

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**AUTOMATIC DETECTION OF CYBERBULLYING
IN SOCIAL NETWORKS**

by
Alican BOZYİĞİT

September, 2021

İZMİR

AUTOMATIC DETECTION OF CYBERBULLYING IN SOCIAL NETWORKS

**A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of
Philosophy in Computer Engineering**

**by
Alican BOZYİĞİT**

September, 2021

İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "AUTOMATIC DETECTION OF CYBERBULLYING IN SOCIAL NETWORKS" completed by ALİCAN BOZYİĞİT under supervision of ASSOC. PROF. DR. SEMİH UTKU and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

.....
Assoc. Prof. Dr. Semih UTKU

Supervisor

.....
Prof. Dr. Efendi NASİBOĞLU

Thesis Committee Member

.....
Assoc. Prof. Dr. Derya BİRANT

Thesis Committee Member

.....
Assoc. Prof. Dr. Deniz KILINÇ

Examining Committee Member

.....
Assoc. Prof. Dr. Gıyasettin ÖZCAN

Examining Committee Member

.....
Prof. Dr. Özgür ÖZÇELİK

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to acknowledge a number of people for their help and support during the production of this thesis.

My first and big appreciation goes to my supervisor, Assoc. Prof. Dr. Semih Utku, for his supervision, noble guidance, kind support with full encouragement, and valuable suggestion. I am incredibly grateful that you took me on as a student and continued to have faith in me over the years.

Thank you to my thesis committee members, Prof. Dr. Efendi Nasiboğlu and Assoc. Prof. Dr. Derya Birant. Your detailed feedback and suggestions have been very important to me during my thesis progress.

My friends and colleagues, Mr. Yiğit Diker, Dr. Can Atılgan, and Dr. Barış Tekin Tezel, deserve my special thanks for being valuable fellows in my path of being a philosopher. Also, I would like to thank all my colleagues in the Department of Computer Science.

My warm and heartfelt thanks go to my father Mr. Metin Bozyiğit, mother Mrs. Elif Bozyiğit, sister Dr. Fatma Bozyiğit, and brother Mr. Hakan Bozyiğit for their strong support as well as regular encouragement in every step to make me in the present stage. Especially, I am indebted to my father for helping me improve my analytical thinking in my childhood and my mother for helping me take my education in proper schools.

Finally, I thank with love to my dear wife, Mrs. Bahar Bozyiğit. She has been a great companion, loved, supported, encouraged, entertained, and helped me get through this difficult period in the most positive way. I am so glad to have you in my life.

Alican BOZYİĞİT

AUTOMATIC DETECTION OF CYBERBULLYING IN SOCIAL NETWORKS

ABSTRACT

Cyberbullying has become a major problem that affects children and youngsters especially. In this thesis, it is aimed to detect cyberbullying content in social networks automatically. In this direction, a comprehensive dataset, which includes social media features (e.g., number of the sender followers), was systematically prepared. Then, the characteristics of cyberbullying are inspected by analyzing the natural language processing features (e.g., the number of title words) and social media features on the prepared dataset. It is seen that some of the social media features are strongly related to cyberbullying. Additionally, some association rules between social media features and cyberbullying were captured, such as users that have more followers on social networks are disinclined to post online bullying content. The obtained results show that social media features would be promising in automatically detecting harmful content in social networks. Accordingly, machine learning algorithms experimented on two different variants of the prepared datasets. The first variant includes only textual features, whereas the second variant consists of the determined social media features and textual features. It is observed that each experimented machine learning algorithm gives more successful prediction performance on the variant containing social media features. Further experiments in machine learning were conducted by implementing word embedding approaches in the feature extraction to increase the performance of the applied machine learning algorithms. Lastly, an open web service that uses the trained machine learning models for cyberbullying detection was published to motivate programmers to develop real-time applications without studying or knowing the machine learning process.

Keywords: Cyberbullying detection, social media analysis, machine learning, text mining

SOSYAL AĞLARDA SANAL ZORBALIĞIN OTOMATİK OLARAK TESPİT EDİLMESİ

ÖZ

Son zamanlarda sanal zorbalık özellikle çocukları ve gençleri etkileyen önemli bir sorun haline gelmiştir. Bu tezde, sosyal ağlardaki sanal zorbalık içeriklerinin otomatik olarak tespit edilmesi amaçlanmıştır. Bu doğrultuda sosyal medya özelliklerini (örneğin paylaşım yapan kişinin takipçi sayısı) içeren kapsamlı bir sanal zorbalık veri seti sistematik olarak hazırlanmıştır. Daha sonra sanal zorbalık karakteristiklerini tespit etmek için hazırlanan veri seti üzerinde doğal dil işleme özellikleri (örneğin paylaşımdaki kelime sayısı) ve sosyal medya özellikleri analiz edilmiştir. Analiz çalışmasının sonuçları değerlendirildiğinde bazı sosyal medya özellikleri ve siber zorbalık aktiviteleri arasında kuvvetli bir ilişki olduğu görülmektedir. Ayrıca, sosyal ağlarda daha fazla takipçisi olan kullanıcıların çevrimiçi zorbalık içeriği paylaşmaktan kaçınması gibi bazı birliktelik kuralları çıkartılmıştır. Elde edilen sonuçlar, sosyal medya özelliklerinin sanal zorbalık içeriklerinin otomatik olarak tespit edilmesinde faydalı olabileceğini göstermiştir. Bu doğrultuda makine öğrenmesi algoritmaları, hazırlanan veri setinin iki farklı varyantı üzerinde denenmiştir. Birinci veri seti varyantı sadece metinsel özellikleri içerirken, ikinci varyant sosyal medya özellikleri ve metinsel özellikleri birlikte içermektedir. Uygulanan her makine öğrenmesi algoritmasının sosyal medya özelliklerini içeren varyant üzerinde daha başarılı sınıflandırma performansı verdiği gözlemlenmiştir. Sonrasında uygulanan makine öğrenimi algoritmalarının performansını artırmak için özellik çıkarmada kelime temsil yaklaşımları uygulanmıştır. Son olarak, programcılar makine öğrenimi sürecini çalışmadan sanal zorbalık alanında gerçek zamanlı uygulamalar geliştirmeye motive etmek için eğitilen makine öğrenimi modellerini kullanan açık bir web servis yayınlanmıştır.

Anahtar kelimeler: Sanal zorbalık, sosyal medya analizi, makine öğrenmesi, metin madenciliği

CONTENTS

	Page
Ph.D. THESIS EXAMINATION RESULT FORM	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
CHAPTER 1 – INTRODUCTION	1
1.1 General Introduction	1
1.2 Motivation	2
1.3 Organization of the Thesis	4
CHAPTER 2 – BACKGROUND	6
2.1 Cyberbullying	6
2.1.1 The Consequences of Cyberbullying	7
2.1.2 Real Life Cyberbullying Stories	8
2.1.3 The Ways of Preventing and Detecting Cyberbullying	9
2.2 Machine Learning	10
2.2.1 Reinforcement Learning	11
2.2.2 Unsupervised Learning	12
2.2.3 Supervised Learning	13
CHAPTER 3 – RELATED WORK	15
3.1 Preliminary Works	15
3.2 Cyberbullying Detection Using Text Mining	16
3.3 Cyberbullying Detection Using Contextual Features	17
3.4 Cyberbullying Detection for Turkish	19

3.5 General Overview.....	19
CHAPTER 4 – DATA PREPARATION	22
4.1 Data Collection	22
4.2 Data Elimination	25
4.3 Data Annotation	26
CHAPTER 5 – FEATURE ANALYSIS	30
5.1 Social Media Features	30
5.1.1 Chi-square Test	31
5.1.2 Association Rule Mining	33
5.2 Natural Language Processing Features	36
CHAPTER 6 – AUTOMATIC DETECTION OF CYBERBULLYING.....	38
6.1 Preprocessing	38
6.2 Feature Extraction	40
6.3 Applied Machine Learning Algorithms	41
6.4 Experimental Results	42
6.5 Further Experiments: Word Embedding	47
6.5.1 Word2Vec	47
6.5.2 FastText.....	49
6.5.3 Cyberbullying Detection with Word Embedding Approaches	49
CHAPTER 7 – CYBERBULLYING DETECTION WEB SERVICE.....	53
7.1 Web Service Architecture	53
7.2 Web Service Endpoints.....	54
CHAPTER 8 – CONCLUSION	56
8.1 Summary	56
8.2 Future Work	57
REFERENCES.....	59

LIST OF FIGURES

	Page
Figure 1.1	The percentage of children being cyberbullied..... 3
Figure 2.1	The categories of machine learning algorithms 11
Figure 2.2	The architecture of reinforcement learning 12
Figure 2.3	The flow of supervised learning 13
Figure 4.1	The data collection software system 23
Figure 4.2	ER diagram of the database..... 24
Figure 4.3	The data elimination process 25
Figure 4.4	The mock-up user interface of the data annotation application 26
Figure 4.5	The activity diagram of user annotation..... 27
Figure 4.6	ER diagram of the web application's database 28
Figure 5.1	Chi-Square scores of the features 32
Figure 5.2	The relationship between sender followers and cyberbullying 33
Figure 5.3	Chi-Square scores of the features 37
Figure 6.1	Pseudo code of the text preprocessing method..... 39
Figure 6.2	The brief flow chart of grid search..... 43
Figure 6.3	The accuracy of the experimented classifiers 45
Figure 6.4	The architecture of continuous bag-of-words 48
Figure 6.5	The architecture of skip-gram..... 48
Figure 6.6	The word embedding parameter optimization algorithm..... 50
Figure 6.7	The experimental results of Word2Vec 51
Figure 6.8	The experimental results of FastText..... 52
Figure 7.1	The architecture of the developed web service..... 54

LIST OF TABLES

		Page
Table 3.1	The summary of reviewed studies.....	20
Table 4.1	A sample from the dataset.....	29
Table 5.1	Social media features	30
Table 5.2	The discovered cyberbullying rules.....	35
Table 5.3	Natural language processing features	36
Table 6.1	Number of textual features and optimal parameters of classifiers ...	44
Table 6.2	The detailed results of the experimented classifiers.....	46
Table 7.1	The endpoints of web service.....	55
Table 7.2	The input parameters for web service request	55

CHAPTER 1

INTRODUCTION

In this chapter, there is a general introduction to this thesis in the first section. In the second section, the motivation of the study is detailed. The organization of the thesis is presented in the last section.

1.1 General Introduction

Social media has become a popular and essential communication tool in today's modern world. According to the research conducted by Chaffey (2020), more than three billion people use social networks for communication. It is clear that social media apps provide various advantages to their users; however, they can be also used for malicious purposes. One of the malicious purposes is cyberbullying, which is defined as bullying a person or a group of people using digital technologies (Slonje & Smith, 2008).

In this thesis, it is aimed to detect cyberbullying content in social networks automatically. Most of the conducted studies in the literature only use text mining techniques in the same vein as sentiment analysis works; however, social media posts are interactive and context-dependent (Modha et al., 2020). Accordingly, utilizing social network features, e.g., the number of users who liked a post, would be useful for cyberbullying detection. In this direction, a novel comprehensive dataset that consists of many social media and textual features (Bozyiğit et al., 2021) was systematically created for this study.

In order to analyze the characteristics of cyberbullying, feature analysis was applied to social media features and natural language processing features of the created dataset. Firstly, the Chi-Square Test (McHugh, 2013) was used to capture the relationship between the features and online bullying events. In the test results, it is seen that some of the social media features have a significant relationship with cyberbullying. Then, Association Rule Mining (Zhang & Zhang, 2003) was applied

to the dataset for representing the captured relationships as rules. The results show that using social media features can make a noteworthy contribution to a machine learning process.

For automatic classification of cyberbullying posts, various text mining techniques experimented on the created dataset standalone and also together with social media features. The accuracy of some experimented machine learning algorithms is close to 90%, which means they can correctly classify nine of ten social media contents in terms of cyberbullying. Moreover, it is observed that using both social media features and textual features significantly improves the classification performance of the algorithms, so confirm the findings of feature analysis.

Accordingly, an open web service that uses the trained machine learning models (Bozyiğit, 2021) was published for online bullying detection within the scope of this thesis. Thus, any programmer intended to develop cyberbullying applications can easily integrate this web service into their application even without studying or knowing the machine learning process. Moreover, the developed applications can motivate other programmers to produce more real-time apps in this field and consequently increase awareness in cyberbullying detection.

1.2 Motivation

Cyberbullying has stronger and more outlasting effects compared to the classical bullying concept since it can reach many people quickly. Besides, removing online documents that include such harmful content can be time-consuming or impossible in some cases. Although online bullying does not seem to cause any direct physical harm to the victim, it can potentially cause psychological disorders such as depression, loss of self-confidence, lack of concentration, and even suicide attempts (Hosseinmardi et al., 2014).

Over the past decade, cyberbullying has become a widespread problem affecting especially children and youngsters. In a recent research study (Cook, 2020) in which an

interview was conducted by asking parents whether their children being cyberbullied, the percentage of children being bullied in different three years is presented in Fig. 1.1. The obtained results show that this problem is growing rapidly and independent from the development level of the countries. For instance, in Sweden, accepted as one of the developed countries, the cyberbullying rate has reached a critical point by steadily increasing between 2011-2018.

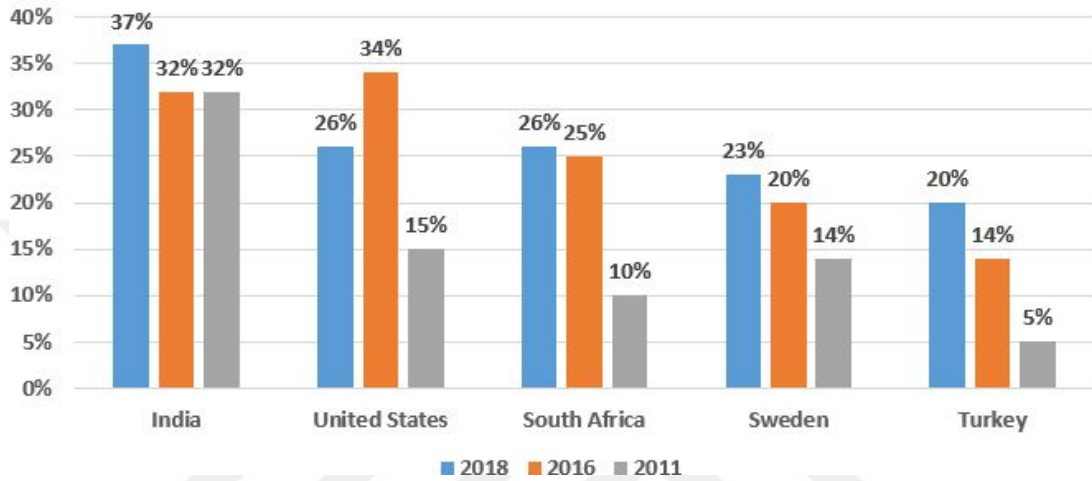


Figure 1.1 The percentage of children being cyberbullied

The awareness of cyberbullying is increased in many countries due to its effects explained in this section. Accordingly, many researchers (e.g. Yin et al. (2009), Reynolds et al. (2011), etc.) presented studies using machine learning techniques to detect cyberbullying automatically. Nevertheless, most of the research in this field is conducted for the English language. Additionally, the conducted studies generally used text mining techniques similar to the studies of sentiment analysis. The fact is that social media posts are interactive and context-dependent; hence they should not be considered a standalone text (Modha et al., 2020).

One of this thesis motivations is whether the relevance between cyberbullying and social media variables (e.g., the number of users who favorite/like document) can increase the success of detecting online bullying documents. In this direction, the objective is to test the classification of online bullying contents automatically by using social media variables with classical text mining approaches. Accordingly, a big dataset consisting of many social media and text features (Bozyiğit et al., 2021) was

created for the Turkish cyberbullying content.

The reason for collecting the Turkish dataset is that there are only a few studies about cyberbullying detection in Turkey. However, the rate of online bullying has reached 20% by increasing 300% between 2011 and 2018 (according to (Cook, 2020)). Providing a comprehensive dataset can encourage other researchers to study this problem and raise awareness about this issue in Turkey. Despite the dataset, this study is a global research since any researcher can test the importance of social media features on the published dataset and apply this study's motivation in their future works.

Lastly, it is clear that the detection of cyberbullying is a real-life problem, and so the provided works in this field should be usable-integrable by real-time applications. In the literature, there are many machine learning studies in the field of cyberbullying detection, and some of these studies publish their sources (e.g., datasets and scripts) online. Although publishing data sources and codes are valuable in the literature, integrating these codes into real-time applications is ignored. Accordingly, most of the existing studies in data science remain theoretical or pseudo-code. Hence, providing an open web service in this field is another important motivation for this thesis.

1.3 Organization of the Thesis

In the first chapter of the thesis, the thesis is briefly introduced. Next, the background of cyberbullying and machine learning processes are presented in the second chapter. In the third chapter, related works about the automatic detection of cyberbullying are presented. In the fourth chapter, the conducted data preparation steps, including data collection by developing a software system and labeling process using a crowdsourcing approach, are described. Then, the detailed analysis of features in the prepared dataset is explained in the fifth chapter. In the sixth chapter, the experimented machine learning techniques and their results are presented. In the

seventh chapter, the published open web service details for online bullying detection are provided. Finally, conclusions and future works are presented in the last chapter.



CHAPTER 2

BACKGROUND

In this chapter, the background of the study is explained. Firstly, the definition of cyberbullying, its consequences and ways to prevent these actions are detailed in Section 2.1. Then, machine learning that is commonly used to classify harmful contents is presented in Section 2.2.

2.1 Cyberbullying

Cyberbullying has become a widespread problem affecting especially children and youngsters. It is a form of psychological violence that can be defined as all of the behaviors of individuals or groups such as humiliation, slander, gossip, harassment, threats, exclusion, and insulting in an electronic or digital environment to cause discomfort or harm to others (Bauman et al., 2013). The rapid spread of sharing on the internet further aggravates this psychological violence. This violence can be realized in various ways as follows.

- Sending insulting, sarcastic, angry, vulgar, sexually abusive, or violent messages to others in online environments such as social networks or chat rooms.
- Sharing a person's personal information on the internet without his consent or knowledge.
- Spreading gossip about someone on social networks or sharing private life issues with everyone.
- Preparing defamatory and derogatory web pages related to a person.
- Opening a fake account on behalf of someone else and impersonating this person.
- Following all accounts of a person on social networks and making negative comments on their posts.

The defined types of cyberbullying actions are realized on different platforms. The first platforms in which cyberbullying is observed frequently are popular social networks such as Twitter, Facebook, Instagram, and Snapchat. Other platforms where cyberbullying is most common: text messages (SMS) sent directly via devices, email providers, chat rooms, games, and broadcasting applications.

The details of cyberbullying are presented in the following subsections. Firstly, the consequences of cyberbullying on victims are detailed in Subsection 2.1.1. Then, some true stories about online bullying are presented in Subsection 2.1.2. Lastly, the ways for preventing and detecting these malicious actions are presented in 2.1.3.

2.1.1 The Consequences of Cyberbullying

The consequences of being bullied on online platforms are both emotional and behavioral. The emotional consequences are anger, anxiety, sadness, and disappointment generally. These feelings affect the victim in every aspect of their life, such as school life, friends, and family relationships. Some victims of cyberbullying state that they have lost their sense of trust towards their friends and feel lonely.

The emotional effects of cyberbullying can be observed in schools mainly. These negative effects in the educational life of victims are indifference to lessons and low concentration. These emotional consequences are also vital signs for teachers and parents to detect cyberbullying.

The behavioral consequences of cyberbullying are mostly telling friends these actions or escaping from school. Moreover, emotions such as anger and frustration caused by cyberbullying may lead some students to seek revenge. On the other hand, victims also can be furious or uneasy in their home, and there would be constant changes in their mood, sleep, and appetite patterns.

2.1.2 Real Life Cyberbullying Stories

According to the (Caulfield, 2012), a ten-year-old girl was bullied in her school so that she requested her mother to change the way of her education. She was unwilling to go to school to avoid the scary kids, but her mother refused this request. Moreover, bullying was keep going home via social network applications. People around the little girl had called her "fat" and "ugly" firstly, and later they called her "bitch" which she did not even know its meaning. At the end of the story, the victim's sister found her hanged with a scarf and destruction left to the victim's family, especially his sister.

A kindhearted boy who had a good-looking girlfriend was also the target of online bullying attacks (Mendoza, 2016). He was getting insulting and harmful messages from cyberbullies in his school, which made him suffer from depression. As a result of these bullying actions, the boy lost his life energy and excitement in anything. Consequently, he hung himself to avoid these attacks at fifteen years old.

A ten-year-old primary school student in the USA committed suicide after a video of her fighting with a child was shared on social media (Osborne, 2017). The victim's mother stated that her daughter received support in rehabilitation for two weeks after the cyberbullying incident. Two weeks later, the little girl came home from the rehabilitation center and committed suicide by hanging herself.

A famous 14-year-old girl exposed to cyberbullying in Australia ended her life by committing suicide (Knox, 2018). The death of the child, who is the advertising face of the famous hat brand in the country, shocked the country. The victim's family called that "people, who think that what they say is in some way a joke or that they can be superior to others by harassment and bullying, please come to our funeral and see the outcomes of your actions".

The final story is about a professional wrestler and star of a series on Netflix, died because of being exposed to cyberbullying (Diaz, 2020). She is believed to have committed suicide based on the note that she left behind. Before her death, she posted a series of content on her social media account that implies that she was cyberbullied.

In her last post, she wrote "goodbye" by sharing her photo with her cat on Instagram story.

After all, it is seen that cyberbullying affects many people, especially children, regardless of victims' features such as hometown, language, race, fame, physical appearance. In addition, many other real-life cyberbullying stories can be accessed on digital platforms. Accordingly, cyberbullying should be accepted as a dangerous and widespread disease and should be acted on with this awareness.

2.1.3 The Ways of Preventing and Detecting Cyberbullying

Individuals, parents, and authorities should try to prevent online bullying in the first place before it happens. As an individual, firstly, limiting access to accounts on social networks to people they do not trust would be a proper solution. Additionally, people should not share their personal information on social networks that are accessible to others. Despite these precautions, if an individual is being cyberbullied, they should immediately share it with another person they trust. On the other hand, individuals should not help to the spreading of posts that will offend someone. They should remember their responsibility to respect the rights of others when using their right to access the Internet. Lastly, people should be aware that every step they take in the digital world has real-life consequences such as arresting.

Parents should specify rules for their children to use computers, mobile phones, and other technological tools. For instance, defining which sites they can visit and which can not be approved should be a good rule. Additionally, parents should explain the behaviors they disapprove of internet usage along with the reasons. Moreover, they should keep tracking their children to adopt the defined rules since it is crucial to ensure that they are adopted to these rules. On the other hand, parents should tell their kids to avoid posts that could harm themselves or others and guide them on the public content they share.

Authorities also play an essential role in raising awareness of students about

cyberbullying. Internet usage should not be seen as out of school activity, and the responsibility should not be left to parents only. Authorities should take steps to make it easier for students to share their cyberbullying experiences with school staff. On the other hand, governments should develop an understanding that encourages proper internet use, not prohibitive. Accordingly, the adverse effects of cyberbullying and its consequences should be presented in many resources such as short movies, reports, conferences, etc.

Even if the efforts to prevent cyberbullying are carried out effectively, some people still continue their bullying activities. In this case, it is crucial to detect cyberbullying content on online platforms quickly. Generally, the detection process is determined by reporting harmful posts to the related social network application. The moderators of these applications check the reported posts and remove the posts that include harmful content. Unfortunately, the defined process can take more than one day, and the cyberbullying contents can be spread out to many people in a short time. At this point, the importance of automatic cyberbullying detection systems emerges. Generally, machine learning techniques are used to detect harmful contents automatically. The background of machine learning for classifying content is presented in the following section.

2.2 Machine Learning

Machine learning is a field of study on computer algorithms that learn and evolve automatically through experience (Mitchell et al., 1997). In general, machine learning is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data," without explicitly programming to make decisions or make predictions. These algorithms are applied in various applications (e.g., spam filtering, sentimental analysis, image recognition, product recommendation) where it is difficult or impossible to develop traditional algorithms. Generally, the machine learning algorithms are categorized as given in Fig. 2.1.

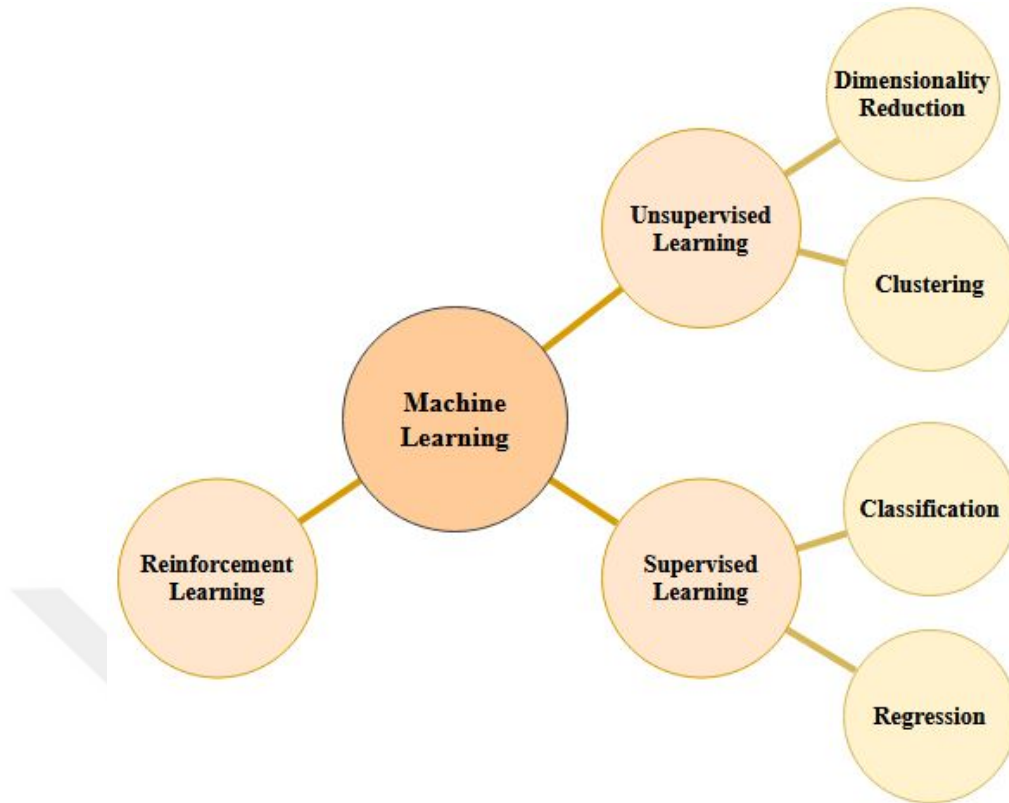


Figure 2.1 The categories of machine learning algorithms

Machine learning approaches are divided into three categories based on their differences in the learning phase (Li, 2017). These categories, reinforcement learning, unsupervised learning, and supervised learning, are detailed in the following subsections.

2.2.1 Reinforcement Learning

Reinforcement learning has a similar learning approach to a human who learns from their mistakes and uses their reasoning and thinking more efficiently to avoid making the same mistake again (Sutton & Barto, 2018). In this type of (unsupervised) machine learning, a reward-punishment method is used to teach an AI system. The machine is rewarded if it makes the right move and punished if it makes a mistake. The goal here is to maximize the total reward. The general architecture of reinforcement learning is presented in Fig. 2.2.

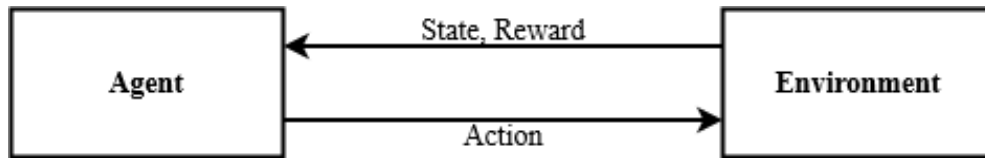


Figure 2.2 The architecture of reinforcement learning

In a reinforcement model, the programmer should assign positive values to desired actions and negative values to undesired actions to encourage the machine to take proper actions for given states (Sutton & Barto, 2018). Thus, the machine is programmed to maximize the reward long-term to reach an optimum solution. In this way, the machine learns to do the right thing by learning from its mistakes without requiring human supervision.

Reinforcement learning is commonly used in the fields of artificial game intelligence, robot navigation, and real-time decision systems. Additionally, the finance sector and inventory management are other fields where reinforcement learning can be utilized.

2.2.2 Unsupervised Learning

Unsupervised learning is a machine learning technique where there is no need to supervise the model (Barlow, 1989). Instead, the model runs on its own to discover the information or find any unknown pattern in data. Additionally, these learning algorithms are used for data categorization. There are two subcategories of unsupervised learning that are clustering and dimensionality reduction.

Clustering is essential in unsupervised learning, and data scientists commonly use it. Generally, these algorithms deal with finding a structure or pattern in a collection of uncategorized data (Xu & Wunsch, 2005). It processes the given data and finds natural clusters (groups) if there are any.

The other unsupervised approach, dimensionality reduction, is also a very important method for data science. The first reason behind this importance is that real-life data

has many dimensions (attributes), and as the size grows, the time and resources need more time for processes from data cleaning to model building increase (Wang & Sun, 2015). Additionally, there is a high correlation between some features in almost every data set, which can cause unnecessary information and overfitting problem in models.

Some unsupervised machine learning techniques used in applications are clustering the dataset into groups based on their similarity and anomaly detection that discovers unusual data points in the given dataset. Additionally, dimensionality reduction is used for compressing data sources in the projects. Moreover, there is association rule mining identifies sets of items that often occur together in the given dataset.

2.2.3 Supervised Learning

Supervised learning is a machine learning technique where the algorithm learns the mapping function from the given input and output pairs (Hastie et al., 2009). The flow of supervised learning is presented in Fig. 2.3.

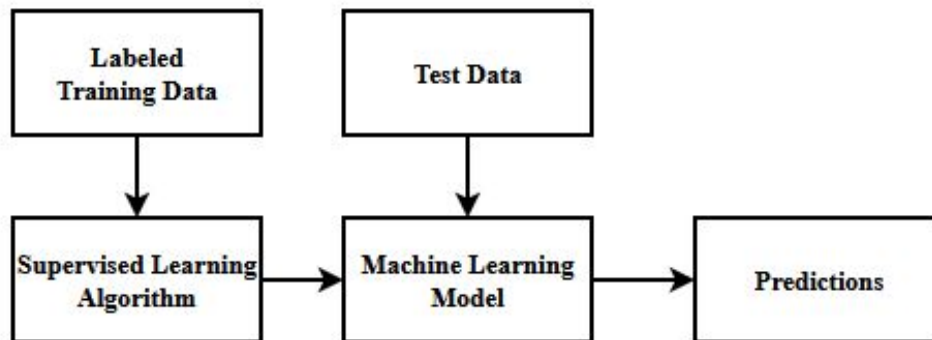


Figure 2.3 The flow of supervised learning

In the flow of supervised learning, a supervised algorithm is trained with labeled data in the initial step. After the training process, the algorithm produces a prediction model for the given problem. The performance of the produced model is evaluated with test data. For each test data, the produced model makes a prediction where the accuracy shows the algorithm's performance.

Supervised machine learning consists of both classification and regression

algorithms. In a classification problem, the output variable is a categorical value (nominal) such as spam or not-spam, so the model is trained with these categorical values to classify new mail (Nasteski, 2017). In a regression problem, the output variable is a continuous value (numeric) such as product price, so the model is trained with these numeric values to approximate new product prices (Hastie et al., 2009).

Both regression and classification algorithms are commonly used in the sentimental analysis, spam filtering, and cyberbullying detection since the class of datasets in these problems are categorical values that need to be supervised. Note that regression algorithms can also be used in classification problems, but vice versa is not valid. In this thesis, algorithms of these supervised learning approaches are used for the automatic detection of cyberbullying. The details of the used algorithms are detailed in Chapter 6.

CHAPTER 3

RELATED WORK

In this chapter, related work in the field of cyberbullying detection is presented. In the first section, the preliminary studies are presented. The studies using only text mining techniques are given in the second section. In the third section, studies using contextual and interaction features are explained. Turkish online bullying detection studies are presented in the fourth section. Finally, a general overview of the related work is given in the last section.

3.1 Preliminary Works

In considering the literature reviewed, it is seen that the first study conducted for the automatic detection of cyberbullying was presented by Yin et al. (2009). In the work, feature extraction techniques (N-grams and Term Frequency-Inverse Term Frequency (TF-IDF)) and a linear kernel classifier were applied on three different data sets to detect harassment on online platforms. One of the datasets was obtained from a chat-style platform (Kongregate) while the others were obtained from discussion-style communities (MySpace and Slashdot). Although the experimental results are not sufficient, the study was motivating for further research.

The following research in this field was presented by Reynolds et al. (2011). The authors experimented with C4.5, k-nearest neighbors (kNN), support vector machine (SVM) classifiers on a dataset consisting of the user comments on Formspring.me. Evaluation results show that the C4.5 decision tree algorithm outperforms both kNN and SVM classifiers with a 78.5% accuracy rate.

In another preliminary study, Dinakar et al. (2011) proposed a two-step detection approach. In the first step, it is determined whether content belongs to a sensitive subject or not. Then, in the second step, the content is assigned to a specific topic (e.g., intelligence, sexual) using a multi-class classifier. The proposed approach experimented with 4,500 YouTube comments and the accuracy of classification is

between 70% and 80%.

Dadvar et al. (2012) presented research using a gender-based approach. In the approach, two different vocabularies (feature sets) were used for two genders. In the result of the study, it was observed that the gender-specific method slightly improves the accuracy of machine learning techniques.

3.2 Cyberbullying Detection Using Text Mining

After the preliminary works, many studies applying different techniques were presented in this field. Kontostathis et al. (2013) designed a model to identify the most commonly used cyberbullying terms on the collection of posts from Formspring.me (question-and-answer-based social network) using the Essential Dimensions of Latent Semantic Indexing (EDLSI). The experimental results of the study show that the model achieves an average precision of 91.25%.

Ptaszynski et al. (2015) conducted a study based on an approach consisting of brute force search algorithms and learning classifiers. The proposed method extracts patterns from sentences and uses them in the step of classification. It is stated that the method outperforms previous methods on the dataset provided by the Human Rights Center.

Zhang et al. (2016) aimed to develop a universal cyberbullying detection model with high performance using deep learning methods. They used the pronunciation of words within the input texts as the features for a Convolutional Neural Network (CNN) model. The CNN model was tested on the social media posts collected from Twitter and Formspring.me. The accuracy scores show that pronunciation-based CNN outperforms baseline CNN with randomly generated word embedding.

Romsaiyud et al. (2017) proposed an automatic online bullying detection system that uses a clustering approach with Naive Bayes (NB). In the proposed method, abusive contents are clustered using K-Mean, and then NB is used for classifying a post into the clustered contents. In the experimental results, it is stated that the

proposed system can successfully classify contents into categories that are obtained from clusters.

Haidar et al. (2017) conducted a solution using machine learning and natural language processing techniques for the detection of Arabian contents. In the study, the dataset consisting of Arabian posts was collected from Twitter and Facebook. NB and SVM experimented on the created dataset, and the F-measure scores of the experimented classifiers are %90 and %92, respectively.

Rosa et al. (2018) compared the capabilities of Fuzzy Fingerprints (FFP) and well-known machine learning methods (Logistic Regression, NB, and SVM) to specify cyberbullying instances in unbalanced datasets. It was concluded that FFP outperforms baseline classifiers with 0,42 F-measure score, which is 1% better than the second-best performing classifier (SVM).

Febriana & Budiarto (2019) proposed a hate speech detection model for the Indonesian Language. For this study, the authors collected more than one million tweets by using Twitter API. After the preprocessing steps on the dataset, Latent Dirichlet Allocation (LDA) was conducted for extracting topics from the dataset. Five clusters were obtained at the end of the experimental study; however, the authors state that they can not specify these clusters to meaningful categories.

3.3 Cyberbullying Detection Using Contextual Features

Literature search shows that the current studies on cyberbullying detection mainly use textual content written by the users as a base input. However, there are other essential features in addition to social media comments. Consequently, some researchers address the direction for incorporating user information (e.g., age, gender, previous comments) in detecting cyberbullying content.

Dadvar et al. (2014) proposed a hybrid model on a dataset that includes YouTube comments and user features (number of comments, uploads, and subscriptions). The

authors compared the performance of an expert system (Multi-Criteria Evaluation System (MCES)), three classifiers (NB, C4.5, and SVM), and the hybrid model. According to evaluation results, the hybrid approach (MCES+NB) with a discrimination capacity of 0.76 outperformed the other experimented classifiers and baseline MCES.

Al-garadi et al. (2016) proposed a cyberbullying detection method by identifying characteristic features (e.g., user ID, username, user biography). In the study, various classifiers were applied to the collected data from Twitter. The experimental results show that random forest (RF) using synthetic minority oversampling technique (SMOTE) obtains the best performance with 93.6% F-measure score.

Escalante et al. (2017) used profile-based (aggressive/non-aggressive) models for early recognition of aggressiveness in written content. In this study, class term-occurrence information was used to obtain a non-sparse, discriminative model for documents, where profiles can be divided into sub-profiles such as sexual predator, or aggressive text. The experimental results show that sub-profile-based representations outperform profile-based representations.

Dadvar & Eckert (2018) conducted a study based on deep learning-based models. The authors implemented various deep learning models such as Long Short-Term Memory (LSTM), Convolutional Neural Network, Bidirectional LSTM, etc. The implemented methods experimented on the YouTube dataset consists of 54,000 comments with sender demographic information. In the experimental results, a 64% F-measure score was obtained, which was increased to 76% by utilizing expert knowledge.

Cheng et al. (2019a) highlight the effects of characteristics embedded in the user-generated content on cybercrime detection. They performed kNN, RF, SVM, and Logistic Regression on the dataset crowded via the Twitter streaming API. Moreover, they proposed a new model (PI-Bully) describing the users' idiosyncrasies to improve the prediction of cyberbullying behaviors. According to evaluation results, it is seen that the proposed model achieves the best performance with over 80% recall.

In their following study, Cheng et al. (2019b) study the cyberbullying problem within a multi-modal context (XBully) by collaboratively exploiting social media data. Thus, the authors created an experimental dataset that includes various social media resources such as image, video, user profile, time, and location. Experiments on multi-modal social media datasets (e.g., Instagram) show that the proposed approach performs better than the state-of-the-art cyberbullying detection models.

3.4 Cyberbullying Detection for Turkish

The majority of the reviewed studies are designed to analyze only texts written in English, and only infrequent researches are conducted for different languages. The prominent research for the Turkish language is presented by Özel et al. (2017). In the study, streaming data is collected from the micro blogging platform, Twitter, to create an experimental dataset. They applied the bag of words method to form vectors for each tweet and experimented with various machine learning algorithms (C4.5, NB, SVM, and kNN) to determine whether the user tweets are harassing or not. It is concluded that NB outperforms other classifiers with a 79% accuracy rate in terms of F-measure.

The further study that uses Twitter as a data source and produces a Turkish dataset for cyberbullying detection is presented by Bozyiğit et al. (2019). Researchers evaluated a wide range of neural network models. Before performing the algorithms, the authors applied information gain feature ranking to decrease the feature space dimension while eliminating irrelevant words. After feature selection, the performance of the algorithms in terms of F-measure reaches 91%.

3.5 General Overview

The summary of reviewed studies in this chapter is presented in Table 3.1.

Table 3.1 The summary of reviewed studies

Works	Supported Language	Datasets	Features	Proposed Method
Yin et al. (2009)	English	Kongregate, MySpace, Slashdot	Textual, Sentimental, Contextual	Linear Kernel Classifier
Reynolds et al. (2011)	English	FormSpring.me	Textual	C4.5
Dinakar et al. (2011)	English	Youtube comments	Textual	LinearSVM
Dadvar et al. (2012)	English	MySpace	Textual, User-based	SVM
Kontostathis et al. (2013)	English	FormSpring.me	Textual	EDLSI
Ptaszynski et al. (2015)	Japanese	Human Rights Center	Textual	Pattern Extraction
Zhang et al. (2016)	English	Twitter, Formspring.me	Textual	CNN
Romsaiyud et al. (2017)	English	Twitter	Textual	Clustering+NB
Haidar et al. (2017)	Arabian	Twitter, Facebook	Textual	SVM,NB
Rosa et al. (2018)	English	Formspring.me	Textual	FFP
Febriana & Budiarto (2019)	Indonesian	Twitter	Textual	LDA
Dadvar et al. (2014)	English	YouTube comments	Textual, User-based	MCSE+NB
Al-garadi et al. (2016)	English	Twitter	Textual, User-based	RF+SMOTE
Escalante et al. (2017)	English	PAN'12, Kaggle, UANL	Textual, User-based	Neural Network
Dadvar & Eckert (2018)	English	YouTube	Textual, User-based	Deep Learning
Cheng et al. (2019a)	English	Twitter	Textual, User-based	PI-Bully
Cheng et al. (2019b)	English	Instagram, Vine	Textual, Collaborative	XBully
Özel et al. (2017)	Turkish	Twitter	Textual	NB
Bozyiğit et al. (2019)	Turkish	Twitter	Textual	Neural Network

According to the literature survey, it is observed that there is no study, having a public dataset, that directly analyzes the importance of social media features for cyberbullying detection. Thus, a dataset with many features was prepared for cyberbullying by collecting thousands of Turkish tweets. Then, the relationship between social media features and online bullying events was analyzed using the created dataset. Lastly, the determined features were used to improve machine learning algorithms. The reason for collecting Turkish Tweets is to raise awareness in

Turkey. However, it is important to remark that this study should be considered global research since other researchers would test the importance of social media features on the published dataset and apply this study's motivation in their studies.



CHAPTER 4

DATA PREPARATION

Creating a comprehensive dataset that includes many social media contents and their features is an essential part of this thesis for analyzing the relationship between online bullying events and social media features. If the analyzed relationship is strong enough, the social media features in the dataset can increase the cyberbullying detection performance of machine learning algorithms. Besides, there is a lack of quality cyberbullying datasets that have building and annotation process details (Rosa et al., 2019). Therefore, making this generated dataset with building and annotation process details publicly available would be a valuable contribution to the literature.

In this study, the data preparation process consists of three steps; data collection, data elimination, and data annotation. In the first step, a software system was developed to collect data regularly. After the data collection process was completed, redundant data was eliminated. In the last step, data annotation was carried out by using the developed crowdsourcing web application. After the data preparation, a balanced dataset consisting of 5,000 labeled Turkish contents with many social media features was prepared for cyberbullying detection. This unprocessed dataset with labels was publicly published in Comma Separated Value format (Bozyiğit et al., 2021) . The steps of the data preparation process are detailed in the following sections.

4.1 Data Collection

The software application was developed in the n-tier design architecture to collect data regularly from the official Twitter Application Programming Interface (Twitter, 2020). The developed software consists of three tiers called Scheduled Task, Api Controller, Cyberbullying Context. The diagram of the software architecture is given in Fig. 4.1.

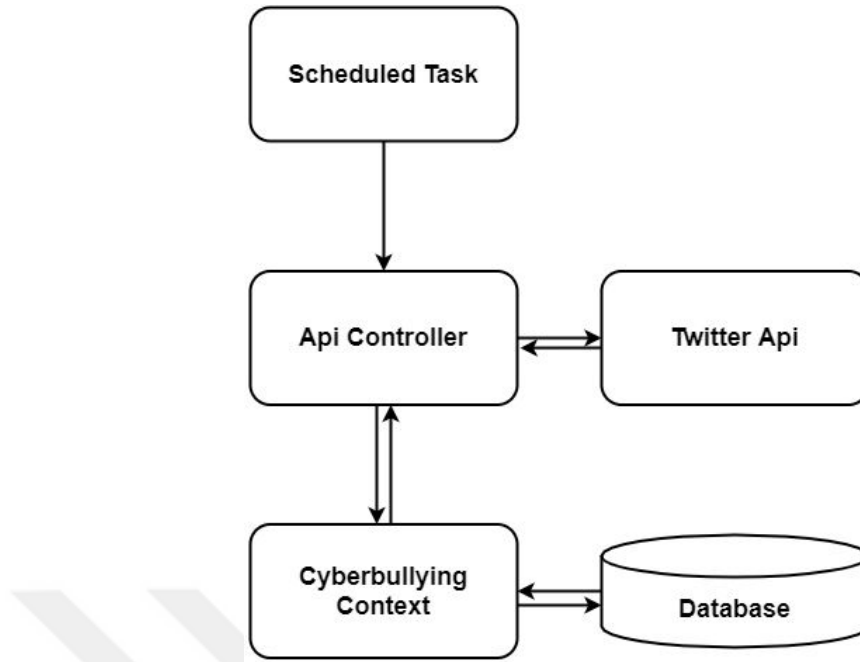


Figure 4.1 The data collection software system

Scheduled Task is a console application that was developed for running it regularly on a Windows operating system as a task. The task was set to run every 15 minutes for the defined days. The only mission of this task is calling *Api Controller* periodically.

Api Controller is a class library developed for collecting data from official Twitter Api and processing the collected data. This class library mainly consists of two methods. The first method is *SetCrediantals* which establishes a connection between the client and the official Twitter Application Programming Interface. The second method is *CollectRecentTweets* that collects Turkish Tweets posted in the last fifteen minutes from Twitter API. Additionally, *CollectRecentTweets* processes the collected data and sends them as models (User, Tweet, Mention) to *Cyberbullying Context*.

Cyberbullying Context, a class library using entity framework (Lerman, 2010), was developed to retrieve and return models from the relational database. The Entity-Relationship (ER) diagram of the database is presented in Fig. 4.2.



Figure 4.2 ER diagram of the database

In the ER diagram, there are three entities such that Tweet, User, and Mention. There is a one-to-many relationship between User and Tweet entities since each tweet has one sender (Twitter user) and one user can have multiple tweets. On the other hand, there is a many-to-many relationship between Mention and other entities since one tweet can mention many users and one user can be cited by many tweets. These entities have many attributes that are used in the machine learning process, and these attributes are detailed in the chapter of Feature Analysis (Chapter 5).

4.2 Data Elimination

Creating a balanced dataset that contains almost equal numbers of samples from cyberbullying and non-harmful posts is one of this study's objectives. However, most of the contents (tweets) in the collected dataset are negative (non-harmful) under normal circumstances. Consequently, extra effort would be required for creating a balanced dataset from the collected imbalanced data. Therefore, data elimination was applied to reduce the working hour in the annotation step. The process of data elimination is presented in Fig. 4.3.

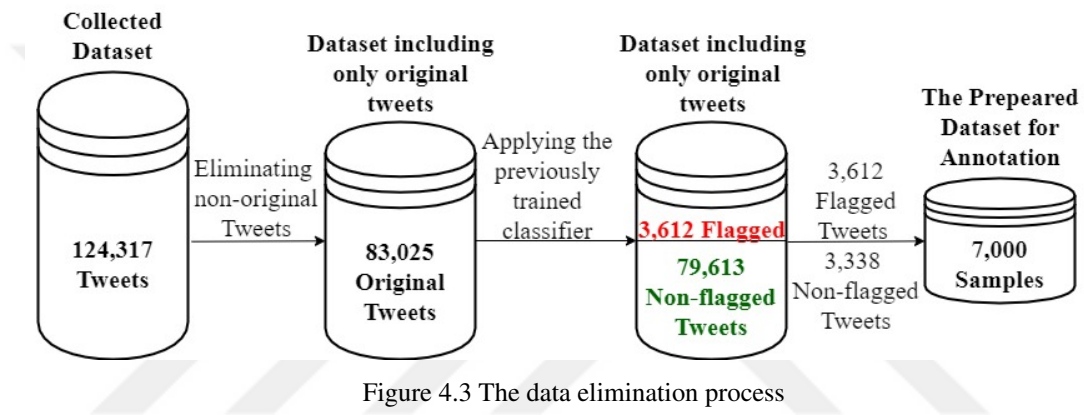


Figure 4.3 The data elimination process

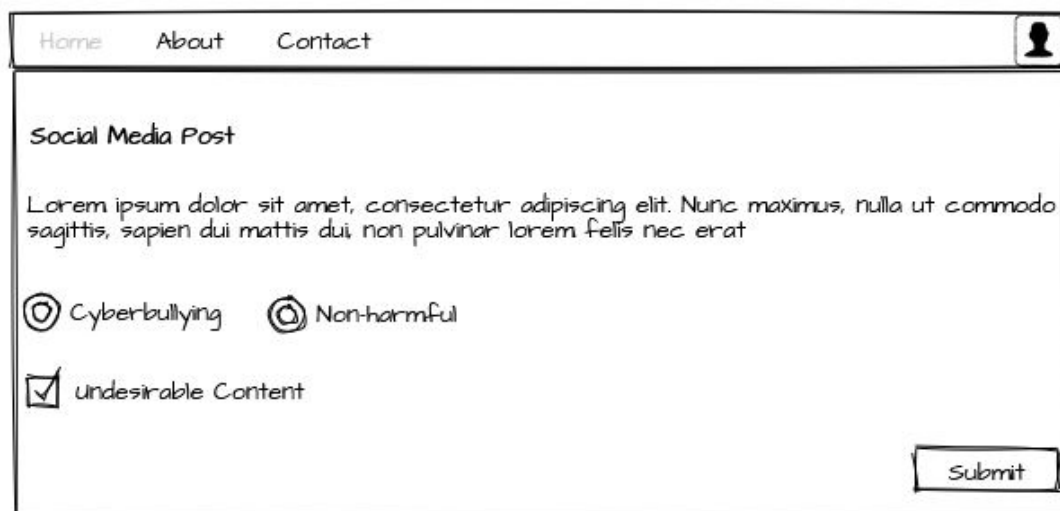
Firstly, approximately one-third of the collected tweets which are non-original (retweet or spam) were eliminated. Then, the original 83,025 tweets were flagged in terms of cyberbullying by applying the previously trained classifier proposed in our previous work (Bozyiğit et al., 2019).

According to the used classifier, 3,612 tweets were flagged as potential cyberbullying content. Then, nearly 3,400 random non-flagged tweets with 3,612 flagged tweets were kept, and the rest of the samples were eliminated. Thus, the dataset size was reduced to 7,000 records, and the number of data to be examined was significantly reduced.

4.3 Data Annotation

Data annotation is defined as the process of labeling records in a dataset to make the dataset usable for supervised machine algorithms (Zampieri et al., 2017). The annotation process is a repetitive but crucial activity that requires attention since mislabeling affects the created dataset's quality and the applied machine learning algorithms' performance. Thus, a crowdsourcing approach was used for data annotation in this study to conduct this process more reliable and faster.

In this direction, a web application where multi-user can label records simultaneously was developed by using the Model-View-Controller pattern (Pop & Altar, 2014). The mock-up user interface of the published web application is presented in Fig. 4.4.



The mock-up user interface is a web browser window. The top navigation bar contains links for 'Home', 'About', and 'Contact', along with a user profile icon. The main content area is titled 'Social Media Post' and displays a block of placeholder text: 'Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc maximus, nulla ut commodo sagittis, sapien dui mattis dui non pulvinar lorem felis nec erat'. Below the text, there are two radio button options: 'Cyberbullying' and 'Non-harmful'. A checkbox labeled 'Undesirable Content' is also present and is checked. A 'Submit' button is located in the bottom right corner of the form.

Figure 4.4 The mock-up user interface of the data annotation application

When a registered user (annotator) logs in to the developed web application, they are redirected to the home page. On the home page, random tweets (from the *Cyberbullying Context*) which were not annotated by the user yet are presented. Accordingly, the user determines whatever the presented tweet is cyberbullying or not. Users can also mark a given content as undesirable to prevent adding uncertain or political issues to the dataset. The activity diagram of the user annotation process is presented in Fig. 4.5.

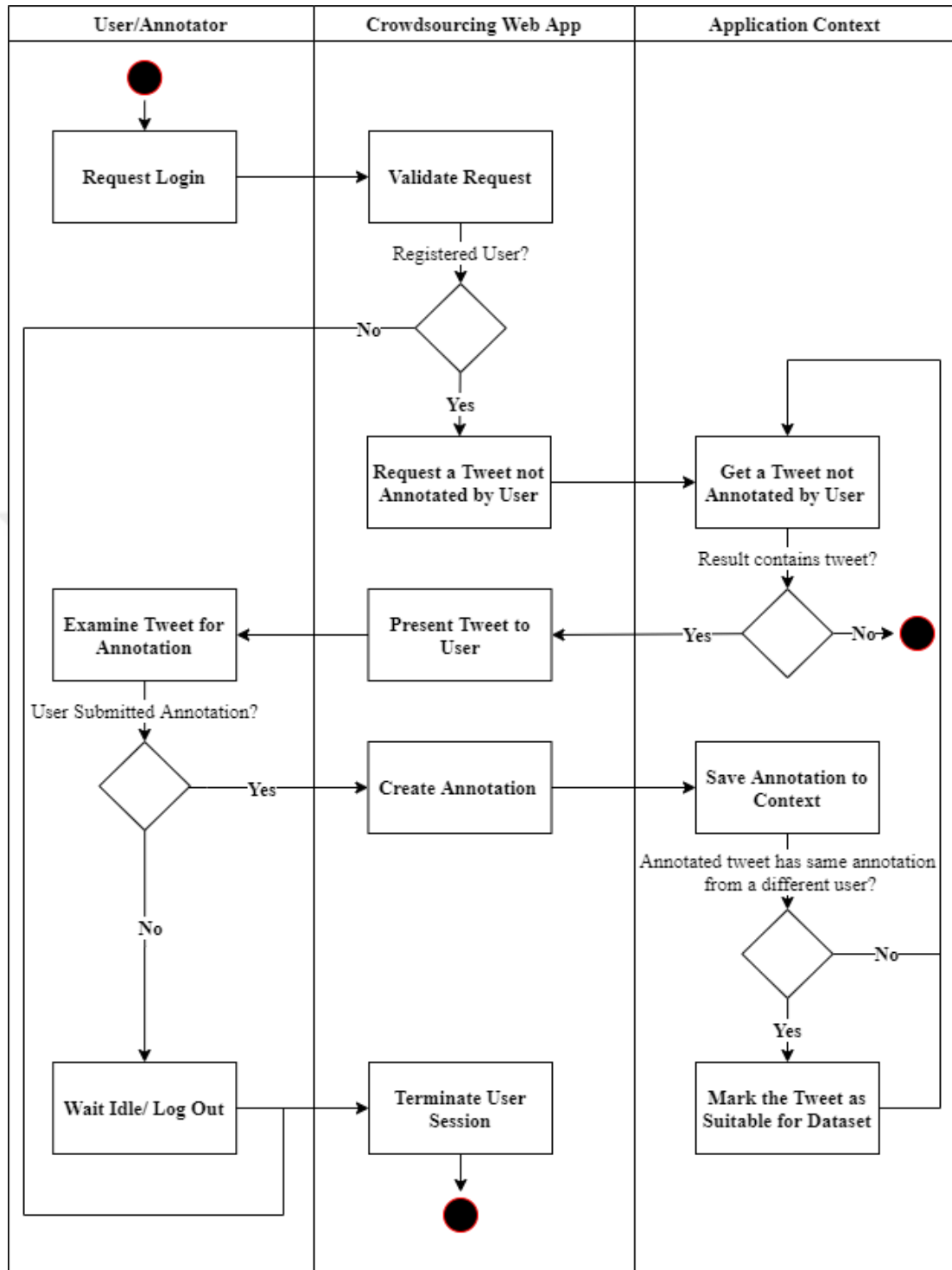


Figure 4.5 The activity diagram of user annotation

The application context in Fig. 4.5 is the database layer of the crowdsourcing application. The ER diagram of the application database is presented in Fig. 4.6.

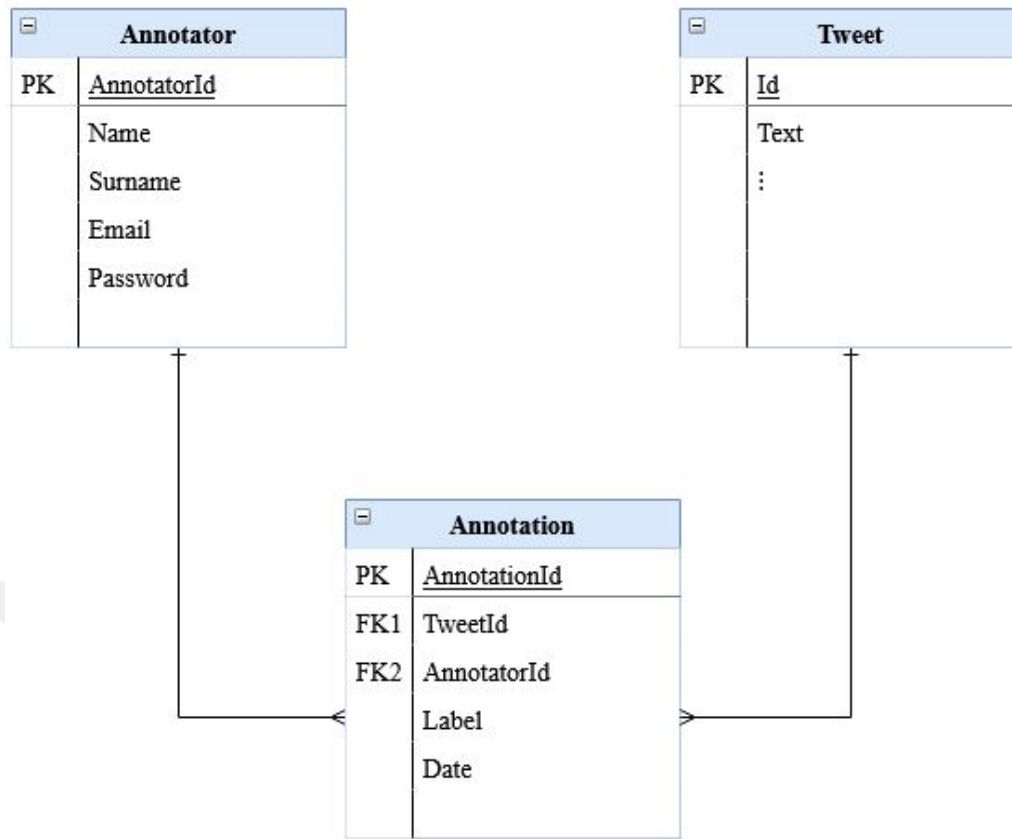


Figure 4.6 ER diagram of the web application's database

In the ER diagram, there are three entities that Annotator, Annotation, and Tweet. Note that Tweet is the entity obtained from *Cyberbullying Context* and Annotator entity keeps the registered users for this study. There is a one-to-many relationship between Tweet and Annotation entities since each annotation is about one Tweet, and one Tweet can have various annotations. Similarly, there is a one-to-many relationship between "Annotator" and "Annotation" entities since a user can conduct multiple annotations, and an annotation record is unique about a tweet and a user.

In the annotation process, it is important to provide the specified instructions to annotators, and also ensure that the annotators are experts in the field of cyberbullying (Rosa et al., 2019). Thus, three data scientists with a master's degree from computer science-related departments were registered as users/annotators to the developed crowdsourcing web application. The definition of cyberbullying was explained to these users with the examples that occurred in social networks. Then, users were asked to annotate the shown tweets by the given criteria. The annotation process was

completed when 5,000 (the desired dataset size) tweets that include at least two of the same annotations from different users were obtained. These 5,000 tweets, which include the same annotations, were selected for the dataset (Bozyiğit et al., 2021) to be used in the machine learning methods. Thus, annotations were conducted more reliably in a short time.

A sample from the prepared dataset is presented in Table 4.1. Note that the texts' English translation (Text_En) is not available in the dataset; it is just provided for information purposes only in this article. The detailed information about attributes/features is given in the following chapter.

Table 4.1 A sample from the dataset

Attribute	Type	Value
TweetId	integer	*****10770000
Text	string	<i>"bende biliyordum adi şerefsiz"</i>
Text_En	string	<i>"I know it too, you dishonorable trash"</i>
IsSelfMentioned	bool	<i>false</i>
Retweets	integer	0
Hashtags	integer	0
Medias	integer	0
Mentions	integer	2
SenderId	integer	*****2980691582000
AccountDuration	integer	2
SenderFollowers	integer	555
SenderFollowings	integer	485
SenderStatuses	integer	1937
SenderFavorites	integer	5889
SenderLocation	string	<i>"izmir"</i>
IsCyberbullying	bool	<i>true</i>

CHAPTER 5

FEATURE ANALYSIS

Feature engineering makes classification performance marginally better in current cyberbullying detection practices (Rosa et al., 2019). Therefore, this chapter directly examines which features are related to cyberbullying incidents and are helpful for machine learning methods. In the first section, social media features in the prepared dataset are presented and analyzed. Then, the determined natural language processing features for the dataset are discussed in the second section.

5.1 Social Media Features

The social media attributes of the samples (tweets) and their senders in the prepared dataset are presented with descriptions in Table 5.1.

Table 5.1 Social media features

Attribute	Description
<i>IsSelfMentioned</i>	Did the sender mentioned themselves in the post?
<i>Retweets</i>	The number of times a post was shared by other users.
<i>Favorites</i>	The number of users who liked a post.
<i>Hashtags</i>	The number of topics (special tags) in a post.
<i>Medias</i>	The number of photos, videos, etc. in a post.
<i>Mentions</i>	The number of references to the other users in a post.
<i>AccountDuration</i>	The membership duration of the sender on Twitter.
<i>SenderFollowers</i>	The number of users following the sender of a post.
<i>SenderFollowings</i>	The number of users followed by the sender of the post.
<i>SenderStatuses</i>	The number of posts shared by the sender.
<i>SenderFavorites</i>	The number of times the sender liked other posts.
<i>SenderLocation</i>	The location of the sender (shared by the sender).

To analyze social media features, firstly, A Chi-square test (McHugh, 2013) was used to test the relationship between samples' class (presenting a sample contains a harmful content or not) and their features in the prepared dataset. Then, association

rule mining processes were conducted on the created dataset for discovering interesting relations between social media features and cyberbullying events. The details of the applied Chi-Square test and association rule mining are detailed in the following subsections.

5.1.1 Chi-square Test

A Chi-square test (McHugh, 2013) was used to test the relationship between samples' class (presenting a sample contains a harmful content or not) and their features in the prepared dataset. The formula of the chi-square test is given in Equation 5.1.

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (5.1)$$

Chi-Square determines the independence of two variables by measuring how E (expected count) and O (observed count) of two variables differ from each other. Chi-Square is based on the hypothesis that two variables are independent. Accordingly, the expected probability of two variables occurring together is the multiplication of these variables' independent probabilities. If the expected value is far from the observed value, it gives a higher Chi-Square value and contradicts the hypothesis.

For instance, assume that ten percent of the dataset samples are self-mentioned. Note that the dataset is balanced where half of the records are positive (cyberbullying contents), and the other half are negative (non-harmful). In this case, the probability of a sample being positive is 0.5, and the probability of a sample being self-mentioned is 0.1. Hence, the expected count of positive self-mentioned samples is $5000(\text{dataset size}) * 0.5 * 0.1$, which is 250. According to Equation 5.1, the Chi-Square score would be zero if the observed count of positive self-mentioned records equaled 250 (expected count). Zero Chi-Square score means there is no relationship between the tested attribute and the class. On the other hand, the

difference between the observed count and the expected count violates the rule of independence and so increases the Chi-Square score. In other words, a higher Chi-Square score means there is more relation between the test attribute and the class.

Chi-square score for each social media attribute and the class of samples are shown in Fig. 5.1.

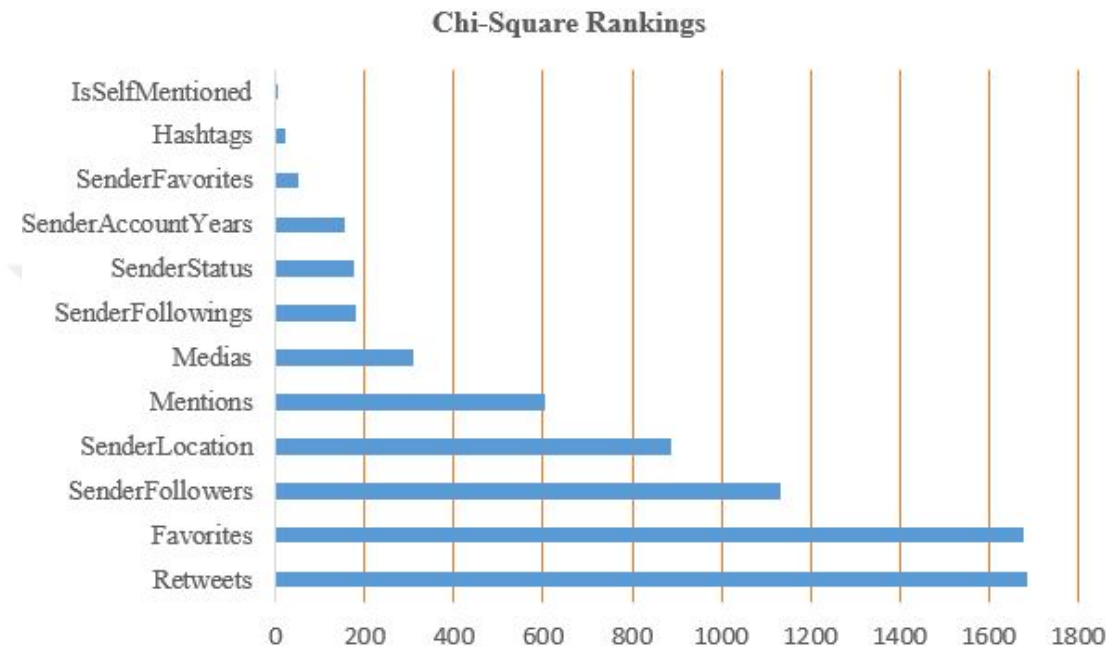


Figure 5.1 Chi-Square scores of the features

According to the Chi-square results, it is observed that samples' class is highly dependent on social media features of the samples such that *Retweets*, *Favorites*, *SenderFollowers*, and *SenderLocation*.

Furthermore, discretization was applied to the dataset for visually analyzing the relationship between social media features and cyberbullying. Discretization is a process that groups given numerical attributes into a smaller number of buckets/bins. *SenderFollowers* was discretized to ten bins with an equal size that means each bin contains nearly 500 samples. In Fig. 5.2, the number of online bullying posts and non-harmful posts are presented for each discretized bin. For instance, it is seen that the number of non-harmful posts is less than 100 in the first bucket that contains the samples whose sender has less than 24 followers. The figure shows that users that have more followers on social networks are not prone to post online bullying content.

The reason behind the inference can be senders' fear of losing their followers. On the other hand, it cannot be concluded that users that have a few followers tend to cyberbully since the dataset was balanced by eliminating most harmful posts.

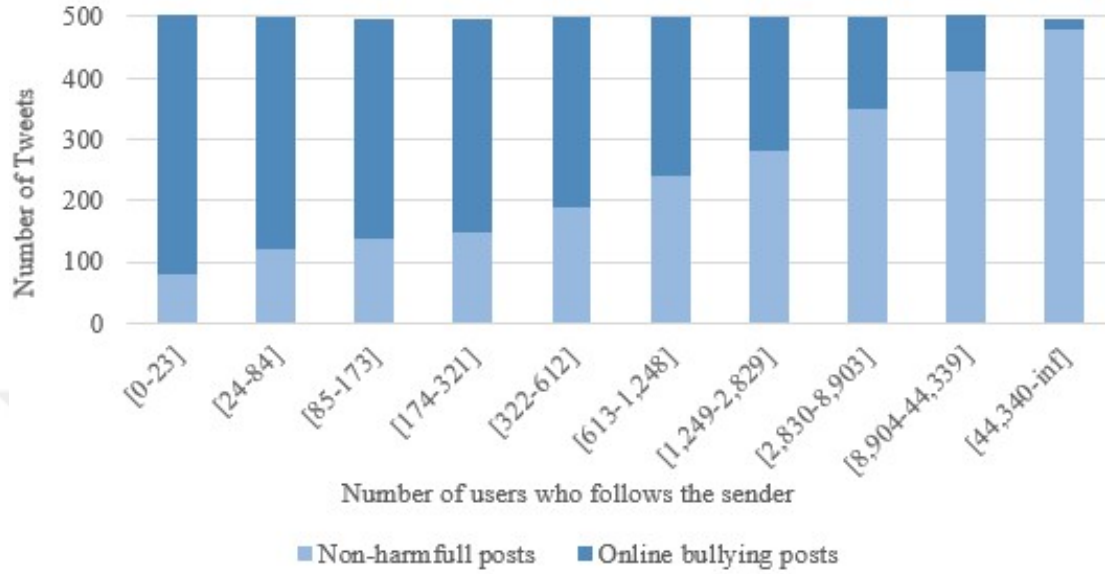


Figure 5.2 The relationship between sender followers and cyberbullying

Retweets and *Favorites* have similar distributions with *SenderFollowers*; while the number of retweets-favorites increases, the cyberbullying ratio decreases. It can be concluded that users in Turkey are mostly not prone to like or share cyberbullying content. Additionally, a sender's location has a significant relationship between cyberbullying since it was observed that cyberbullies generally share a fake location or no address in their profile. To sum up, the analysis results show that some features would be promising for cyberbullying detection. Thus, the promising features and the text of tweets were used in machine learning techniques, explained in the following chapter.

5.1.2 Association Rule Mining

Association rule mining is a machine learning method that analyzes the co-occurrence of events and discovers relationships between data as rules (Zhang & Zhang, 2003). An association rule consists of two parts: an antecedent (if) and a

consequent (then). For instance, an association rule $potato \rightarrow burger$ in shopping transactions implies that “if a customer buys potato, then he probably also buys a burger”. To extract association rules, some measures of interestingness are used such as support, confidence, lift, leverage, etc.

Assume that A and B are item sets and T is a set of records of the given dataset. The **support** ($supp$) value of the association rule $A \rightarrow B$ is the ratio of records containing A and B together to the entire dataset (Azevedo & Jorge, 2007). The formula of the support is given as follows.

$$sup(A \cup B) = \frac{|A \cup B|}{|T|} \quad (5.2)$$

The **confidence** ($conf$) value of the $A \rightarrow B$ association rule determines the accuracy of the rule (Azevedo & Jorge, 2007). It is the ratio of the records containing A to the records containing B as given in the following equation.

$$conf(A \rightarrow B) = \frac{sup(A \cup B)}{sup(A)} = \frac{|A \cup B|}{|A|} \quad (5.3)$$

Generally, support and confidence are primary metrics that are used for association rule mining. However, confidence can be misleading in some situations, such as some nominal value outnumbering other values in a feature. Hence, **lift** and **leverage** measures (Zheng et al., 2001), which take statistical dependence into account, are experimented within this study.

$$lift(A \rightarrow B) = \frac{sup(A \cup B)}{sup(A) \times sup(B)} \quad (5.4)$$

The value of leverage (lev) can be between -0.25 and 0.25. The higher leverage value means that the related rule is more interesting. A leverage value that is less than 0 is accepted as insufficient, and the related rule should not be considered. The equation of leverage is given as follows.

$$lev(A \rightarrow B) = sup(A \cup B) - sup(A) \times sup(B) \quad (5.5)$$

Before applying association rule mining processes, discretization was applied to the dataset for converting numerical attributes to nominal. Each numeric social media attribute was discretized to ten bins with an equal size. Then, association rule mining processes were conducted on the created dataset according to the defined measures of interestingness. In each process, the minimum support value was set to 5%, which means at least 250 records of the dataset should support a rule. The discovered remarkable rules with the defined measure values are given in Table 5.2 .

Table 5.2 The discovered cyberbullying rules

Id	Rule	conf	lift	lev
R1	$Cyberbullying = TRUE \rightarrow Retweets = [0, 1)$	0.86	1.48	0.14
R2	$Cyberbullying = TRUE \rightarrow Favorites = [0, 1)$	0.72	1.47	0.12
R3	$Cyberbullying = TRUE \rightarrow Retweets = [0, 1) \quad Favorites = [0, 1)$	0.72	1.49	0.12
R4	$Cyberbullying = TRUE \rightarrow SenderFollowers = [0 - 23]$	0.13	1.83	0.03
R5	$Cyberbullying = TRUE \rightarrow$ $SenderFavorites = [0 - 30] \quad SenderAccountYears = [0, 1)$	0.16	1.63	0.03
R6	$Cyberbullying = TRUE \rightarrow SenderLocation = NA$	0.09	1.57	0.02
R7	$Favorites = [572 - 2233) \rightarrow Cyberbullying = FALSE$	0.94	1.87	0.02
R8	$Favorites = (2233 - inf) \rightarrow Cyberbullying = FALSE$	0.96	1.92	0.02
R9	$SenderFollowers = [43, 340 - inf) \rightarrow Cyberbullying = FALSE$	0.96	1.92	0.05

R1, R2, and R3 were discovered from the rules in which leverage values greater than 0.10. These rules imply that cyberbullying posts generally have zero retweets and favorites. So, it is deduced that social media users in Turkey mostly do not favor and share cyberbullying actions.

R4, R5, and R6 were obtained from the rules in which lift values are greater than 1.5. It is concluded from R4 and R5 that mostly accounts that post cyberbullying content are newly created and do not have many followers. Additionally, it is seen that the sender location frequently is not shared by users who have at least one cyberbullying event, according to R6.

Lastly, *R7*, *R8*, and *R9* were captured from the rules which of confidence values are greater than 90%. According to these rules, we can say that posts having many favorites are mostly non-harmful and senders having many followers not intend to do cyberbullying actions.

Note that many rules which have Cyberbullying class in their consequent are ignored in this study. The reason behind that is that the dataset was biased by eliminating most non-harmful posts that cause an increase in the rate of positive cyberbullying posts artificially.

5.2 Natural Language Processing Features

The analyzed natural language processing (NLP) features in this section are presented with descriptions in Table 5.3.

Table 5.3 Natural language processing features

Attribute	Description
<i>Words</i>	The number of words in a post.
<i>TitleWords</i>	The number of title case words in a post.
<i>UpperCasewords</i>	The number of full upper-case words in a post.
<i>AvgWordLength</i>	Average length of the words used in the text of the post.
<i>Letters</i>	The number of letters in the post.
<i>UpperCaseLetters</i>	The number of upper-case letters in the post.
<i>PunctuationMarks</i>	The number of punctuation marks in the post.
<i>Symbols</i>	The number of symbols in the post.
<i>Emojis</i>	The number of emojis in the post.

The NLP features are calculated for each sample of the dataset by using the developed algorithm. The algorithm firstly finds out the number of letters, upper case letters, punctuation marks, symbols, and emojis by iterating over characters of the sample. In the second part, the algorithm splits the given text into terms; then, it calculates the number of words, title words, and upper-case words by iterating over

the obtained terms. Lastly, the average length of words is determined by dividing the number of letters by the number of words.

To analyze NLP features, A Chi-square test (detailed in Subsection 5.1.2) was used to test the relationship between samples' class (presenting a sample contains harmful content or not) and the generated NLP features. Chi-square score for each NLP attribute and the class of samples are shown in Fig. 5.3. Note that "MostRelatedSMF" in the figure presents the most related social media feature that is the number of retweets.

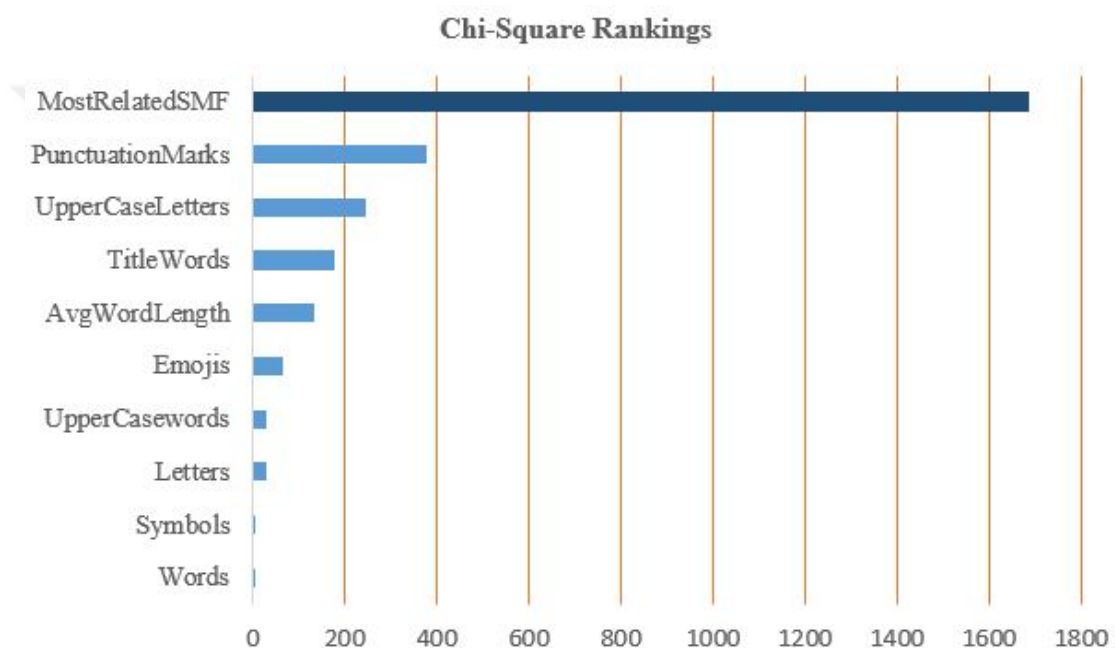


Figure 5.3 Chi-Square scores of the features

According to the Chi-square results, it is observed that samples' class is not highly dependent on any NLP features as comparison with social media features. Accordingly, it can be concluded that there would be no significant relationship between NLP features and cyberbullying in this study. The reason behind this fact would be the lack of massive data or more specific features such as the number of exclamation points etc. Because of the explained reasons, there is no further analysis for NLP features, and these features are not used in the machine learning process.

CHAPTER 6

AUTOMATIC DETECTION OF CYBERBULLYING

In this chapter, machine learning algorithms' performance on the prepared dataset was evaluated using the related social media features with text mining approaches. After performing pre-processing and feature extraction (from the text of posts) on the dataset, some supervised machine learning algorithms were applied by using parameter optimization. Lastly, the experimental results of the study were presented. The steps of the machine learning process are explained in the following subsections.

For the implementation of machine learning steps, Scikit-learn, an efficient open-source machine learning library for Python by Pedregosa et al. (2011), was mainly used. Spyder (2020), an open-source scientific environment, was chosen for Integrated Development Environment (IDE), and all tests were run on a computer equipped with 16 GB Ram and a 3.4 GHz processor with six cores.

6.1 Preprocessing

The prepared dataset consists of the text of posts (Tweets) and their social media features. Some pre-processing and normalization steps have been applied to the dataset to improve the machine learning results.

The prepared dataset includes numerical social media features that have different intervals. For instance, the number of users who liked a tweet can be thousands; however, the number of media in this tweet can be a few. An imbalance between values of features can affect the performance of machine learning algorithms badly. Thus, the min-max normalization (Al Shalabi et al., 2006) was applied to the numerical social media features of samples in the dataset to remove instability. For the numerical attribute A where its interval $[min_A, max_A]$, the normalized value of v is calculated using the formula given in Equation 6.1.

$$v' = \frac{v - min_A}{max_A - min_A} \quad (6.1)$$

In the second step, numerical characters, punctuation marks, and weblinks were removed from the text of posts in the dataset. Additionally, lower case conversion was applied to the text of the posts. Thus, fewer and meaningful tokens would be obtained from the content in the feature extraction process. Additionally, misspelled online bullying terms were corrected using the preprocessing method which was developed within the scope of this thesis. The pseudo-code of the applied text preprocessing is presented in 6.1.

```

TextPreprocessing (string input, list of Slangwords)
{
  Remove numerical characters, punctuation marks, web links from input.
  Apply lowercase conservation input.
  Tokens  $\leftarrow$  Get tokens from input
  For each token in tokens
    Remove repeating characters (more than two) in token
    For each slangWord in SlangWords
      maxCharacter  $\leftarrow$  max (token.length, slang.length)
      distance  $\leftarrow$  LevenshteinDistance(token, slangWord)
      If ( (maxCharacter - distance) / maxCharacter > threshold)
        token  $\leftarrow$  slangWord
        break
      End if
    End loop
  End loop
}

```

Figure 6.1 Pseudo code of the text preprocessing method

In the text preprocessing method, the standard text cleaning operations are applied to the given input. Then, the tokens are obtained by splitting the input by space. For each token, characters that repeat more than two are removed since a word in the Turkish language does not include repeating letters that consecutively repeat more than two times. Lastly, the similarity between each token and slang words (a list that was collected in our previous work (Bozyiğit et al., 2019)) is calculated by using Levenshtein Distance (Heeringa, 2004). If the similarity is greater than the threshold, the token is replaced with the related slang word. The reason behind using Levenshtein similarity for the detection of misspelled slang words is that there is no open-source Turkish natural language processing tool that corrects misspelling perfectly.

Note that stemming was not applied in this study since Turkish is an agglutinative language, stemming can change the meaning of the term. For instance, “şerefsiz” (dishonorable) is a term in Turkish and its stem is “şeref” (honor) that has the opposite meaning.

At the end of the pre-processing process, it is observed that the experimented machine learning algorithms’ training time decreased dramatically by means of numerical normalization in social media attributes. Additionally, the features that have equal importance were obtained since the normalization process provides numerical stability to social media attributes that have different intervals. Lastly, applying text cleaning, lowercase conservation, and misspelled slang words correction reduced the noise, a large amount of additional meaningless information, in the dataset. Consequently, a more stable dataset that has less noisy data were obtained before the machine learning stage.

6.2 Feature Extraction

In the classification of text such as social media messages, it is crucial to represent documents in the dataset as vectors by extracting features. In this study, the text of samples (Tweets) in the dataset was converted into feature vectors by applying the bag of words approach (Wallach, 2006). Accordingly, the weight of the terms (features) is expressed numerically in the document-term matrix.

The terms’ weights were calculated using the Term Frequency - Inverse Document Frequency ($TF - IDF$) (Tokunaga & Makoto, 1994). Term frequency is the number of times a word appears within a document. Inverse document frequency is the logarithmic ratio of the number of samples in the dataset to the number of samples containing the relevant word (DF). The weight of a term is the multiplication of these two frequency types that is given in Equation 6.2.

$$TF - IDF = TF * \log(N/DF) \quad (6.2)$$

There are 21,938 features extracted from the text of the samples. However, irrelevant extracted features can be misleading that decreases the accuracy of machine learning models. Thus, recursive feature selection (Venkatesh & Anuradha, 2019) using a chi-square score was applied for each applied machine learning algorithm to select relevant textual features.

6.3 Applied Machine Learning Algorithms

In the experimental study, popular supervised machine learning algorithms were applied to the prepared dataset after preprocessing and feature extraction processes. The experimented algorithms were intentionally selected from different approaches such as lazy, regression-based, and ensemble. The applied algorithms in this study are explained as follows.

Support Vector Machines (SVM): It is a highly preferred supervised algorithm that can be used for both regression and classification problems. The objective of SVM is to find a hyperplane in N-dimensional space (N equals the number of attributes) that correctly separates the data points (Suykens & Vandewalle, 1999). Since there would be many hyperplanes that distinctly classify data points, the algorithm aims to find the plane with the maximum margin (distance) between both classes' samples. SVM's strengths are useful in high dimensional spaces, memory efficient, and versatile (having different decision functions).

Logistic Regression (LR): It is a statistical model that is used for binary classification in machine learning. LR is a regression analysis that estimates the relationships between a target feature and other independent features. LR uses the sigmoid function to model binary dependent features and then estimates the parameters of the model (Wright, 1995). The main advantages of this regression are being very efficient for training and easy to implement.

K-Nearest Neighbor (KNN): It has a lazy-learner (instance-based) algorithm where classification is conducted by using the similarity metric (Peterson, 2009). In

this algorithm, the class of a new sample is determined by looking at its K nearest neighbors in the training data set. The sample is assigned to the class that the majority of the K nearest neighbors belong. It is required to analyze the dataset to find the optimal value for K and similarity distance. The advantages of the algorithm are being robust to the noisy data and having no training phase.

Naive Bayes Multinomial (NBM): It is a supervised learning algorithm based on Bayes Theorem. NBM calculates the conditional probability $p(C_k|x_1, \dots, x_n)$ of each class C for the given sample where (x_1, \dots, x_n) are features of the sample (Kibriya et al., 2004). This Multinomial variation of Naïve Bayes takes the terms' number of occurrences (also it works for $TF-IDF$) in a document into account. NBM does not require big training data, and its computational time is fast enough to be used to make real-time predictions.

Adaptive Boosting (AdaBoost): It is an ensemble learning method that was initially created for increasing the performance of binary classifiers (Freund & Schapire, 1997). In this algorithm, firstly, a classifier is trained on the original dataset, and then additional copies of the classifier are created. The classifier's clones are fitted on the same dataset where the weights of incorrectly classified instances are adjusted. Thus, the cloned classifiers focus on more complicated cases.

Random Forest (RF): It is another applied ensemble learning approach in this study. RF trains the specific number of decision trees on various datasets' sub-samples created using random sampling with replacement (Liaw et al., 2002). Then, the average of predictions among the trained decision trees is used to improve predictive accuracy. This algorithm is one of the most used data science approaches since it is simple and successful even without parameter optimization.

6.4 Experimental Results

The machine learning algorithms experimented on two different variants of the prepared datasets. The first variant, named as D_T , includes only textual features. On

the other hand, the second variant, D_{T+S} , consists of the determined social media features and textual features.

It is important to remind that 21,938 textual features were extracted from the text of Tweets by applying the bag of words approach. Then, recursive feature selection was applied to these extracted features for each used classifier model. In this direction, the number of textual features varies depending on the related machine learning algorithm. Additionally, grid search (Bergstra & Bengio, 2012), the process of scanning the data to configure optimal parameters for a given model, was applied to the used machine learning algorithms (except NBM) in the training phase. The brief flow chart of the grid search is given in Fig. 6.2.

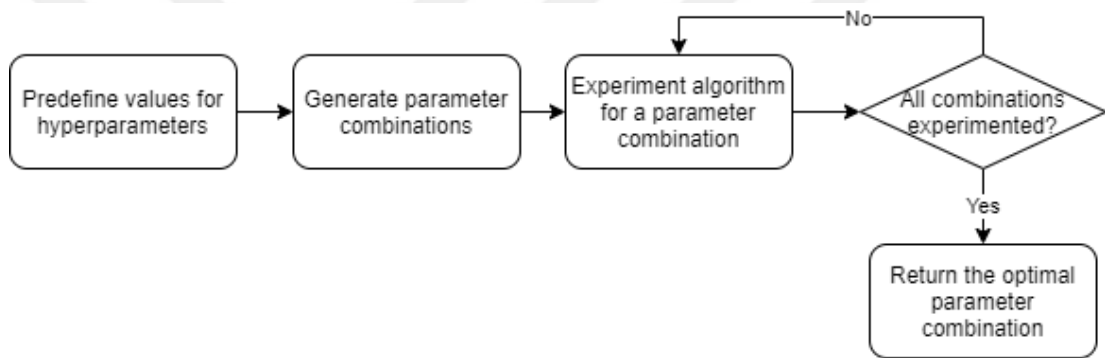


Figure 6.2 The brief flow chart of grid search

In the grid search process, each parameter combination for a given model is generated from the user-defined parameter values. Then, the process applies the model on the dataset for each parameter combination and selects the parameter combination with which the model has the best results. For instance, values were defined for KNN parameters as $K = \{1, 2, 3, \dots, 20\}$ and $metric = \{euclidean, manhattan, minkowski\}$ in this study. Grid search determined that KNN gives the best results on D_T when $K = 6$ and $metric = euclidean$.

As a result, the optimal parameters obtained by grid search and the number of textual features for each machine learning algorithm are presented in Table 6.1.

Table 6.1 Number of textual features and optimal parameters of classifiers

Classifier	Textual Features	D_T Optimal Parameters	D_{T+S} Optimal Parameters
SVM	500	regularization parameter=50, kernel= <i>rbf</i> , kernel coefficient=0.01	regularization parameter=5, kernel= <i>linear</i>
LR	500	regularization parameter=100, penalty= <i>l2</i>	regularization parameter=100, penalty= <i>l2</i>
KNN	500	K=6, metric= <i>euclidean</i>	K=3, metric= <i>euclidean</i>
NBM	1000	-	-
AdaBoost	500	learning rate=0.1, number of estimators=1000	learning rate=0.1, number of estimators=1000
RF	2000	number of estimators=200, max features= <i>log2</i> ,	number of estimators=250, max features= <i>log2</i>

Tenfold cross-validation (Rodriguez et al., 2009) was used to evaluate the experimented classifiers' performance on the prepared dataset's variants. The advantage of tenfold cross-validation is reducing bias by testing each sample once and using each sample nine times in training. In tenfold cross-validation, the dataset is divided into ten subsets that contain equal samples, and training is repeated ten times for the experimented machine learning algorithm. Each time, one of the subsets is used as the test set, while the other nine subsets' union is used as a training set. Then the mean accuracy of ten trials is accepted as the accuracy of the experimented classifier.

Accuracy was selected for the primary performance measure since the used dataset is perfectly balanced. Simply, accuracy is the ratio of the number of correctly classified samples in the test to the total number of tested samples. Accordingly, the brief experimental results are presented in Fig. 6.3.

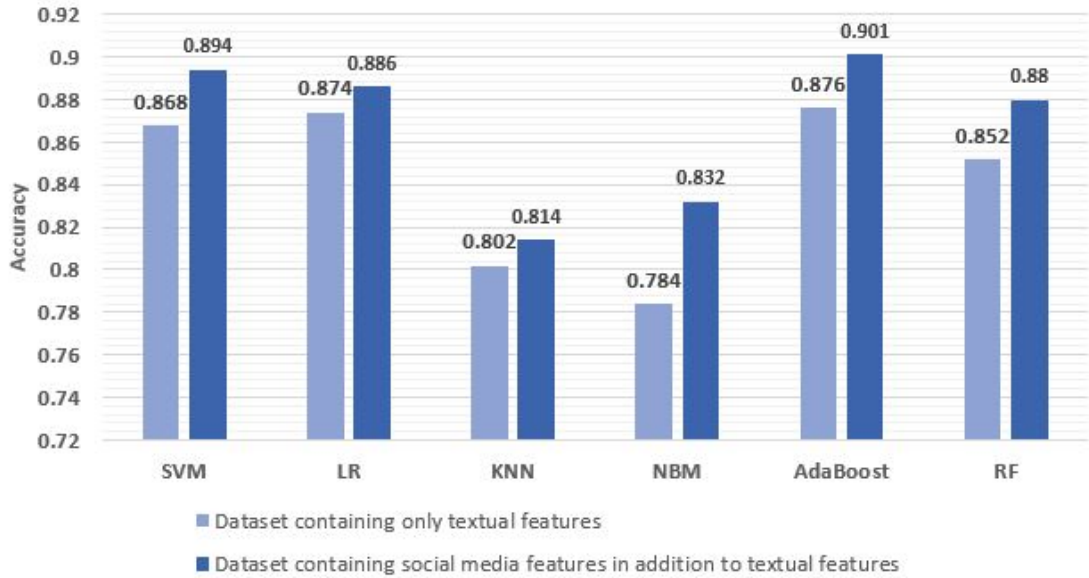


Figure 6.3 The accuracy of the experimented classifiers

In addition to accuracy, the experimented classifiers' performance scores were calculated according to other measures such as Precision, Recall, and F1-score. Precision is the ratio of the correctly classified positive samples to all samples classified as positive, while recall is the ratio of the correctly classified positives to all positive samples in the dataset (Davis & Goadrich, 2006). The harmonic mean (weighted average) of recall and precision is F1-score (traditional F-measure). Furthermore, the average elapsed time while classifying test sets (500 samples) in the tenfold cross-validation was recorded for each classifier. The detailed experimental results are presented in Table 6.2.

In the experimental results, it is seen that AdaBoost, experimented on D_{T+S} , is the most successful algorithm for the classification of online bullying contents. Except for KNN and NBM, the accuracy of other experimented machine learning algorithms is close to 90% which means they can correctly classify 9 of 10 social media contents in terms of cyberbullying. NBM and KNN have relatively lower accuracy than others since NBM performance decreases when a dataset has correlated features and KNN is not efficient for a high dimensional dataset that causes complicated distance calculations (Pechenizkiy, 2005). The reason behind AdaBoost being the most accurate classifier is that AdaBoost was implemented by combining several SVM

models. On the other hand, RF, which experimented on D_{T+S} , has the best Recall score, which means RF would be a better alternative if detecting positive samples is more critical than misclassifying negative samples. In terms of prediction time, it is observed that each classifier is a bit slower on D_{T+S} in comparison with D_T since D_{T+S} includes additional social media features. Lastly, it is seen that AdaBoost is the slowest classifier for predicting samples; however, AdaBoost is still fast enough to classify 500 samples in less than 230 milliseconds.

Table 6.2 The detailed results of the experimented classifiers

Classifier	Dataset	Prediction Time(ms)	Precision	Recall	F1-score	Accuracy
SVM	D_T	39	0.881	0.828	0.854	0.868
SVM	D_{T+S}	42	0.895	0.876	0.885	0.894
LR	D_T	0.1	0.863	0.866	0.864	0.874
LR	D_{T+S}	1	0.879	0.875	0.877	0.886
KNN	D_T	59	0.876	0.668	0.758	0.802
KNN	D_{T+S}	104	0.872	0.703	0.778	0.814
NBM	D_T	0.1	0.728	0.853	0.786	0.784
NBM	D_{T+S}	0.9	0.789	0.871	0.828	0.832
AdaBoost	D_T	215	0.890	0.836	0.862	0.876
AdaBoost	D_{T+S}	226	0.904	0.884	0.894	0.901
RF	D_T	71	0.857	0.820	0.838	0.852
RF	D_{T+S}	72	0.844	0.909	0.875	0.880

The critical point of the experimental study results is that each experimented machine learning algorithm gives a more successful prediction performance on the dataset containing social media features in addition to the textual features. For instance, using the social media features in the dataset enhances the SVM performance by 3% compared to the classical pure text mining approaches. It is clearly seen that social media features provide some useful information (e.g., users that have more followers in social networks are not prone to post online bullying content) that textual features cannot provide. Remark that information provided by social media features are discussed in chapter . As a result, one of the thesis's motivations, "social media features can increase the success of detecting online

bullying documents." is validated in the experimented study.

6.5 Further Experiments: Word Embedding

In this section, it is aimed to increase the accuracy of the experimented machine learning algorithms by replacing the used classical bag of words approach with word embedding in textual feature extraction. Word embedding is an approach where words are converted into real-valued vectors having the specified size (Levy & Goldberg, 2014). In this approach, each word of the given corpus is represented by a vector. The values of a vector are obtained from the usage of the related term in a sentence. Thus, synonyms or near-synonymous words have similar vector representations in word embedding models since they are used similarly in contexts. Furthermore, syntactic and semantic connections are also obtained using the word embedding approach. For instance, the relationship between male and female is automatically captured in this approach such that the result of vector operation "King – Man + Woman" would be very close to "Queen".

In the literature, there are many word embedding techniques/methods using various techniques such as neural networks, document statistics, etc. In the scope of this experimental study, Word2Vec and FastText, which are well-known word embedding approaches, experimented with automatic cyberbullying detection. Word2Vec and FastText are detailed Subsection 6.5.1 and 6.5.2 respectively, then the experimental results of these approaches are given in Subsection 6.5.3.

6.5.1 Word2Vec

In the Word2Vec algorithm, a neural network model is used for obtaining word relationships from a given text corpus (Church, 2017). Two different models can be used in this algorithm to get the vector representations of words. These models are continuous bag-of-words (named as CBOW) or continuous skip-gram.

In the CBOW model, the surrounding words of the given term are used as inputs in a neural network to predict the term (Kenter et al., 2016). The architecture of the CBOW model is shown in the Fig. 6.4, where w_t is predicted term in a sentence which consists of ordered tokens $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$.

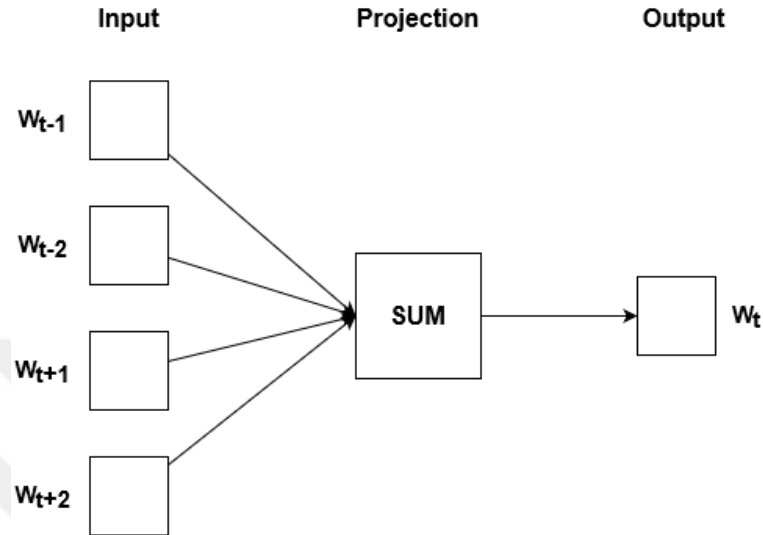


Figure 6.4 The architecture of continuous bag-of-words

In the continuous skip-gram model, the given term is used as input parameter in a neural network model to predict surrounding words (Lazaridou et al., 2015). The architecture of the continuous skip-gram model is shown in the Fig. 6.5 where w_t is predicted term in a sentence which consists of ordered tokens $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$.

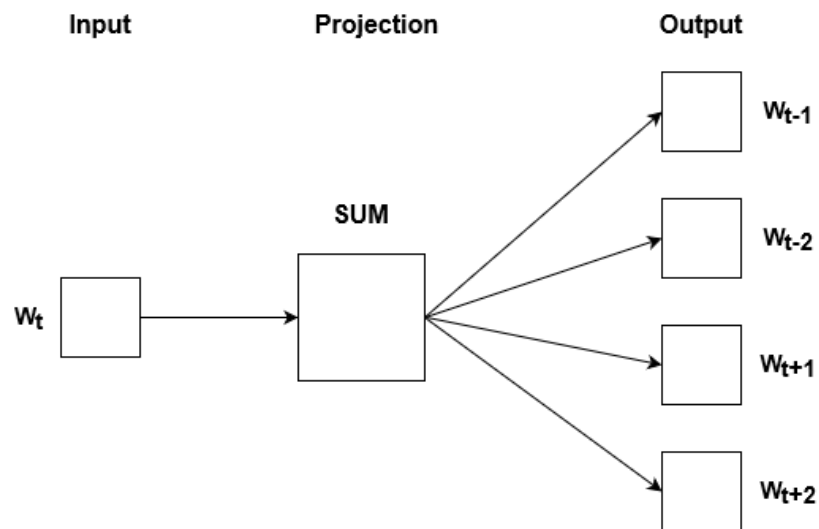


Figure 6.5 The architecture of skip-gram

6.5.2 *FastText*

FastText algorithm is an extension of Word2Vec that supports both models; CBOW and continuous skip-gram. However, its difference from Word2Vec is using the internal structure of a word to improve the accuracy of vector representations (Joulin et al., 2016). Indeed, FastText finds an embedding for previously unseen words since this algorithm uses the character n-grams of words in the corpus.

FastText processes a word into an array of character n-grams where n is generally between three and six. For instance, the term "iyidir" is decomposed as 3-grams in FastText such that {"iyi", "yid", "idi", "dir"}. Accordingly, the embedding vector of a term is calculated by summing vectors of the individual 3-grams (decomposition of the related term).

In most natural language processing researches, it is stated that FastText may be a better choice than Word2Vec when the data source is small. The reason behind that is FastText can find word embedding for unseen words, and it handles suffixes and derivational affixes. On the other hand, FastText takes more time than Word2Vec due to applying more operations. Moreover, some studies state Word2Vec over-performs FastText. Thus, both of these algorithms experimented within the scope of the thesis.

6.5.3 *Cyberbullying Detection with Word Embedding Approaches*

For cyberbullying detection, FastText and Word2Vec were applied to the prepared dataset using parameter optimization. The experimented word embedding methods were trained with nearly 350,000 tweets which were collected in an extended period. The parameter optimization was conducted to both word embedding methods for two crucial parameters; vector size (dimension of the word vectors) and window size (maximum distance between the current and predicted word within a sentence). The pseudo-code of the parameter optimization was given in Fig. 6.6.

```

TuneWE(embeddingMethod, allTweets, trainingX, validationX, trainingY, validationY)
{
  For w=2 to 8
    For v=50 to 300|
      embeddingModel ← embeddingMethod(sentences=allTweets, size=v, window=w)
      trainingXWordVectors←Text2WordVectors(trainingX["Text"], embeddingModel)
      validationXWordVectors←Text2WordVectors(validationX["Text"], embeddingModel)
      trainingData←hstack(trainingXWordVectors, trainingX["SocialMediaFeatures"])
      validationData←hstack(validationXWordVectors, validationX["SocialMediaFeatures"])
      optimalClf←TuneSVC(trainingData, trainingY)
      predictions←optimalClf.predict(validationData)
      result←metrics.accuracy_score(predictions, validationY)
    end
  end
}

```

Figure 6.6 The word embedding parameter optimization algorithm

Firstly, an embedding model was created in the algorithm from all collected tweets using the given method (Word2Vec or FastText), window, and vector sizes. The text of each sample in training and validation datasets was then converted to word vectors using the created model. Accordingly, validation and training datasets are prepared by stacking their samples' word vectors and social media features. Then, the optimal support vector classifier for the training is determined using the grid search approach. Lastly, the accuracy of the tuned support vector classifier that uses word vectors was calculated with respect to its predictions on validation data. The experimental results of Word2Vec and FastText are presented respectively in Fig. 6.7 and Fig. 6.8 respectively.

The accuracy of the tuned support vector classifier was obtained 87.8% using Word2Vec with window size seven and vector size 150. The tuned classifier's accuracy reached 85% using FastText word vectors with window size seven and vector size 250. However, the optimal support vector classifier's accuracy that uses the bag of words approach was 89.4% in the previous experimental study. It is seen that word embedding models are still slightly unsuccessful compared to the bag of words approach. The main reason behind this fact would be a relatively small training corpus or inapplicability of context.

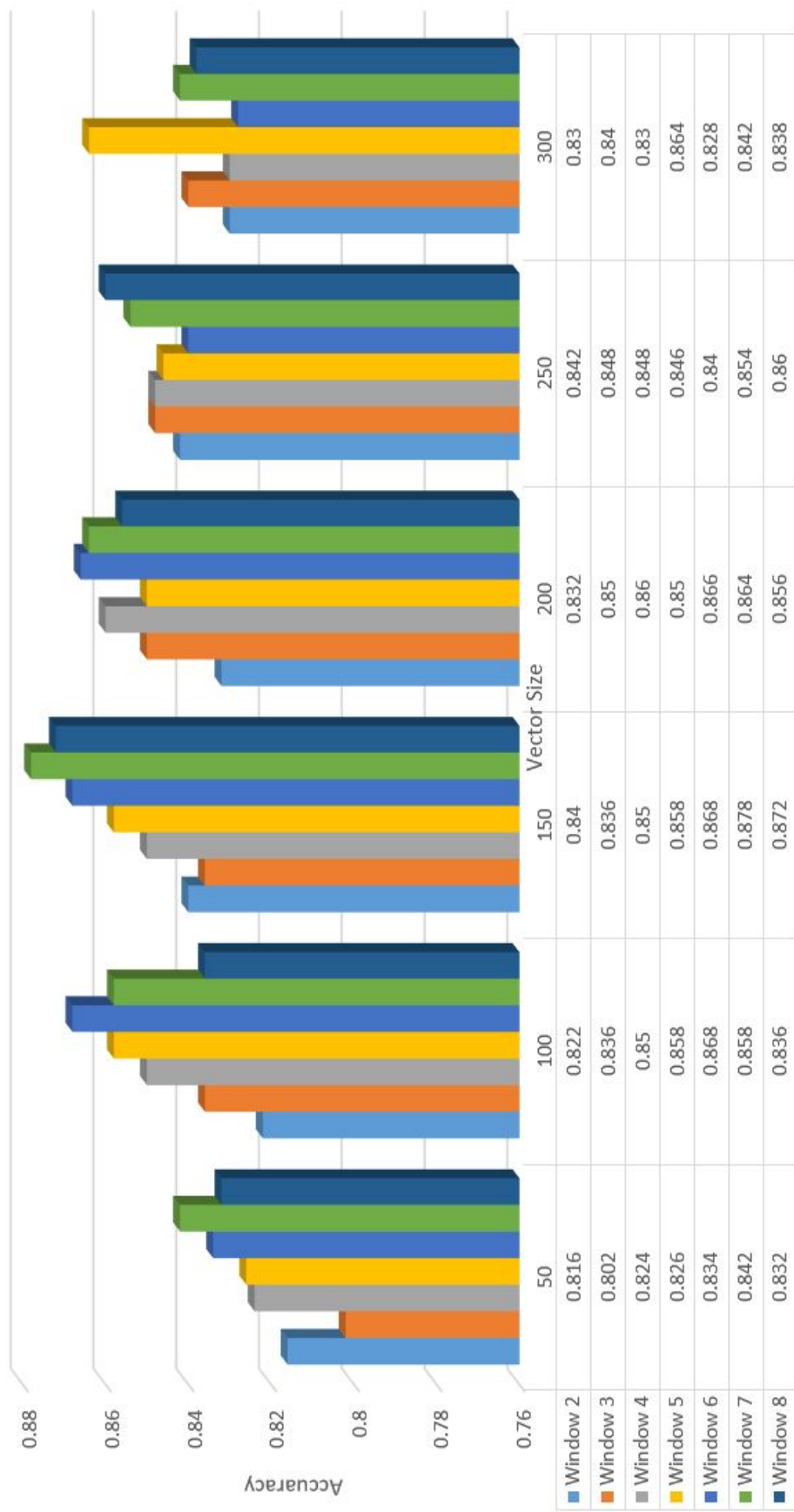


Figure 6.7 The experimental results of Word2Vec

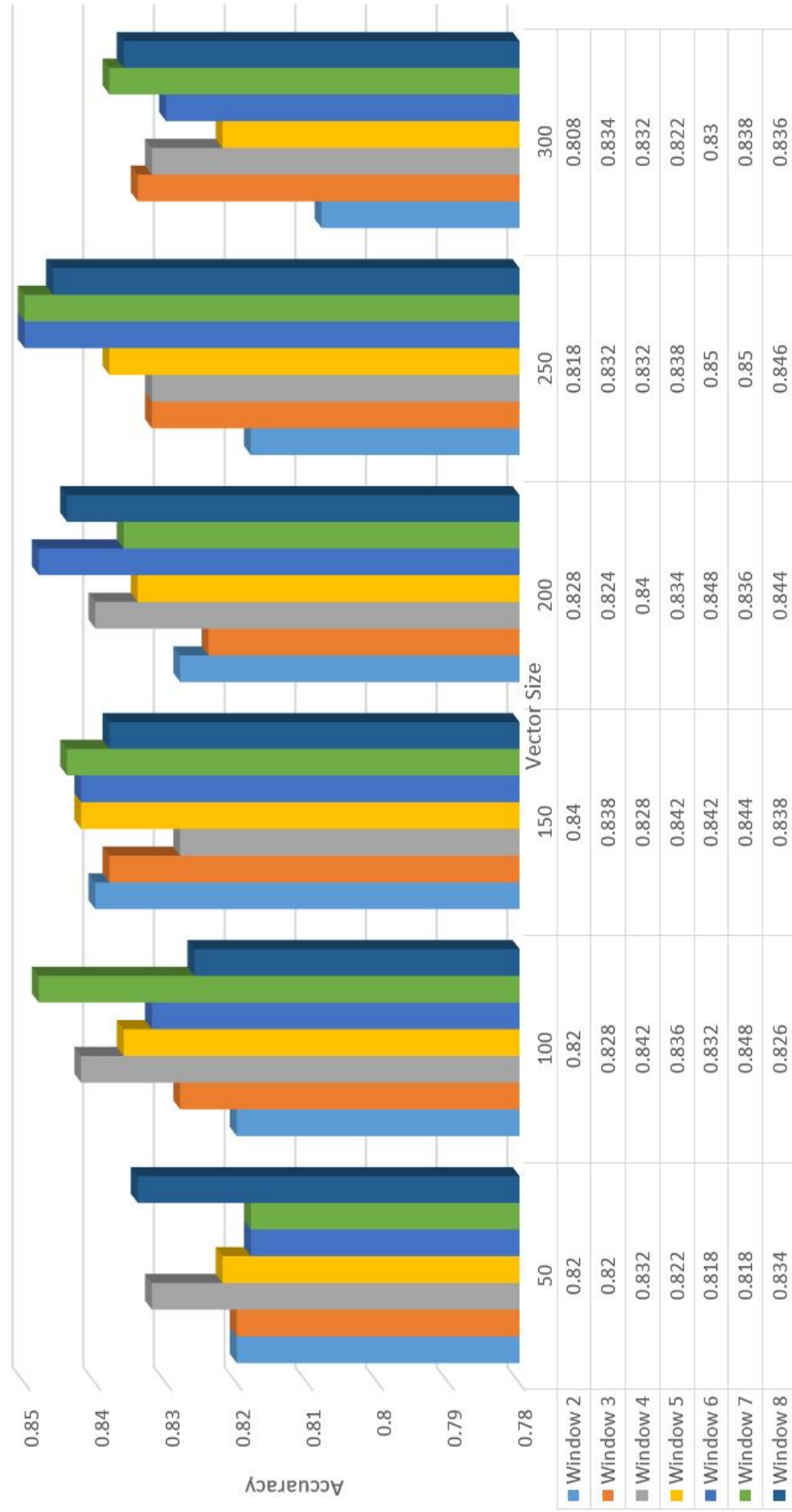


Figure 6.8 The experimental results of FastText

CHAPTER 7

CYBERBULLYING DETECTION WEB SERVICE

In the literature, there are many machine learning studies in the field of cyberbullying detection, and some of these studies publish their sources (e.g., datasets and codes) online. These sources are helpful for other researchers who want to bring the existing works to better points by studying new scripts on the published datasets or using the published codes in their datasets. Although publishing data sources and codes are valuable in the literature, integrating these codes into real-time applications is ignored. Accordingly, most of the existing studies in data science remain theoretical or pseudo-code. However, the detection of online bullying is a real-life problem, and the provided works in this field should be usable-integrable in real-time applications. A RESTful web service, which is named "CyberbullyingDetection", is developed to detect cyberbullying contents within the scope of the thesis.

By providing this web service, it is aimed to provide developers to classify online cyberbullying content on any platform. Accordingly, a programmer intended to develop cyberbullying applications can easily integrate this web service into their application without studying or knowing the machine learning process. Hence, this case motivates programmers to produce more real-time apps in this field and consequently increase awareness in cyberbullying detection, which is one of the objectives of this thesis.

The developed web service is published online on my personal website (Bozyiğit, 2021). The details of the developed web service are presented in the following sections. Firstly, the architecture of the web service is given in Section 7.1. Then, the endpoints of web services are detailed in Section 7.2.

7.1 Web Service Architecture

A web service would be defined as a set of open protocols used to transfer data between software systems or applications (Roman et al., 2005). Hence, software

applications developed using different programming languages and ran on various operating systems can use the same web service to exchange data. Web services, which are produced according to the architecture of REST (Representational State Transfer), are known as RESTful web services (Rodriguez, 2008). In these web services, HTTP (e.g., get, post, etc.) requests are used for implementing the REST concept, and JavaScript Object Notation (JSON) is used as resource description.

In this study, the RESTful web service was developed basically using ASP.Net Core MVC Framework and the top-ranking cyberbullying detection classifier in the machine learning process. The architecture of the web service is given in Fig. 7.1.

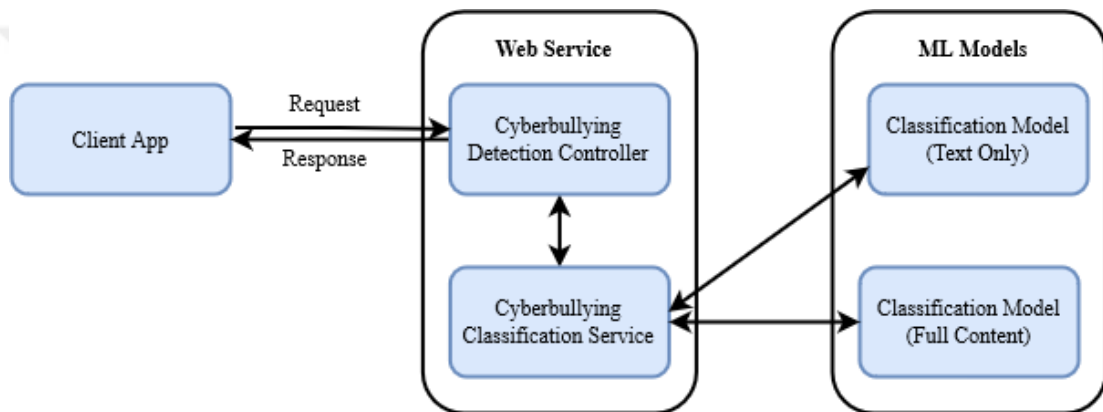


Figure 7.1 The architecture of the developed web service

A client app, which can be any application such as browser extension, android application, etc., makes a request to the web service. The web service handles the incoming request in the controller and sends the request to the classification service. Then, the classification service runs the related machine learning model according to the called endpoint. Accordingly, the executed machine learning model's prediction with the probability for the request is returned as a response to the client app. The details of the web service endpoints are presented in Section 7.2.

7.2 Web Service Endpoints

There are two end points of the web service; `ClassifyTextOnly` and `ClassifyFullContent`. The details of these endpoints are presented in Table 7.1.

Table 7.1 The endpoints of web service

Method	End Point	Description
Post	/api/cyberbullying/classifyTextOnly	Classifies the requested post according to only text.
Post	/api/cyberbullying/classifyFullContent	Classifies the requested post according to social media features along with text.

Both endpoints accept input data in JSON format from the request body. The input parameters for the request body are given in Table 7.2.

Table 7.2 The input parameters for web service request

Parameter	Type	Description
text	string	The text of the post.
retweets	int	The number of times a post was shared by other.
favorites	int	The number of users who liked a post.
senderAccountYears	int	The membership duration of the sender on the app.
senderFavorites	int	The number of times the sender liked other posts.
senderFollowings	int	The number of users following the sender of a post.
senderFollowers	int	The number of users followed by the sender of the post.
senderStatutes	int	The number of posts shared by the sender.
senderLocation	string	The location of the sender.

It is sufficient to send only the text data to the ClassifyTextOnly endpoint since this endpoint classifies a social media post according to the just text. On the other hand, the ClassifyFullContent endpoint classifies a social media post according to the post's social media features and text. These endpoints return a response in the same format as JSON data. There are two parameters of the response: the prediction of the classifier and the score, which is the probability of the prediction.

CHAPTER 8

CONCLUSION

In this chapter, the conclusion of the thesis is presented. In the first section, the thesis is summarized and the main research questions are addressed. The directions and recommendations for future researches are stated in the second section.

8.1 Summary

Cyberbullying, which is defined as bullying a person or a group of people using digital technologies (Slonje & Smith, 2008), has become a widespread problem affecting especially children and youngsters. In this thesis, it is aimed to detect cyberbullying content in social networks automatically. Firstly, a comprehensive dataset, which includes social media features(e.g., number of the sender followers), was prepared systematically due to the lack of quality cyberbullying datasets that have building and annotation process details (Rosa et al., 2019). Additionally, publishing the prepared dataset publicly available is one of this thesis' contributions.

Before implementing an automatic detection system, the characteristics of cyberbullying are inspected by analyzing the natural language processing features (e.g., the number of title words) and social media features on the prepared dataset. In the feature analysis, it is seen that natural language processing features are weakly related to cyberbullying. On the other hand, it is observed that some of the social media features are strongly related to online bullying. Accordingly, it was addressed that these related features would be promising for cyberbullying detection.

For automatic cyberbullying detection, an extensive machine learning process was conducted. After preprocessing and feature extraction steps, well-known supervised learning algorithms experimented on two variants of the created dataset; D_T that includes only textual features and D_{T+S} that consists of the determined social media features and textual features. Each experimented machine learning algorithm gives a more successful prediction performance D_{T+S} . Accordingly, It is concluded that

social media features provide useful information (e.g., users who have more followers in social networks are not prone to post online bullying content) that textual features cannot provide for classifiers. Additionally, the accuracy of some experimented machine learning algorithms is close to 90%, which means they can correctly classify 9 of 10 social media contents in terms of cyberbullying.

Further experiments in machine learning were conducted by implementing word embedding approaches in the feature extraction to increase the performance of the applied machine learning algorithms. Accordingly, well-known word embedding methods, FastText and Word2Vec, were trained with nearly 350,000 social media posts, and then they experimented with the machine learning algorithms. However, it is observed that word embedding models are slightly unsuccessful compared to the classical bag of words approach. The reason behind this inefficiency would be a relatively small training corpus (350,000 posts) compared to other word embedding works.

Lastly, a web service is developed and published to detect cyberbullying contents within the scope of the thesis. The motivation behind this web service is that most of the existing studies in this field remain theoretical or pseudo-code by ignoring the integration of their works into real-time applications. However, the detection of online bullying is a real-life problem, and the provided works in this field should be usable-integrable in real-time applications. Thus, a programmer intended to develop cyberbullying applications can easily integrate this web service into their application without studying or knowing the machine learning process.

8.2 Future Work

For future work, firstly, extending the used data source for word embedding approaches to millions of records could increase the accuracy of the trained machine learning algorithms. On the other hand, a fuzzy rule-based system that directly uses the analysis results of the social media futures could be implemented for

cyberbullying detection. Additionally, there are some studies that detect fake accounts in social networks (e.g., Erşahin et al. (2017)), using the information about whether the sender's account is fake would be quality input for automatic detection of cyberbullying. Lastly, it would be a valuable study that creates datasets belonging to different countries/regions that use the same language to compare the effects of social media features demographically.



REFERENCES

- Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63, 433–443.
- Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9), 735–739.
- Azevedo, P. J., & Jorge, A. M. (2007). Comparing rule measures for predictive association rules. In *Machine Learning: ECML 2007* (1st ed.) (43-52). Berlin: Springer.
- Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, 1(3), 295–311.
- Bauman, S., Cross, D., & Walker, J. L. (2013). *Principles of cyberbullying research: Definitions, measures, and methodology* (1st ed.). Oxfordshire: Routledge.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1), 281–305.
- Bozyiğit, A., Utku, S., & Nasiboğlu, E. (2019). Cyberbullying detection by using artificial neural network models. In *2019 4th International Conference on Computer Science and Engineering (UBMK), IEEE*, 520–524.
- Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179, 115001.
- Bozyiğit, A. (2021). *Open web service for online bullying detection*. Retrieved July 19, 2021, from <https://www.alicanbozyigit.com/projects/cyberbullyingdetection>.
- Caulfield, P. (2012). *Bullying didn't drive ashlynn conner, 10, to suicide: authorities*. Retrieved July 20, 2021, from <https://bit.ly/2UjAYSM>.
- Chaffey, D. (2020). *Global social media research summary july 2020*. Retrieved July 16, 2020, from <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research>.

- Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019a). Pi-bully: Personalized cyberbullying detection with peer influence. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5829–5835.
- Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019b). Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 339–347.
- Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1), 155–162.
- Cook, S. (2020). *Cyberbullying facts and statistics for 2020*. Retrieved July 5, 2020, from <https://www.comparitech.com/internet-providers/cyberbullying-statistics/>.
- Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *arXiv preprint arXiv:1812.08046*.
- Dadvar, M., Jong, F. d., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, University of Ghent.
- Dadvar, M., Trieschnigg, D., & de Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence* (1st ed.) (275-281). Cham: Springer.
- Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240.
- Diaz, J. (2020). *Hana kimura, japanese wrestler and star of netflix series, dies at 22*. Retrieved July 20, 2021, from <https://www.nytimes.com/2020/05/24/world/asia/hana-kimura-aew-dead.html>.
- Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *The social mobile web*, (1st ed.) (11-17). California: AAAI.

- Erşahin, B., Aktaş, Ö., Kılınç, D., & Akyol, C. (2017). Twitter fake account detection. In *2017 International Conference on Computer Science and Engineering (UBMK)*, *IEEE*, 388–392.
- Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., & Villaseñor-Pineda, L. (2017). Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89, 99–111.
- Febriana, T., & Budiarto, A. (2019). Twitter dataset for hate speech and cyberbullying detection in indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)*, vol. 1, *IEEE*, 379–382.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119–139.
- Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content. In *2017 1st Cyber Security in Networking Conference (CSNet)*, *IEEE*, 1–8.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (2nd ed.) (9-41). New York: Springer.
- Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University Library Groningen, Groningen.
- Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., & Mishra, S. (2014). Towards understanding cyberbullying behavior in a semi-anonymous social network. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, *IEEE*, 244–252.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

- Kenter, T., Borisov, A., & De Rijke, M. (2016). Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.
- Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence* (1st ed.) (488-499). Berlin: Springer.
- Knox, P. (2018). *Evil trolls target tragic amy 'dolly' everett's heartbroken pal just days after model, 14, killed herself over cyberbullying*. Retrieved July 20, 2021, from <https://www.thesun.co.uk/news/5337151/amy-everett-akubra-model-death-final-drawing-online-trolls-bullying-friend-targeted/>.
- Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*, 195–204.
- Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.
- Lerman, J. (2010). *Programming Entity Framework: building data centric apps with the ADO. NET Entity Framework* (2nd ed.). California: O'Reilly Media, Inc.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2177–2185.
- Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.
- Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3), 18–22.
- McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, 23(2), 143–149.
- Mendoza, M. (2016). *Evil trolls target tragic amy 'dolly' everett's heartbroken pal just days after model, 14, killed herself over cyberbullying*. Retrieved

- July 20, 2021, from <https://www.mysanantonio.com/news/local/article/Alamo-Heights-High-School-student-was-a-victim-of-6743320.php>.
- Mitchell, T. M., et al. (1997). *Machine learning* (1st ed.). New York: McGraw-hill.
- Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems with Applications*, 161, 113725.
- Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4, 51–62.
- Osborne, S. (2017). *Schoolgirl hangs herself after bullying video posted on social media platform*. Retrieved July 20, 2021, from <https://bit.ly/3seMRFY>.
- Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017). Detection of cyberbullying on social media messages in turkish. In *2017 International Conference on Computer Science and Engineering (UBMK), IEEE*, 366–370.
- Pechenizkiy, M. (2005). The impact of feature extraction on the performance of a classifier: knn, naïve bayes and c4. 5. In *Conference of the Canadian Society for Computational Studies of Intelligence, Springer*, 268–279.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883.
- Pop, D.-P., & Altar, A. (2014). Designing an mvc model for rapid web application development. *Procedia Engineering*, 69, 1172–1179.
- Ptaszynski, M., Masui, F., Kimura, Y., Rzepka, R., & Araki, K. (2015). Extracting patterns of harmful expressions for cyberbullying detection. In *Proceedings of 7th Language & Technology Conference: Human Language Technologies as a*

Challenge for Computer Science and Linguistics (LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions, 370–375.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, vol. 2, *IEEE*, 241–244.

Rodriguez, A. (2008). Restful web services: The basics. *IBM developerWorks*, 33, 18.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 569–575.

Roman, D., Keller, U., Lausen, H., De Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., & Fensel, D. (2005). Web service modeling ontology. *Applied ontology*, 1(1), 77–106.

Romsaiyud, W., na Nakornphanom, K., Prasertsilp, P., Nurarak, P., & Konglerd, P. (2017). Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*, *IEEE*, 242–247.

Rosa, H., Carvalho, J. P., Calado, P., Martins, B., Ribeiro, R., & Coheur, L. (2018). Using fuzzy fingerprints for cyberbullying detection in social networks. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, *IEEE*, 1–7.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. V., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333–345.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2), 147–154.

Spyder (2020). *Spyder ide*. Retrieved December 21, 2020, from <https://www.spyder-ide.org/>.

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (2nd ed.). Cambridge: MIT Press.
- Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300.
- Tokunaga, T., & Makoto, I. (1994). Text categorization based on weighted inverse document frequency. In *Special interest groups and information process society of Japan (SIG-IPSI)* (1st ed.) (33-39). Pennsylvania: Citeseer.
- Twitter (2020). *Twitter developer*. Retrieved November 12, 2020, from <https://developer.twitter.com/en/docs>.
- Venkatesh, B., & Anuradha, J. (2019). A hybrid feature selection approach for handling a high-dimensional data. In *Innovations in Computer Science and Engineering*, 365–373.
- Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, 977–984.
- Wang, F., & Sun, J. (2015). Survey on distance metric learning and dimensionality reduction in data mining. *Data Mining and Knowledge Discovery*, 29(2), 534–564.
- Wright, R. E. (1995). Logistic regression. *Reading and Understanding Multivariate Statistics*, 217–244.
- Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678.
- Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2, 1–7.
- Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. *arXiv preprint arXiv:1710.04989*.

Zhang, C., & Zhang, S. (2003). *Association rule mining: models and algorithms* (1st ed.). New York: Springer.

Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., & Dillon, E. (2016). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE*, 740–745.

Zheng, Z., Kohavi, R., & Mason, L. (2001). Real world performance of association rule algorithms. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 401–406.

