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

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF INFORMATICS

OF

THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGÜR URAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF SCIENCE

IN

CYBER SECURITY

MAY 2019

Approval of the thesis:

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

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

Prof. Dr. Deniz Zeyrek Bozşahin

Dean, **Graduate School of Informatics**

Prof. Dr. Aysu Betin Can

Head of Department, **Cyber Security**

Assist. Prof. Dr. Cengiz Acartürk

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

Assoc. Prof. Dr. XXX

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

**Examining Committee Members:**

Prof. Dr. XXX (\*)

Computer Engineering Dept., METU

Prof. Dr. XXX (\*\*)

Medical Informatics Dept., METU

Prof. Dr. XXX

Cognitive Science Dept., METU

Prof. Dr. XXX

Computer Engineering Dept., Boğaziçi University

Prof. Dr. XXX

Computer Engineering Dept., Bilkent University

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

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

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

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

**Name, Last name : ÖZGÜR URAL**

**Signature :**

# ABSTRACT

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

Ural, Özgür

MSc., Department of Cyber Security

Supervisor: Assist. Prof. Dr. Cengiz Acartürk

May 2019, XX pages

Thesis Abstract – max 250 words

Internet contains various time dependent information. Every day, security experts face lots of security events that affect people, institutions and governments. Staying up to date on security events requires that an information analyst constantly scan many sources, leading to information overload. For example, an information analyst might want to be aware of cyber security incidents such as a DDOS attack on a government agency website. If they can earlier detect and understand these threats, the more time left to mitigate and forensically investigate them. Therefore, they need to have a situation awareness of the existing security events and their possible effects. However, given the large number of events, it can be difficult for security analysts and researchers to handle this flow of information in an adequate manner and answer the following questions in near real-time: what are the current security events? How long they last? Can we detect them with using data sources in Turkish language? In this thesis, I try to answer these issues with using Twitter social network and Hürriyet Turkish newspaper that contain a massive amount of valuable information. However, because of the high volume, extracting meaningful information can be challenging. For this reason, I propose an automatic, Turkish language specific software system that can detect cyber security events in near real time over the Twitter stream in Turkish language and Hurriyet newspaper stream.

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

# ÖZ\_\_

TÜRKÇE TWITTER AKIŞI VE TÜRKÇE GAZETE VERİLERDEN SİBER GÜVENLİK OLAYLARININ OTOMATİK TESPİT EDİLMESİ

Ural, Özgür

Yüksek Lisans, Siber Güvenlik Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Cengiz Acartürk

Mayıs 2019, XX sayfa

Tez özeti – en fazla 250 kelime

İnternet çeşitli zamana bağlı bilgiler içerir. Her gün, güvenlik uzmanları insanları, kurumları ve hükümetleri etkileyen birçok güvenlik olayıyla karşı karşıya. Güvenlik olaylarından haberdar olmak bir bilgi analistinin sürekli olarak birçok kaynağı taramasını ve böylece aşırı bilgi yüklenmesini gerektirir. Örneğin, bir bilgi analisti, bir devlet kurumu web sitesinde DDOS saldırısı gibi siber güvenlik olaylarının farkında olmak isteyebilir. Bu tehditleri daha erken tespit edip anlayabilirlerse, onları hafifletmek ve adli olarak soruşturmak için daha fazla zaman kalır. Bu nedenle, mevcut güvenlik olayları ve olası etkileri hakkında bir durum bilgisine sahip olmaları gerekir. Ancak, çok sayıda olaya bakıldığında, güvenlik analistlerinin ve araştırmacıların bu bilgi akışını yeterli şekilde ele almaları ve aşağıdaki soruları gerçek zamanlı olarak cevaplamaları zor olabilir: mevcut güvenlik olayları nelerdir? Ne kadar sürüyorlar? Veri kaynaklarını Türkçe dilinde kullanarak tespit edebilir miyiz? Bu tezde, Twitter sosyal ağını ve çok miktarda değerli bilgi içeren Hürriyet Türk gazetesini kullanarak bu sorunları çözmeye çalışıyorum. Bununla birlikte, yüksek hacimlerden dolayı anlamlı bilgilerin çıkarılması zor olabilir. Bu nedenle siber güvenlik olaylarını neredeyse gerçek zamanlı olarak Twitter üzerinden Türk dilinde ve Hürriyet gazetesi akışında tespit edebilen otomatik Türkçe'ye özel bir yazılım sistemi öneriyorum.

Anahtar Sözcükler: xx, yy, zz… (en fazla 5 anahtar kelime)

# DEDICATION

To My Family

# ACKNOWLEDGMENTS

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

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

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

To my wife, …..

# TABLE OF CONTENTS

[ABSTRACT iv](#_Toc5616971)

[ÖZ\_\_ vi](#_Toc5616972)

[DEDICATION vii](#_Toc5616973)

[ACKNOWLEDGMENTS viii](#_Toc5616974)

[TABLE OF CONTENTS ix](#_Toc5616975)

[LIST OF TABLES x](#_Toc5616976)

[LIST OF FIGURES xi](#_Toc5616977)

[LIST OF ABBREVIATIONS xii](#_Toc5616978)

[1 INTRODUCTION 13](#_Toc5616979)

[1.1 Motivation 13](#_Toc5616980)

[1.2 Problem Definition 13](#_Toc5616981)

[1.2.1 Initial Subject Definition 13](#_Toc5616982)

[1.3 Objectives 13](#_Toc5616983)

[1.4 Contributions of the Thesis 13](#_Toc5616984)

[1.5 Use Cases 13](#_Toc5616985)

[1.6 Outline 13](#_Toc5616986)

[2 BACKGROUND INFORMATION 14](#_Toc5616987)

[2.1 What is anInformation Security Analyst? 14](#_Toc5616988)

[2.2 What does an Information Security Analyst do? 15](#_Toc5616989)

[2.3 What is Natural Language Processing? 15](#_Toc5616990)

[2.4 Why do we need Natural Language Processing? 15](#_Toc5616991)

[2.5 Sample Tweets Related with a Security Incident 15](#_Toc5616992)

[2.6 Why is NLP Hard? 15](#_Toc5616993)

[2.7 Twitter Social Network 15](#_Toc5616994)

[2.7.1 Mechanism and Specific Terminology 15](#_Toc5616995)

[2.7.2 Working with Twitter Data 15](#_Toc5616996)

[2.7.3 Twitter, a Social Network? 15](#_Toc5616997)

[2.8 Hürriyet Turkish Newspaper as a Data Source 15](#_Toc5616998)

[3 LITERATURE REVIEW 16](#_Toc5616999)

[4 SYSTEM ARCHITECTURE AND DESIGN 16](#_Toc5617000)

[4.1 Approach 16](#_Toc5617001)

[4.2 Taxonamy 16](#_Toc5617002)

[4.3 Data Collection 16](#_Toc5617003)

[4.4 Data Preprocessing 16](#_Toc5617004)

[4.5 Event Detection 16](#_Toc5617005)

[5 IMPLEMANTATION AND EVALUATION 17](#_Toc5617006)

[5.1 Implemantation 17](#_Toc5617007)

[5.2 Events Evaluation 17](#_Toc5617008)

[5.2.1 Yahoo Data Breach ie 17](#_Toc5617009)

[6 RESULTS 17](#_Toc5617010)

[7 CONCLUSION AND FUTURE WORK 17](#_Toc5617011)

[7.1 Conclusion 17](#_Toc5617012)

[7.2 Future Work 17](#_Toc5617013)

[REFERENCES 18](#_Toc5617014)

[APPENDICES 29](#_Toc5617015)

[APPENDIX A 29](#_Toc5617016)

[APPENDIX B 30](#_Toc5617017)

# LIST OF TABLES

[Table 1: Decision Making Terminology 1](#_Toc444263431)

[Table 2: xxxx 7](#_Toc444263432)

# LIST OF FIGURES

[Figure 1: xxx 5](file:///C:\Users\Enformatik\Desktop\tEZ%20öRNEK.docx#_Toc444263433)

[Figure 2:xxx 8](file:///C:\Users\Enformatik\Desktop\tEZ%20öRNEK.docx#_Toc444263434)

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **DDOS** | Distributed Denial of Service |
| **DOS** | Denial of Service |

**CHAPTER 1**CHAPTER

# INTRODUCTION

## Motivation

Security awareness tools helps security analysts to protect a company's sensitive and mission-critical data from being stolen, damaged or compromised by hackers. In the case of a newly discovered vulnerability, the delay between the disclosure and the moment when the security practitioner is aware of it is crucial. For example, attack such as the MongoDB ransomware can be mitigated as soon as the user is aware of the flaw and the possible patch. In order to learn such information, the analyst has a wide range of sources available: specialized press, tech forums and even specialized communication protocol for the dissemination of cyber threat information.

Social media can be also used for such a purpose. Characteristics of Twitter make it the foremost choice in the case of real-time event detection with 321 million monthly active users as reported by the Statista website.

Major events are discussed on social media and often the first reference for such event is on those social networks. In our case, an event is a phenomenon happening during a certain time period which stimulates people to post about it on social media.1 For example, denial of service attacks are often first reported by users of the website or service under attack. Users of online social networks witnessing or participating in an event are naturally incentivized to discuss it on social media (e.g., tweeting “I cannot reach X website” ). Moreover, Twitter act as a platform where not only cybersecurity media but also mainstream media will tweet about ongoing events and where people can react to such events. Nonetheless, dealing with document such as tweets for the purpose of detecting cyber security related events poses a wide variety of challenges. First, in order to have a general overview of the current situation, the framework must be able to proceed in real-time and in a timely manner a high-volume data. Hence, all the algorithms used should be fast, scalable and possibly distributed.

Secondly, given the nature of tweets, documents retrieved are often unusable: there is a lot of duplicates, off-topic, badly written and incomplete documents. As we can get very valuable information (e.g., information on a data breach), we can also get a lot of irrelevant information (e.g., how I feel today). One of the main challenges lies in correctly filtering the noise from the retrieved content. Finally, the lexical field of cyber security event is growing and constantly changing which makes the tracking of new terminology a hard and time-consuming task. Having an autonomous system that leverage social network’s data for security event detection would be extremely useful for a security analyst. Moreover, the ability to have an evolving scope of detected cyber security related events represent a substantial advantage.

There are several academic researches to achieve such system. During the analyze the twitter stream, such systems filters English data and analyze it to gather valuable information. Twitter users can tweet in any languages they wish. Turkish people tend to use social media. According to statistia website, there are 9 million active Twitter users in Turkey. Turkish tweets can be also used to detect cyber security events. However Turkish language specific steps and requirements shall be added to the detection system and algorithms.

To address these challenges, I designed and developed a software system capable of detecting and monitoring cyber security 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, I added Hürriyet Turkish newspaper stream to analyze and detect security events. The software solution’s infrastructure supports adding new data resources. For example, it can be added LinkedIn, Facebook, Ekşisözlük website streams to gain more accurate results.

## Problem Definition

### Initial Subject Definition

## Objectives

## Contributions of the Thesis

## Use Cases

## Outline

**CHAPTER 2**

# BACKGROUND INFORMATION

## What is anInformation Security Analyst?

An information security analyst is someone who takes measures to protect a company's sensitive and mission-critical data, staying one step ahead of cyber attackers. They do this by coming up with innovative solutions to prevent critical information from being stolen, damaged or compromised by hackers.

Note the differences between a Security Analyst and a Security Administrator:

* Security Analysts - are responsible for analyzing data and recommending changes to higher ups, but do not authorize and implement changes. Their main job is keeping attackers out.
* Security Administrators - ensure that systems are working as designed by making changes, applying patches and setting up new admin users. Their main job is keeping systems up.

## What does an Information Security Analyst do?

## What is Natural Language Processing?

## Why do we need Natural Language Processing?

## Sample Tweets Related with a Security Incident

## Why is NLP Hard?

## Twitter Social Network

### Mechanism and Specific Terminology

### Working with Twitter Data

### Twitter, a Social Network?

## Hürriyet Turkish Newspaper as a Data Source

**CHAPTER 3**

# LITERATURE REVIEW

**CHAPTER 4**

# SYSTEM ARCHITECTURE AND DESIGN

## Approach

## Taxonamy

## Data Collection

## Data Preprocessing

## Event Detection

**CHAPTER 5**

# IMPLEMANTATION AND EVALUATION

## Implemantation

## Events Evaluation

### Yahoo Data Breach ie

**CHAPTER 6**

# RESULTS

**CHAPTER 7**

# CONCLUSION AND FUTURE WORK

## Conclusion

## Future Work

# REFERENCES

[1] B. McKenna, “Symantec’s Thompson pronounces old style IT security dead,” *Netw. Secur.*, vol. 2005, no. 2, pp. 1–3, 2005.

# APPENDICES

# APPENDIX A

**TITLE**

xxx

# APPENDIX B

**TITLE**

xxx