occurrences of a victim shows anomaly¹³ according to its historical values, our solution detects them as a potential cybersecurity event and shows that events in the user interface portal.

¹³ Anomaly detection is a machine learning technique. In data mining, anomaly detection is the identification of rare events, items or observations which raise suspicions by differing significantly from most of the data. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions. (Zimek & Schubert, 2017)

CHAPTER 4

IMPLEMENTATION

4.1 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. These are Twitter API Stream to Database, Hurriyet API Stream to Database, ITU NLP API Normalization and Security Events Web Portal Processes.

4.1.1 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 is a sample screenshot of the database browser of SQLite which stores the gathered data. As you can see in the figure, source, date, username, text, and status columns are filled with data whereas the title column is empty for all rows, because Tweets has no title.

Database Structure Browse Data Edit Pragmas Execute SQL

Table: databaseTable

	Source	Date	UserName	Title	Text	Status
	Filter	Filter	Filter	Filter		Filter
1	twitter	2019-05-19	r4bisofi	—	RT @mkulunk: Liderimiz Sayın Erdoğan'ın yanında gelecek 1 asrı İnşaa etmeye devam edeceğiz. İçimizde hiç sönmeyen, Kuvai Milli Ruhu ile Mil...	2
2	twitter	2019-05-19	Forplay06	—	RT @filiz175: Habur Oslo Megri Megri diye zirvalayanlar, Devletimiz Terör Örgütüne silahlari gómün, teslim olun, işi kan dökmeden neticelens...	2
3	twitter	2019-05-19	LegendaryLego	—	RT @LegendaryLego: Karşınıza çıkacak "Canavar" kendiniz yaratınız Başakşehir!! Ağılamayın...	2
4	twitter	2019-05-19	veyselinedeniz	—	RT @HilmiKurnaz7: Bu edepsiz sabahtan akşamaya kadar Kürt sayfalarını kapatıp daha sonra sayfasında paylaşıyor girin sayfasına göreceksiniz 2...	2
5	twitter	2019-05-19	MuhammedLa...	—	RT @zaferates74: @medyaadami @NecatiOzkan @MuhammedLatifi17 veri kopyalama ve casus yazılım konusunda tecrübe oldukları belli, daha ö... 2	2
6	twitter	2019-05-19	qafasiguzel	—	RT @umitkocasakal_: Bunun sonucunda ülkemiz, küresel virus ve mikroplara açık hale gelerek yeniden emperyalizmin etkisi, saldırır ve kugatma...	2
7	twitter	2019-05-19	aysnrgl_gurbuz	—	Elime bilgisayarı almazken isim düşüncəe hacker bile olabilecek kadar ustalaşmama hayranım😊 2	2
8	twitter	2019-05-19	EnginAk68142...	—	@Ordinaryusbey Tepede tanına çirkef pislik Camiası Türk futbolunun içinde ki virus ☐ fetosaray 2	2
9	twitter	2019-05-19	sosyalismet1	—	@sedefecer sedef ecer. takıldı kaldı bir saatdir. virus mü var ne? 2	2
10	twitter	2019-05-19	ismailgezzer	—	Askerden sonra virus saldırısına uğradım sanırım 🙁 2	2
11	twitter	2019-05-19	hayalmarasli	—	@bhtypryzasan @Melissaaavci @kecirge @ozayprofil @eceminizz1 Ahahahahahaha 😂 @bhtypryzasan kanki sakin ol virus 😂😂😂 2	2
12	twitter	2019-05-19	Busezkssaa	—	bir gün beni siber suçlarından içeri alıcaklar ama bakalım ne zaman olucak bu 2	2
13	twitter	2019-05-19	ozlemalikilic	—	@ormerturantv72 Ömet bey inanamıyorum; gerçekten bunları siz mi yazınızı yoksa hesabınız mı hacklendi? 2	2
14	twitter	2019-05-19	enformedya	—	Siber düşmanlar iş arayanları güneşi kesiyor https://t.co/GzgiJp8Rpu https://t.co/1YNjTMyqe0 2	2
15	twitter	2019-05-19	Meraklimel...	—	@oztrk_aydn Resmen solcular sagolar ayrılmışlar ama solcuların içinde bir kara virus var.. aman orda kalsın . 2	2
16	twitter	2019-05-19	_selenga_	—	RT @zaferates74: Veri kopyalama ve casus yazılım konusunda tecrübe oldukları belli oluyor, Rahmetli gazeteci #TELATÇABUK yaşamış zaten.D... 2	2
17	twitter	2019-05-19	WelayetNews	—	Son yılda İran'a karşı 33 milyon siber saldırı yapıldıhttps://t.co/Fj1PZaTPhJ https://t.co/Zz4tCzKVKE 2	2
18	twitter	2019-05-19	GerceklerKara	—	RT @WelayetNews: Son yılda İran'a karşı 33 milyon siber saldırı yapıldıhttps://t.co/Fj1PZaTPhJ https://t.co/Zz4tCzKVKE 2	2
19	twitter	2019-05-19	benji37614541	—	@cirkinoglu bunlar içimizdeki virus gibidir sömürük yedikçe doymuyorlar taaki türkiyeyi mahvedene kadar 2	2
20	twitter	2019-05-19	merveeeeeee...	—	Siber terörist ve Siber suçlar adında bir makale yazıyorum. Hackerlar ve hacker gruplarından bahsetmek olmaz diyere... https://t.co/Y6mnzWAyxb 2	2
21	twitter	2019-05-19	gnayyksel	—	AKP'nin parali tweet silicileri(!)ş反正ında veya benim,hesabima gönderdikleri virus mesai yapıyor(!)Zira onları hi... https://t.co/pqNKSfmO7d 2	2
22	twitter	2019-05-19	just_a_bts	—	RT @ibtsblogs: hoseok şu illeti virus gibi yayıyor çocuklara namjoon'da ondan görüp yürekleri ağızlarına getiriyor, yakında fandomın leslerinin... 2	2
23	twitter	2019-05-19	bulgarsesmisiye	—	@blondesikhead AKP stepneliği bu hastalığı derinleştiriyor olmalı virus bir kere girmiş iлерiyor her daim vitesi artırarak. 2	2
24	twitter	2019-05-19	THT_Zentron	—	RT @Official_THT: Malware Analiz Alanında Eğitilecek Asistan Alımları Açılmıştır ! 2	2
25	twitter	2019-05-19	eceosmanoglu	—	Keşke hacker arkadaşım olsayı da herkesin instagramını patlatsaydık 2	2
26	twitter	2019-05-19	frkkko	—	@sellismi Virus gibisiniz ya😂 2	2
27	twitter	2019-05-19	jeoxygenss	—	tüm olayı izledim. resmen kızın üzerine insan salmışsin. buna siber zorbalık deniyor. başka yerlerde duyarlımış... —... https://t.co/fC7Rjp9Fqt 2	2
28	twitter	2019-05-19	Lakyuz	—	RT @OnlyAnkaragucu: Getirildiği günden itibaren hiçbir sorunu çözmemen, tek amacı ranta hizmet etmek olan, karaborsayı yasallaştıran, taraf... 2	2
29	twitter	2019-05-19	yunusemire041	—	hazari Sampiyon FavorisenDe virus 2	2
30	twitter	2019-05-19	guvencyy	—	@coachaykutt Hesabın mı hacklendi yoksa sen ciddimisin 2	2

Go to: 1

Figure 18 Sample Stored Twitter Data in the Database File

4.1.2 Hurriyet API Stream to Database Process

This process continually gathers Hurriyet API stream. Then preprocess the gathered data and write them to the database. Figure 13 is a sample screenshot of the database browser of SQLite which stores the gathered data. As you can see in the figure, source, date, title, text, and status columns are filled with data whereas username column is empty for all rows, because newspaper data has no username.

Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: databaseTable | New Record | Delete Record

Source	Date	erNar	Title	Text
Filter	Filter	Filter	Filter	Filter
1 hurriyet	2019-05-13	—	Kuantum bilgisayarlar siber güvenlikte neleri değiştirecek?	Bu yılın başında IBM tarafından piyasaya sürülen 'Q System One' ile ticari uygulanabilirliği ortaya çıkan kuantum hesaplama, siber güvenlikte şe
2 hurriyet	2019-05-12	—	Hem aydınlatıyor hem izlettiyor	TEKNOLOJİYLE birlikte hayatımızda güvenlik önlemleri almak oldukça kolaylaştı. Akıllı kilitlerden akıllı alarm sistemlerine kadar pek çok çözüm su
3 hurriyet	2019-05-10	—	Anneler Günü gelen tehlkiye aman dikkat!	Akıllı telefonlardan ve tabletlerden alışveriş yapmak, tüketicilere alışveriş merkezlerinde bulması zor olan o hediyeye, online alışveriş sitelerinde
4 hurriyet	2019-05-10	—	Dünya armatörlerinin ilk kadın başkanı dümene geçiyor	- TURMEPA Başkanı Şadan Kaptanoğlu, 14-15 Mayıs 2019'da Atina'da gerçekleştirilecek genel kurulda BİMCO'nun (Baltık ve Uluslararası Denizcilik k
5 hurriyet	2019-05-10	—	Jeopolitik siber saldırı sayısında sıçrama yaşandı	Üç aylık APT trend özette, Kaspersky Lab'ın özel tehdit istihbarat araştırmasından ve diğer kaynaklardan elde edilen Facebook, Çarşamba günü blockchain içeriğine ilişkin yaşasını kal
6 hurriyet	2019-05-09	—	Bitcoin 6 bin doların üzerine çıktı	Geçen ay kendi kripto para birimini piyasaya sürmeye hazırladığı iddia edilen Facebook, Çarşamba günü blockchain içeriğine ilişkin yaşasını kal
7 hurriyet	2019-05-09	—	Son dakika... DEAŞ'ın Çanakkale'ye saldırısı planı ortaya çıktı!	24-25 Nisan günü Çanakkale'de yapılan Anzak Günləri'nde terör saldırısı yapmayı planladığı değerlendirilen Suriye uyruklu DEAŞ'lı şahıs Abdulke
8 hurriyet	2019-05-09	—	Harmandalı oynayan robot Avrupa finalinerine taşıdı	"Harmandalı zeybeği" oynayan robot ile Türkiye birinciliğine ulaşan Manisa öğrenciler, aynı başarıyı Avrupa'ya da taşımak istiyor. Manisa Fen Lise
9 hurriyet	2019-05-09	—	Bilgisayar korsanları 41 milyon dolarlık bitcoin ele geçirdi	Üyeleri için Bitcoin ve diğer kripto para birimlerini depolayan Binance Exchange'in 41 milyon dolar değerindeki 7 bin bitcoini bilgisayar korsanları
10 hurriyet	2019-05-08	—	Şehit polis Muammer Ateş için Bayburt'ta hazır tören	Bayburt İl Emniyet Müdürlüğü önünde düzenlenen törenin ardından doğduğu köy olan Aydintepe ilçesine bağlı Çatıksu köyüne getirilen şehit Ateş
11 hurriyet	2019-05-07	—	Beş Bakanlık sanal tehditlere karşı harekete geçiyor	İşleri Bakanlığı koordinasyonunda, Aile Çalışma ve Sosyal Hizmetler, Gençlik ve Spor, Milli Eğitim ve Ulaştırma ve Altyapı Bakanlıklar, çocukların sa
12 hurriyet	2019-05-07	—	S-400 speküasyonlarına yanıt: Tarihizimi ve konumumuzu an...	BEDELİ KOMŞU ÜLKELER ÖDÜYOR* Hatay gibi farklı dinlerin farklı dil ve mezheplerin asırlardır bir arada barış içinde yaşadığı ilimize gidecek, inşa
13 hurriyet	2019-05-06	—	Son dakika: Cumhurbaşkanı Erdoğan'dan NATO toplantısında...	Erdoğan, NATO Konseyi ve NATO Akdeniz Diyalogu Ortakları Toplantısı'nda, 2019 yılının Akdeniz Diyalogu'nun 25. NATO'nun kuruluşunun 70'inci
14 hurriyet	2019-05-06	—	MegaCortex adlı yeni bir fidye yazılımı tehlike saçıyor	Sophos güvenlik araştırmacıları, daha önce nispeten alt sıralarda yer alan bir tehdit olan MegaCortex fidye yazılıminin 1 Mayıs'tan itibaren başta
15 hurriyet	2019-05-05	—	Son dakika... NATO'dan Türkiye mesajı	Stoltenberg, 6-7 Mayıs tarihlerinde Türkiye'ye geleceğini ziyaret öncesi Brüksel'de AA muhabirinin sorularını cevapladırdı. Türkiye'ni
16 hurriyet	2019-05-05	—	Kuruluşlar siber güvenlik olaylarına müdahalede hazır değil	Çalışmalar, bir siber saldırının 30 gün içinde kontrol altına alınmak adına hızlı ve etkin bir şekilde müdahale eden şirketlerin toplam veri ihlali maliy
17 hurriyet	2019-05-04	—	TOBB Başkanı Hisarcıklıoğlu: e-ihracatta fırsat çok büyük	TOBB Başkanı Rifat Hisarcıklıoğlu, Balıkesir Avlu Kongre ve Kültür Merkezi'nde, Sanayi ve Teknoloji Bakanı Mustafa Varank'ın katılımıyla gerçekle
18 hurriyet	2019-05-04	—	Bakan Varank açıkladı! 1742 proje için başvuru geldi	Sanayi ve Teknoloji Bakanı Mustafa Varank, KOBİ-Gel Destek Programı kapsamında imalat sanayinde dijitalleşme çabası başvurularının 2 Mayıs'
19 hurriyet	2019-05-04	—	Fikri olanın önu açılacak	TİCARET Bakanlığı koordinasyonuyla Türkiye İhracatçılar Meclisi (TİM) tarafından düzenlenen Türkiye İnovasyon Haftası'nın açılış konuşmalarınd
20 hurriyet	2019-05-03	—	ABD'nin İran endişesi	CNN televizyonuna konuşan ismi açıklanmayan ABD'li hükümet yetkilisi, Washington yönetiminin, Tahran'ın baskılara karşılık ABD'nin Orta Doğu'c
21 hurriyet	2019-05-03	—	TUSAŞ Genel Müdürü Kotil: 10 bin mühendise gidiyoruz	Türk Havacılık ve Uzay Sanayii (TUSAŞ) Genel Müdürü Prof. Dr. Temel Kotil, "TUSAŞ olarak yılda yaklaşık 1 milyar dolar Ar-Ge'ye para harcıyoruz
22 hurriyet	2019-05-03	—	Siber güvenlik alanında önemli satın alma	Türkiye'de yerli ve milli akıllı sistemler geliştirme konusunda önemli çalışmalar yürütün Asis Elektronik ve Bilişim Sistemleri A.Ş., İDEF 2019'da di
23 hurriyet	2019-05-03	—	SIM değişikliği dolandırıcılığı dolandırıcılığı tehlike saçıyor	Operatörün, telefon numaranızı suçluların kontrolündeki bir SIM kartına aktarmaya ikna edilmesine SIM değişikliği dolandırıcılığı adı veriliyor. Bu
24 hurriyet	2019-05-03	—	Sadece bir içinde 1,6 milyon siber saldırı yapılmış	Kaspersky Lab; Güney Afrika'nın Cape Town şehrinde düzenlenen Cyber Security Weekend etkinliğinde Ortadoğu, Türkiye ve Afrika'nın yanı sıra
25 hurriyet	2019-05-03	—	Drone'ları hack'lemek meğer 'çocuk oyuncuğu'ymiş!	Kaspersky Lab'ın Güney Afrika'nın Cape Town şehrinde düzenlediği Cyber Security Weekend 2019 etkinliğinde, "Cyber Ninja" olarak da tanınan :
26 hurriyet	2019-05-03	—	Sosyal medya çocuklar için sakınca içeriklerle dolu	Siberzorbalık.org Kurucusu, İletişim Uzmanı ve Sosyal Medya Danışmanı Nurhan Demirel, konu ile ilgili yaptığı açıklamada "Sosyal medya siteler
27 hurriyet	2019-05-03	—	Hacker'lar işi abarttı, 30 milyar kez sızmayı denedi!	Otomatize saldırılar ırkutucu biçim almasına başladı. 'Kimlik bilgisi doldurma' (credential stuffing) adı verilen otomatik saldırılarında, kötü amaçlı kiş
28 hurriyet	2019-05-03	—	10'u söze üst düzey 93 terörist etkisiz hale getirildi	Çatkılı, Bakanlıktaki basın toplantısında yaptığı nisan ayı değerlendirme de, İçişleri Bakanlığı koordinasyonunda yürütülen iç güvenlik operasyon
29 hurriyet	2019-05-03	—	İçeriden gelen siber tehlkiye aman dikkat!	"Siber tehditlerin genel olarak dışardan geleceği varsayırlar. Ancak pek çok kez tehdit içерiden de oluşabilir" tespiti yapan ESET Türkiye Satış M

Figure 19 Sample Stored Hurriyet Newspaper Data in the Database File

4.1.3 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”.

4.1.4 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.

4.2 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.

4.3 Database Architecture of the System

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

4.3.1 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.

4.3.2 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 first day and the second detections are detected in separate days, our HTML portal which shows the detected events shows this detection information with two separate detection because the period of our software solution is one day.

4.3.3 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.

4.3.4 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.

4.3.5 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.

4.3.6 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.

4.4 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.

4.5 Other Technologies Used in the Thesis Study

Except for the mentioned technologies from the previous chapters, we also used other technologies too. In order to develop such a system in the present thesis, a software implementation is required. We used Python Programming Language to implement the system. “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 to Pandas,” 2014) In this thesis, we use Python version 3.6.6.

We used SQLite as database technology. 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.

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 git for version service with GitHub¹⁴ 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.

¹⁴ The GitHub repository of the present thesis software project is available upon request.

4.6 Summary of the Implementation Chapter

In this chapter, we mentioned the implementation details of the present thesis' software project. We used up to date software technologies, methodologies, and software libraries during the implementation process of the project. There are sixteen implementations related code file in the software project. These are config.py, hurriyetApi.py, hurriyetApiToDb.py, ituNlpPipeline.py, kamuKurumlari.json, logging.conf, manager.py, pipeline.token, pipeline_caller.py, securityEventsDataBase.sqlite, securityEventsWebPortal.py, sqliteOperations.py, twitterStreamToDb.py, twtiterPremiumApi.py, twitterSelenium.py and userInterface.html. These files include thousands line of code mostly written with Python programming language. 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 by changing just the configuration file. The software project is licensed with Apache License 2.0. It is a permissive license whose main conditions require preservation of copyright and license notices. Contributors provide an express grant of patent rights. Licensed works, modifications and larger works may be distributed under different terms and without source code.

CHAPTER 5

RESULTS

In this chapter, 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 20, 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.

The figure shows a screenshot of a web browser displaying the user interface of a cybersecurity detection software. The title bar reads "C:/Users/Ozgur/source/repos/MSThesis/userInterface.html". The main content area has a red header bar with the text "Cybersecurity Events". Below this are two tables, each representing a specific date.

2019-05-14

Entity	Representative News Title or Tweet	Count
meksīka	WhatsApp 'casus yazılımı' hakkında neler biliniyor?	4
israil	WhatsApp, bir grup 'seçilmiş' kullanıcısının casus yazılımla hedef alındığını duyurdu	6
ingiltere	Dijitalleşme ile birlikte şirketleri tehdit eden siber riskler	5
avrupa birliği	2019'un siber güvenlik trendleri açıklandı	4

2019-05-19

Entity	Representative News Title or Tweet	Count
iran	RT @WelayetNews: Son yılda İran'a karşı 33 milyon siber saldırısı yapıldı https://t.co/Fj1PZaTPhJ https://t.co/Zz4lCzKVKE	2
daniştāy	Kumpasların 'bilirkişisi' Estonia'da profesör oldu	3
türkiye	Kumpasların 'bilirkişisi' Estonia'da profesör oldu	8
belçika	Kumpasların 'bilirkişisi' Estonia'da profesör oldu	3
tubitak	Kumpasların 'bilirkişisi' Estonia'da profesör oldu	4
estonya	Kumpasların 'bilirkişisi' Estonia'da profesör oldu	6

Figure 20 User Interface of the Cybersecurity Detection Software

5.1 Historical Cybersecurity Event Detection Test with an Independent Dataset - Nic.tr DDOS Attack

To reach the best version of our software solution, we train our software with training data. In order to do that, we select an important cybersecurity event test that can our solution detect that cybersecurity event.

Turkish Internet hit with massive DDoS attack started on 14.12.2015 and continues about two weeks long. Turkey's official domain name servers (Nic.tr) have been under a Distributed Denial of Service (DDoS) attack. We created 3 separate databases using existing keywords. 2310 tweets were found when we pulled the tweets during the 1-year period before the attack. Then we analyzed these data, our solution can successfully find the cybersecurity events that took place for a year. Some of the example detections can be found in Figure 21.

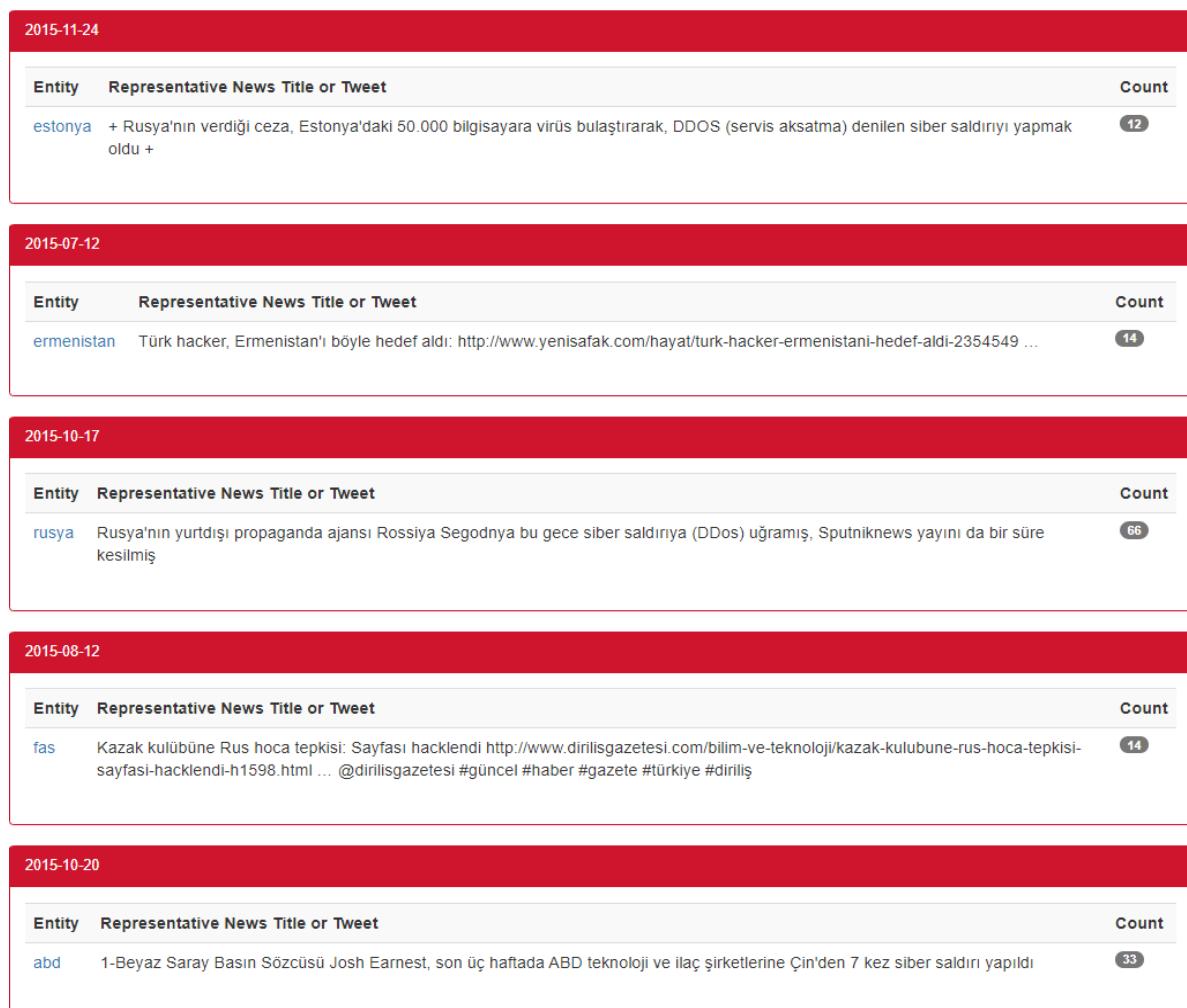


Figure 21 Detected Security Events Before the Nic.tr Attack (2014-2015)

28 tweets were found when we pulled the tweets at the start day of the nic.tr DDOS attack. Results of this day data were important for us because we wanted to see that our solution could detect the event just after the attack happened. Then we analyzed these data, our solution can successfully detect the nic.tr attack as you can see in Figure 22.

2015-12-14		
Entity	Representative News Title or Tweet	Count
nic.tr	Biraz önce ".tr" uzantılı web sitelerine girişte sorun yaşandı. ODTÜ'nün (nic.tr) DNS sunucularına DDOS saldırısı yapıldığı söyleniyor.	3
ant	Biraz önce ".tr" uzantılı web sitelerine girişte sorun yaşandı. ODTÜ'nün (nic.tr) DNS sunucularına DDOS saldırısı yapıldığı söyleniyor.	1
suriye	KAÇIRIP FİDYE İSTEDİLER http://www.edessatv.com/sanliurfa/kacirip-fidye-istediler-h15535.html ... #şanlıurfa #fidye #Suriyeli	2
çin	#istanbul'da #Fidye için kaçırılan #iranlı; "Siz olmasanız ölecektim" http://fb.me/3xAle924U	4
iran	#istanbul'da #Fidye için kaçırılan #iranlı; "Siz olmasanız ölecektim" http://fb.me/3xAle924U	1

Figure 22 Nic.tr Attack Start Day Detected Security Events Samples

The nic.tr attack lasted for about two weeks. Therefore, we analyze that two weeks period (14.12.2015 – 28.12.2015) and we expected to detect the nic.tr attack. About 400 tweets were found when we pulled the tweets for the given period. After running our software solution with that database, the results were satisfactory. Our solution successfully detected the nic.tr attack as you can see in Figure 23.

2015-12-14		
Entity	Representative News Title or Tweet	Count
nic.tr	Ulusal Siber Olaylara Müdahale Merkezi (@TRCert) Nic.tr kesintisinin DDOS saldırısı sonucu meydana geldiğini belirtti. #nictr	8
2015-12-16		
Entity	Representative News Title or Tweet	Count
rusya	ODTÜ/com.tr alan adı veren NİC.TR dDOS saldırısı altında. Siber saldırısı Rusya/Iran kaynaklı olabilir.	8
iran	ODTÜ/com.tr alan adı veren NİC.TR dDOS saldırısı altında. Siber saldırısı Rusya/Iran kaynaklı olabilir.	4
2015-12-26		
Entity	Representative News Title or Tweet	Count
ekonomi bakanlığı	Türk Hacker grubu Ayyıldız Tim, aralarında Rus Ekonomi Bakanlığı'nın sitesinin de bulunduğu 19 bin Rus sitesini hackledi.	10
2015-12-17		
Entity	Representative News Title or Tweet	Count
ant	Ülkenin Yurtdışından erişimi siber saldırı sebebiyle engellenmiş durumda, .tr uzantıları iptal. Haber bile değil.	14
2015-12-15		
Entity	Representative News Title or Tweet	Count
türkiye	Dünya Türkiyenin yeni yaptığı yazılımı konuşuyor nasıl sinyaller kesildi ve musula gece asker sevk edildi. Yazılım konusunda şaşkınlar...	55

Figure 23 Detected Security Events Samples between 14 and 28 December 2015

As we explained in section 3.2 (Determination of Cybersecurity Related Keywords Vector), we used one of the past cybersecurity incidents. We used Term Frequency - Inverse Document Frequency (TF-IDF) analysis of the news and tweets just before the cybersecurity event (premise) and immediately after the event. For immediately after phase, we used two different time intervals for testing. First one is the attack day, and the second one is the two weeks period after the attack. We used the attack day for sensitivity test. Our solution is accepted as successful in terms of sensitivity if it can detect the cybersecurity event at the attack day. We used two weeks of period after attack for certainty. Our solution is accepted as successful in terms of certainty if there is not so many (more than %30) false-positive cybersecurity event detection within two weeks period after a cybersecurity event. According to these success criteria, we train our software solution with the datasets and cybersecurity-related keyword lists. Then update our keyword lists according to the results.

With using these lists, we tested the method and its accuracy in independent data set which is nic.tr DDOS attack dataset in the present section. As can be seen in Figure 22 and Figure 23, our software solution can successfully detect the nic.tr DDOS attack in terms of sensitivity and certainty and passed our test.

5.2 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.

5.2.1 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.

2019-05-14		
Entity	Representative News Title or Tweet	Count
meeksika	WhatsApp 'casus yazılımı' hakkında neler biliniyor?	4
israil	WhatsApp, bir grup 'seçilmiş' kullanıcısının casus yazılımla hedef alındığını duyurdu	6

Figure 24 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 “meeksika” which is the Turkish synonym of Mexico, and “israel” 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 “israel” 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 “meeksika” entity.

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

5.2.2 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.

2019-04-26		
Entity	Representative News Title or Tweet	Count
stm	Uzaktan Hasta Takip Sistemi uygulamalarında tehlike	5

Figure 25 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.

5.2.3 Other Successful Detection Examples

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

2019-03-26		
Entity	Representative News Title or Tweet	Count
abd	ASUS bilgisayarlara 'arz zinciri' saldırısı	2

Figure 26 Supply Chain Attack Targets ASUS Computers Through Backdoored Update

According to the Symantec website, ASUS update system hijacked to send out malicious updates. More than half a million systems have been affected from this attack. (“ASUS Software Updates Used for Supply Chain Attacks | Symantec Blogs,” 2019) Our solution can successfully detect this cybersecurity attack.

azerbaycan	MuddyWater nedir? İşte interneti saran yeni tehlke	4
pakistan	MuddyWater nedir? İşte interneti saran yeni tehlke	3
ürdün	MuddyWater nedir? İşte interneti saran yeni tehlke	5
afganistan	MuddyWater nedir? İşte interneti saran yeni tehlke	6
lübnan	MuddyWater nedir? İşte interneti saran yeni tehlke	8
irak	MuddyWater nedir? İşte interneti saran yeni tehlke	3

Figure 27 MuddyWater Attack

Figure 27 shows the MuddyWater attack detection.

irlanda	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	5
kanada	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	3
avustralya	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	4
endonezya	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	4
amerika birleşik devletleri	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	3
hollanda	MegaCortex adlı yeni bir fidye yazılımı tehlke saçıyor	1

Figure 28 MegaCortex Ransomware

Figure 28 shows a new cyber threat detection named MegaCortex Ransomware.

2019-04-01		
Entity	Representative News Title or Tweet	Count
çin	Siber suçlular şimdi de popüler TV dizilerini kullanıyor	5

Figure 29 Using Popular TV Shows to Spread Malware

Figure 29 exemplifies a new cyber-attack method detected by our software solution.

2019-02-22

Entity	Representative News Title or Tweet	Count
rusya	RTM Bankacılık Truva Atı şirketlere saldırmaya devam ediyor	3

Figure 30 RTM Banking Trojan

Figure 30 is a sample of Banking Trojan news detection.

ant	Anneler Günü gelen tehlikeye aman dikkat!	3
-----	---	---

Figure 31 Mother's Day Cyberattack Alerts

ispanya	Sevgililer Günü'yle birlikte kimlik avı saldırısında patlama yaşanıyor	1
portekiz	Sevgililer Günü'yle birlikte kimlik avı saldırısında patlama yaşanıyor	1
venezuela	Sevgililer Günü'yle birlikte kimlik avı saldırısında patlama yaşanıyor	1
yunanistan	Sevgililer Günü'yle birlikte kimlik avı saldırısında patlama yaşanıyor	1

Figure 32 Valentine's Day Phishing Attacks

Special day-specific cyber-attacks may occur. Figure 31 and Figure 32 are samples of such attacks detected by our software.

fas	Pirate Bay kullanıcılarını bekleyen büyük tehlike	1
-----	---	---

Figure 33 Pirate Matryoshka Virus

Cyber-attackers may use websites. Figure 33 is a sample of that.

ant	Televizyon izleyenleri bekleyen siber tehlike	3
mali	Televizyon izleyenleri bekleyen siber tehlike	1

Figure 34 Cyber Attacks Using Smart TVs

2019-04-03		
Entity	Representative News Title or Tweet	Count
avusturya	Siber saldırganların yeni hedefi: Akıllı Evler	1

Figure 35 Smart Home Cyberattacks

Hardware can also be used by cyber attackers as mentioned in Figure 34 and Figure 35. Our software solution can also detect such attacks.

mali	Siber saldırısı sonrası banka operasyonlarını durdurdu	2
çin	Bakteri ve virüs mücadeleliğinde bağışıklık sisteminizi besleyin	6
ant	Bakteri ve virüs mücadeleliğinde bağışıklık sisteminizi besleyin	4
ingiltere	Siber saldırısı sonrası banka operasyonlarını durdurdu	1
hong	Siber saldırısı sonrası banka operasyonlarını durdurdu	1
malta	Siber saldırısı sonrası banka operasyonlarını durdurdu	1
vemen	Kurallarım vok!	1

Figure 36 Valletta Bank Cyberattack

Detected news related to attacking a bank is in Figure 36.

yunanistan	Gelen bu e-postaları sakın açmayın	1
------------	------------------------------------	---

Figure 37 Business Email Compromise Attack

Entity	Representative News Title or Tweet	Count
amerika birleşik devletleri	Siber dolandırıcılar bu kez Netflix'i alet ediyor	1

Figure 38 Cyberattacks Which Use Netflix Brand

ispanya	Siber saldırganlar şimdi de iş arayanları hedef alıyor	1
brezilya	Siber saldırganlar şimdi de iş arayanları hedef alıyor	1
vietnam	Siber saldırganlar şimdi de iş arayanları hedef alıyor	1
avusturya	Siber saldırganlar şimdi de iş arayanları hedef alıyor	1
rusya	Siber saldırganlar şimdi de iş arayanları hedef alıyor	1

Figure 39 Cyberattacks Targeting Job Seekers

çin	Sosyal paylaşım ağlarına dikkat! Bu yöntemle dolandırıyorlar	6
ant	Sosyal paylaşım ağlarına dikkat! Bu yöntemle dolandırıyorlar	3
bilgi teknolojileri ve iletişim kurumu	Sosyal paylaşım ağlarına dikkat! Bu yöntemle dolandırıyorlar	1

Figure 40 Phishing Attacks Targeting Social Media Users

New kind of phishing attacks is detected by our software as can be seen in Figure 37, Figure 38, Figure 39 and Figure 40.

2019-03-21		
Entity	Representative News Title or Tweet	Count
çin	Reklamcılığın gizli aktörleri: "çerezler"	4
ant	Reklamcılığın gizli aktörleri: "çerezler"	2
bursa teknik üniversitesi	Reklamcılığın gizli aktörleri: "çerezler"	1
türkiye	Reklamcılığın gizli aktörleri: "çerezler"	1

Figure 41 Zombie Cookies

Zombie cookies is a new kind of thread detected by our software. Related news detected by our software as can be seen in Figure 41.

[çin](#) Milyonlarca e-mail adresi interne sızdı! Tehlike kapıda... 6

Figure 42 Millions of Email Addresses Infiltrated the Internet

Important information can be stolen and shared by the attackers on the internet as you can see in Figure 42.

[trt](#) @trthaber birleşik ülke ile prc siber savaşı başladı bakanım daha neler göreceğiz 2

Figure 43 The USA and China Cyberwar Started Tweet

Cyberwars also can be detected by our software solution. Figure 43 is exemplifying that.

Entity	Representative News Title or Tweet	Count
almanya	Almanya'da hacker şoku: Cumhurbaşkanı'nın bile kimlik bilgilerini çaldılar	3

Figure 44 Angela Merkel and Hundreds of German Politicians Hacked

Politicians can be the target of cyber-attacks as you can see in Figure 44. Our software can successfully detect related news of such events.

5.3 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.

5.3.1 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.

ant @omerturantv72 Ömet bey inanamıyorum; gerçekten bunları siz mi yaziyorsunuz yoksa hesabınız mı hacklendi? ④

Figure 45 Sample False Positive Cybersecurity Event Detection

5.3.2 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 her/him.

cin Beni takip eden bütün takipçilerime duyurumdu: Twitter hesabım hacklendiği için abuk sabuk reklam, ilan ve de tele...
<https://t.co/8njVtCkfNQ> ⑦

Figure 46 Sample, not Useful Cybersecurity Event Detection

Even if the detected event is not a personal cybersecurity event, the detected event may not be useful for security events. For example, an event may occur months ago, but a Twitter user or a Twitter bot may share the event in a Tweet as if it occurred newly. The time frame is configurable in our software system. Security analysts should configure the software detection timeframe according to their needs. For example, if a security analyst works for a big cybersecurity technology company and he/she wants to know more detected security events, he/she can set the timeframe longer. However, if another security analyst wants to know only the latest security events, he/she should set smaller timeframe in our software solution.

5.4 Evaluation of the Results

When we run our software with too much the cybersecurity-related seed keywords vector, our software system might receive more tweets than it can handle. Moreover, the false-positive cybersecurity event detection may significantly increase. It decreases the

certainty of our software solution. On the other hand, if we run our software with too few cybersecurity-related seed keywords, our software system might not detect some cybersecurity events as fast as we expect from our software. It decreases the sensitivity. We expect that we can detect an attack on the day of the attack.

Although we can verify with other sources that the detected events are indeed occurring, or occurred, being sure that we have missed any events is very difficult. During our tests, we realized that we could miss small events. However, our solution does not miss any serious attack as far as we know.

Sometimes our solution detects an already detected event as if it is a new cybersecurity event. Because our software uses one day as a period for its frequency calculation. For each day, all calculations start from zero again.

For a limited time, we run our software for testing purposes. We cannot calculate something like detected Tweet per minute or detected news per minute because it is highly related with selected named entity vector or the selected period for the detection. At a sample test run of our software solution, our database of the software includes 437 entries. One hundred eighty-six of them is Twitter Tweets, and 251 of them is from Hürriyet Newspaper. After analyzing the entries in our database, our software solution can detect 29 cybersecurity events. Twenty-two of them are positive detection, and 7 of them are false positive detection. Our software solution's success rate is approximately %76.¹⁵ These statistics show that this methodology works in the detection of cybersecurity events from Turkish texts with an acceptable success rate in term of certainty and sensitivity. Cybersecurity analysts can use our software with preparing our cybersecurity-related keyword vector and named entity vector and selecting a suitable timeframe. Moreover, they can modify the keyword vector or named entity vector as they wish. If we add new data sources in the future, our software can work with bigger datasets and this leads to more accurate detection and it may increase the success rate percent of our software solution in terms of certainty and sensitivity.

¹⁵ These results are for limited and have demonstrative constraints. Success rate is shared just for giving an idea to readers and it is dependent external factors like time frame, cyber-attack types, Tweet number, newspaper news and so on. The success rate can increase or decrease according to these factors.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 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 cybersecurity 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.

6.2 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.

We do not handle the named entity recognition ambiguities yet. We are planning to handle them in the future.

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>
- ASUS Software Updates Used for Supply Chain Attacks | Symantec Blogs. (2019). Retrieved June 28, 2019, from <https://www.symantec.com/blogs/threat-intelligence/asus-supply-chain-attack>
- 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/>
- Differences between a security event vs incident vs non-compliance. (n.d.). Retrieved June 2, 2019, from <https://advisera.com/27001academy/blog/2018/12/03/iso-27001-information-security-event-vs-incident-vs-non-compliance/>
- Eryiğit, G. (2015). *ITU Turkish NLP Web Service.* 1–4. <https://doi.org/10.3115/v1/e14-2001>
- Eryiğit, G., & Torunoğlu-Selamet, D. (2017a). Social media text normalization for Turkish. *Natural Language Engineering,* 23(6), 835–875. <https://doi.org/10.1017/S1351324917000134>

- Eryiğit, G., & Torunoğlu-Selamet, D. (2017b). 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 & 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/>
- Pardot. (2012). The ABCs of A/B Testing | Salesforce Pardot. Retrieved September 8, 2019, from <https://www.pardot.com/blog/abcs-ab-testing/>

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/>

Rajaraman, A.; Ullman, J. D. (2011). ch1 | Database Index | Machine Learning. Retrieved September 7, 2019, from Mining of Massive Datasets website: <https://www.scribd.com/document/251633025/ch1>

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/>

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

Voyant Tools Help. (n.d.). *Voyant-Tools.Org*. Retrieved from <https://voyant-tools.org/docs/#!/guide/cirrus>

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-google-dan-dondu-22290509>

Zimek, A., & Schubert, E. (2017). *Outlier Detection BT - Encyclopedia of Database Systems* (L. Liu & M. T. Özsü, Eds.). https://doi.org/10.1007/978-1-4899-7993-3_80719-1

APPENDICES

APPENDIX A

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 ACESSECRET

ITU NLP API Related Constants Stored in the Config File

CONSUMER_NAME

CONSUMER_PASSWORD

ITU_NLP_API_TOKEN

PIPELINE_ENCODING

DEFAULT_TOOL

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

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,” 2019) 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.

APPENDIX B

NAMED ENTITIES AND CYBERSECURITY RELATED KEYWORDS VECTORS IN TURKISH

CYBERSECURITY RELATED KEYWORDS		
anonymous,	"erişmekte güçlük",	"session hijacking",
"sayfa açılmıyor",	"erişim sağlanamıyor",	"spear-phishing",
"bigisayar korsamı",	"erişim problemi",	"spear phishing",
"bilgisayar korsanları",	"firewall",	"sql injection attack",
"brute force attack",	"fidye yazılımı",	"sql injection saldırısı",
"brute force saldırısı",	"forensics",	"sızma",
"casus yazılım",	"güvenlik duvarı",	"sızmış",
"clone phishing",	"hack",	"sızıntı",
"cross-site scripting",	"hacklendi",	"sistem açığı",
"cross site scripting",	"hacker",	"siber",
"credential reuse",	"istihbarat",	"ulaşlamıyor",
"cyber",	"link spoofing",	"unreachable",
"cyberattack",	"malware",	"whaling",
"cyberwarfare",	"man-in-the-middle",	"website spoofing",
"çöktü",	"mitm",	"yemleme",
"down olmuş",	"virüs bulaştı",	"xss",
"denial-of-service",	"virüs çıkarıyor",	"zafiyet",
"dictionary attack",	"oltalama",	"zararlı yazılım"
"dictionary saldırısı",	"oturum çalma",	"trojan horse"
"dos",	"phising",	"turuva atı"
"ddos",	"ransomware",	"rootkit"
"kullanıcı hesapları ele geçirildi",	"rainbow attack",	"buffer overflow"
"kimlik avı",	"rainbow saldırısı",	"backdoor"
"erişilemiyor",	"sayfa açılmıyor",	"worm"
"erişim güclüğü",	"saldırı",	"data breach"

Named Entity Vectors in Turkish	
Institution and Government Organization Names in Turkey	
Country Names	
"Türkiye",	"Adalet Bakanlığı",
"ABD",	"Adiyaman 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 Adıgüzél Meslek Yüksekokulu",
"Benin",	"Atatürk Araştırma Merkezi",
	"Atatürk Kültür, Dil ve Tarih Yüksek Kurumu
"Beyaz Rusya",	Başkanlığı",
"Bermuda",	"Atatürk Kültür Merkezi",
"Bhutan",	"Atatürk Orman Çiftliği",
"Birleşik Emirlikleri",	Arap "AB Bakanlığı",
	"Avrupa Birliği Eğitim ve Gençlik Programları
"Birleşik Krallık",	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",	"Cumhurbşakanlığı",
"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",	"Esnaflar 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",
"Kirgızistan",	"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",	"Hittit 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 Enfrastruktur 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",
"Misir",	"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 Varlıklarları 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 Kurulu 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",

"Sili",	"Milli Kütüphane Başkanlığı",
"Tacikistan",	"Milli Piyango İdaresi Genel Müdürlüğü",
"Tanzanya",	"Millî Savunma Bakanlığı",
"Tayland",	"Millî 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",
"Zimbabwe"	"PTT",
	"Radyo ve Televizyon Üst Kurulu",
	"Refik Saydam Hıfzıssıhha 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",
"Tarimsal Araştırmalar ve Politikalar Genel Müdürlüğü",
"Tarimsal 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",
"Yüksekokretim Kurulu Başkanlığı",
"YÖK",
"Yüksek Seçim Kurulu Başkanlığı"
"YSK",

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics

Enformatik Enstitüsü / Graduate School of Informatics X

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences

YAZARIN / AUTHOR

Soyadı / Surname : Ural

Adı / Name : Özgür

Bölümü / Department : Cybersecurity

TEZİN ADI / TITLE OF THE THESIS (İngilizce / English) : AUTOMATIC DETECTION OF CYBER SECURITY EVENTS FROM TURKISH TWITTER STREAM AND TURKISH NEWSPAPER DATA

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master X

Doktora / PhD

1. **Tezin tamamı dünya çapında erişime açılacaktır.** / Release the entire work immediately for access worldwide. X
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of two year. *
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of six months. *

* Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature 

Tarih / Date 20.08.2019