

**Graduation Project**

HATE SPEECH DETECTION IN TWEETS FOR TURKISH

**HATE SPEECH DETECTION IN TWEETS FOR TURKISH**

**Submitted By**

**Ahmet Özgür DARENDELİ 1501020057**

Project Advisor

Assis. Prof. Dr. Bahar İLGEN

Department of Computer Engineering

İstanbul Kültür University

2019-2020 Spring



**Graduation Project**

**HATE SPEECH DETECTION IN TWEETS FOR TURKISH**

**Submitted By**

**Ahmet Özgür DARENDELİ 1501020057**

Advisor

Assis. Prof. Dr. Bahar İLGEN

THE EXAMINATION COMMITTEE

Jury Member Signature

1. ………………………….. ………………………..

2. ………………………….. ………………………..

3. ………………………….. ………………………..

# ABSTRACT

NLP (Natural Language Processing) aims to analyze, understand and reproduce the regular structure of natural languages. NLP, in its widest scope, is the analysis of the natural language texts on social media such as tweets, blog posts, forum posts, and the transfer of their analysis to the computer environment.

Social media provides ease in producing discourse, making it widespread and normalizing this discourse in terms of accessibility. Studies shows that people can use social media platforms, especially Twitter, to make hateful, disturbing discourses, and a content such as that can have negative effects on society.

In this thesis study, by focusing on Turkish content, methods to automatically detect hate speech cases on social media that are targeting a certain part of the society, cultural values, ethnic/religious minorities and political views have been researched. Studies in this subject are mostly directed towards the English language. This subject has not been examined and analyzed thoroughly for the Turkish language. In this study, it is aimed to apply data pre-processing and determine classification for a certain dataset via Java programming language.

**Key Words:** Hate Speech, Machine Learning, Text Mining, Classification, Natural Language Processing, Twitter

# ÖZET

Doğal Dil İşleme, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır. En geniş kapsamıyla Doğal Dil İşleme, tweetler, blog yazıları, forum yazıları gibi sosyal medya üzerindeki doğal dillerdeki metinlerin bilgisayar tarafından algılanarak yazılım programında çözümlenmesi ve bilgisayar ortamına aktarılmasıdır.

Sosyal medya, ulaşılabilirliği açısından söylem üretilmesinde, bu söylemin yaygınlaştırılmasında ve normalleştirilmesinde kolaylıklar sağlamaktadır. Araştırmalar, insanların nefret içerikli, rahatsız edici söylemlerde bulunmak için sosyal medya platformlarını, özellikle Twitter’ı kullanabildiğini ve bu tür içeriklerin toplum üzerinde olumsuz etkilere yol açabildiğini göstermektir.

Bu tez çalışması kapsamında, Türkçe içeriğe odaklanarak, sosyal medyada üzerinde, toplumun belirli bir kısmını, kültürel değerleri, etnik/dini azınlıkları ve siyasi görüşleri hedef alan nefret söylemi olaylarının otomatik olarak tespit edilmesini sağlayacak yöntemler araştırılmıştır. Bu alanda yapılan çalışmalar çoğunlukla İngilizce diline yöneliktir. Bu konu, Türkçe dili için derinlemesine incelenmemiş ve irdelenmemiş bir konudur. Bu çalışmada, belirli bir veri kümesine Java programlama dili üzerinden veri ön işleme uygulanıp sınıflandırma yapılması amaçlanmıştır.

**Anahtar Kelimeler:** Nefret Söylemi, Makine Öğrenmesi, Metin Madenciliği, Sınıflandırma, Doğal Dil İşleme, Twitter

# TABLE OF CONTENTS

[ABSTRACT I](#_Toc5951944)

[ÖZET II](#_Toc5951945)

[TABLE OF CONTENTS I](#_Toc5951947)II

[LIST OF TABLES IV](#_Toc5951948)

[LIST OF FIGURES V](#_Toc5951949)

[1. INTRODUCTION 1](#_Toc5951951)

[1.1. Problem Statement *1*](#_Toc5951952)

[1.2. Project Purpose *1*](#_Toc5951953)

[1.3. Project Scope *1*](#_Toc5951954)

[1.4. Report Outline *2*](#_Toc5951956)

[2. RELATED WORK 3](#_Toc5951957)

[3. METHODOLOGY 5](#_Toc5951961)

[3.1. Keyword List Preparation *5*](#_Toc5951962)

[3.2. Dataset Creation *6*](#_Toc5951962)

[3.3. Natural Language Processing (NLP) *7*](#_Toc5951962)

[3.4. NLP Normalization *8*](#_Toc5951962)

[3.5. NLP Tokenazation *8*](#_Toc5951962)

[3.6. NLP Classfication *8*](#_Toc5951962)

[3.7. Sentiment Analysis *8*](#_Toc5951962)

[3.8. WEKA *8*](#_Toc5951962)

[4. EXPERIMENTAL RESULTS 9](#_Toc5951974)

[4.1. Classification with Weka *9*](#_Toc5951962)

[5. CONCLUSIONS 13](#_Toc5951976)

[REFERENCES 14](#_Toc5951977)

# 

# LIST OF TABLES

# Table 3.1: Keyword List........................................................................................................................5

Table 3.2: Dataset Example...................................................................................................................7

Table 4.2: Classification Results.........................................................................................................12

# 

# LIST OF FIGURES

[Figure 4.1: Weka GUI 1](#_Toc5872579)0

[Figure 4.2: Weka Explorer 1](#_Toc5872579)0

[Figure 4.3: Weka Classification 1](#_Toc5872579)1

# INTRODUCTION

## Problem Statement

In recent years, by the fact that social media has an important place in human life, social media platforms like YouTube, Facebook and especially Twitter become active areas for people to share their thoughts. This increase of activity in social media, also brought the content of hate speech with it. The aim in this project is to detection of Tweets that contain hate speech.

## Project Purpose

Hate speech can be described as activities that are targeting a certain part of the society, cultural values, ethnic/religious minorities and political views. Ease in discourse production provided by social media platforms and the increasing use of it in daily life, brought also the spread of hate speech with it. Hate speech can target a specific person or might be against a wide range of society which is the reason that makes this issue a bigger problem. By this reason, hate speech on social media should be detected.

In this project, it is planned to create a system that will automatically detect hate speech cases in Turkish tweets by analyzing giving datasets and searching it’s content to match results with a keyword list.

## Project Scope

As stated in the last section, the scope of this project is to detect hate speech on the data that collected from Twitter.

First step of the project was the creating a keywords list that containing widely used words described as a hate speech in Turkish. According to the determined hashtags, dataset created. Twitter API that provided by Twitter Developer is used to create a dataset. After creating the dataset from tweets, data pre-processing applied to it. By using NLP (Natural Language Processing), spelling mistakes in the dataset corrected to get more accurate results in the following parts. Sentiment analysis done at the step that following the data pre-processing. Weka machine learning tools used to make classification on the dataset.

## Report Outline

In the first section of this thesis study, the motivation of the project introduced to the reader in general terms. In the second section, similar researches on the subject found as a result of the literature researches were included. In the third section, the methods applied during this thesis study were explained. Dataset preparation, natural language processing, sentiment analysis and Weka machine learning tools can be listed as applied methods. In the fourth section, experimental studies and their results are presented.

# RELATED WORK

In this section of thesis study, similar studies that previously conducted in the literature and used as a starting point were included.

In their study, Z. Waseem and D. Hovy[1] created their corpus that contain more than 16k tweets and used it for hate speech detection by using a character n-gram based approach. They provided a list of criteria to annotate their corpus. It is stated in the study that their approach provided accurate results yet problem is far from solved. In addition to that study, another study done by Z. Waseem[2] following their first study, their corpus extended with an additions of new tweets.

In another study on hate speech detection on Twitter, Z. Zhang, D. Robinson and J. Tepper[3] introduced a method for classify hate speech using a deep neural network model. They used publicly available Twitter datasets and also added their own datasets using Twitter API. Racism and sexism determined as targeted characteristics.

In the study that focusing on detecting tweets contain anti-black hate speech, I. Kwok and Y. Wang[4] Naive Bayes classifier implemented. Tweets in their dataset selected from the accounts that classified as racist by different resources. 86% of their tweets labeled as racist since they contains offensive words. Their accuracy in classification evaluated as 76% with an average error rate of 24% by using 10-fold cross-validation method. Since in this study only unigrams employed, it is stated that text sentiments should be considered to evaluate more accurate results.

In another study, S. Malmasi and M. Zampieri[5] used a existing dataset that features English tweets annotated by annotator and categorized into three classes as hate speech, offensive language and non-offensive content. In their experiments, they performed multi-class classification. According to their experimental results, character n-grams and word unigrams performed well, but especially 4-gram model achieved best performance with the 78% accuracy. It is stated that distinguishing profanity in the content is a challenging task.

In another study on the different area in social media to detecting hate speech, dataset of user comments on Yahoo Finance used by N. Djuric, J. Zhou and friends[6]. In this study, two-step method proposed as their approach using embedding in binary classifier as second step to detect if comment contain hateful content.

# 3. [METHODOLOGY](#_Toc470871184)

**3.1. Keyword List Preparation**

**Table 3.1:** Keyword List

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Racism** | **Religious** | **Sexism** | **Characteristics** | **Politics** | **Extras** |
| batılı | allahsız | biseksüel | ahlaksız | darbe | andaval |
| çekik | ateist | eşcinsel | angut | darbeci | avel |
| çingene | dinsiz | fahişe | aptal | hırsız | çakal |
| çomar | gavur | ibne | bencil | monşer | çapsız |
| göçmen | gerici | sapık | beyinsiz | soyguncu | dengesiz |
| hain | kitapsız | sapkın | cadaloz |  | dümenci |
| kaçak | kızılbaş | sübyancı | cibiliyetsiz |  | enayi |
| köylü | misyoner | trans | dallama |  | gerzek |
| kürt | putperest |  | denyo |  | gevşek |
| laikçi | sıkmabaş |  | deyyus |  | hırbo |
| mülteci | yobaz |  | dümbük |  | hırtapoz |
| soysuz |  |  | dürzü |  | hıyar |
| suriyeli |  |  | embesil |  | katekulle |
| zenci |  |  | ergen |  | kaypak |
|  |  |  | ezik |  | keriz |
|  |  |  | gerizekalı |  | özürlü |
|  |  |  | hödük |  | paçoz |
|  |  |  | kahpe |  | salak |
|  |  |  | kansız |  | tırnakçı |
|  |  |  | karaktersiz |  | ucube |
|  |  |  | keko |  | varoş |
|  |  |  | kılıbık |  | velet |
|  |  |  | kıro |  | vizyonsuz |
|  |  |  | lavuk |  | yavşak |
|  |  |  | öküz |  | yıkık |
|  |  |  | onursuz |  | zibidi |
|  |  |  | pislik |  |  |
|  |  |  | şahsiyetsiz |  |  |
|  |  |  | samimiyetsiz |  |  |
|  |  |  | simsar |  |  |
|  |  |  | terbiyesiz |  |  |
|  |  |  | tipsiz |  |  |

As the initial phase of the project, determination of the keywords that containing hate speech selected. These words were obtained by analyzing popular hashtags on Twitter and the posts of social media users during the period of project. In addition, studies on hate speech in Turkish media were also examined.

At the end of the researches, a keyword list with 96 words was created. These words were categorized under 6 headings as racism, religious, sexism, characteristics, politics and extras. The final version of keyword list after this categorization can be seen in the Table 3.1.

**3.2. Dataset Creation**

Twitter search API is used to collect entire tweets in the dataset to create a resource for further use in the project. While doing this process, filtering used for tweets written in Turkish language to get more accurate data. As search criteria, trending hashtags on Twitter in a span of project time were chosen. In addition to that, also results of direct searches by keywords that are containing hate speech were added. After the data collecting process, redundancies in the dataset such as duplicates, retweets removed manually.

The number of hate speech words in the data obtained can be written by Java programming language. In this way, it can be printed ‘1’ on the hate content, and ‘0’ on non-hate content in the dataset.

The dataset created consist of 9479 tweets. The data in the first column decides whether the tweet contains hate speech. The second column consists of the tweets that cleaned by natural language processing.

An example from dataset can be seen in the Table 3.2.

**Table 3.2:** Dataset Example

|  |  |
| --- | --- |
| 1 | @meltemoztek : geldi bu tipsiz yaylı yine. |
| 1 | @zubeydetpc : samimiyetsiz ve gereksiz uzun bir akbank reklamı. |
| 0 | @gltnpolattt : İnsanın hayatında iyi ki var dediği birileri olmalı. |
| 0 | @\_aysegulkarakus : Birbirimize sunabileceğimiz tek anlamlı şey sevgidir. |
| 1 | @gsgsgsfterim : kendi özgür ve hür iradesiyle avmye giden salak var mı. |
| 1 | @graphixmrtii : kapatmış twitter hesabını pislik yobaz. |
| 1 | @aksamabekleriz : biz müslüman yapar mı dedik andaval. |
| 0 | @BesiktasAysegul : Rica ederim amin inşallah kalbi güzel Asi ruhlum. |

**3.3. Natural Language Processing (NLP)**

There are two different languages in the computer world. These languages can be classified as programming languages and natural languages. Natural languages are subcategories of artificial intelligence. Natural Language Processing (NLP), is the processing of the language that people speak by machines. NLP techniques are used to extract meaningful and desired information from the data in the meaning that the computer will understand.

Usage of NLP can vary in different areas. Machine translation is the process of trying to make sense from words and strings. fighting spam is filtering process on the data that extracted from mails. Information extraction is the another area that NLP used to extract data from news, forums and social media in the field of finance. Summarization can be explained as the interpretation of texts whether a brand is perceived good or bad reviews from the costumers. Also question answering, extract meanings from questions and give appropriate answers to them, is another area that NLP used.

By using Zemberek library, vowel and consonant errors, spelling and punctuation mistakes corrected. Also repetition of unnecessary words erased to clean the dataset.

**3.4. NLP Normalization**

In this part of the study, it is aimed to correct the misspelled words in the dataset by adapting the normalization in Zemberek library. Our sentences to change is given as follows “Tmm, yanına gidicem.” Then we apply the Normalization our sentences is “tamam, yanına gideceğim.”.

**3.5 NLP \_Tokenazation**  
Using the Tokenazition prepared by the Zemberek library for Turkish sentences, it enabled us to separate the sentences from each other. It has worked very well, especially in the adaptation to WEKA.

**3.6 NLP\_Classfication**

Using the Classfication prepared by the Zemberek library, we changed the vowel consonant errors in the cum.

**3.7. Sentiment Analysis**

General definition of sentiment analysis or opinion mining can be given as the process of identification and classification of a text to specify its meaning as positive, negative or neutral. Sentiment analysis can make it possible to detect emotions and thought of individuals about the products and services.

Sentiment analysis can be used many fields such as finance, medicine, marketing, media and politics. It is a widely used technique to measure consumers’ satisfaction for a particular product from the product reviews.

Due to insufficient Turkish resources, texts translated from Turkish to English by using Stanford NLP library and analyzed by using Java programming language. This library categorized sentences into three categories as positive, negative and neutral. According to results, sentences analyzed mostly negative in the dataset.

**3.8 WEKA**

It is basically a program that enables us to implement and analyze and classify data preprocessing modules. It has a simple interface and is very easy to use.

We preferred to use Weka so that we can make a correct and wide classification of our data we have cleaned with Natural Language processing. In order to take a more accurate approach, we chose to take the same number of hate and non-hate sentences.

Since it works with arff files in its basic structure, we have converted our file extensions in csv format to arff file with Wekan's own interface. It allows us to easily analyze our sentences in our dataset by separating them into n\_gram models.

# 4. EXPERIMENTAL RESULTS

Correcting mistakes in the sentences done by using Natural Language Process (NLP) on the data that gathered from Twitter in the pre-processing stage. In addition to that, words selected as stop words removed from the dataset while analyzing. With Java programming language, a program written that filter targeted hate speech words from the keywords list on the created dataset from Twitter. Program designed to give output 1 when the targeted word is found and 0 when the sentence doesn’t contain the target word. Weka machine learning tools used to make classifications according to approach.

**4.1. Classification with Weka**

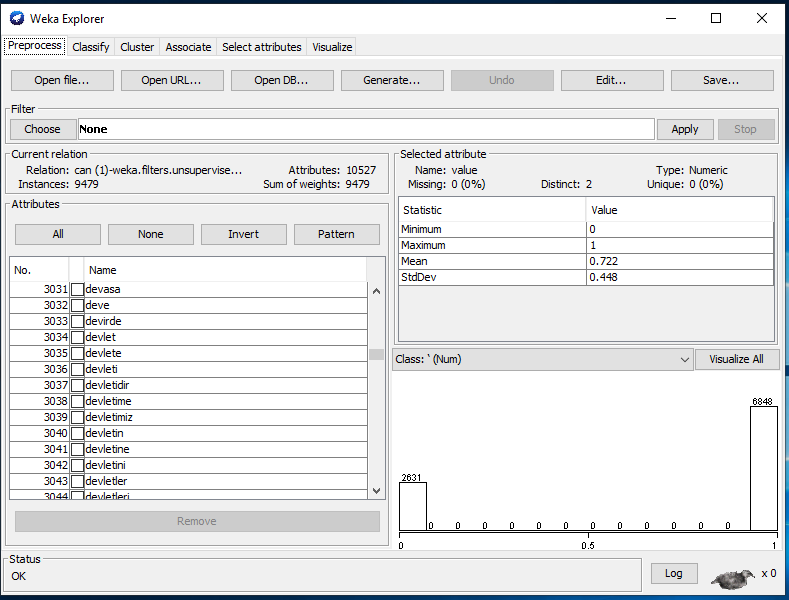
Test and analysis operations done on the Weka program. Weka is a open source machine learning library that developed by Java language. 3.8.3 version of Weka used in this project. The operations done on the Weka explained.

In the Figure 4.1, Weka GUI shown. Program comes with particular tools and libraries automatically when downloaded. To open datasets, Explorer tab can be used directly. Tools tab can be used to select algorithms to classify or select and download desired algorithms and libraries. In addition to that, it also include option to convert .csv files to .arff files. Files in the .csv format can be converted to .arff format from the arffViewer option in the Tools tab.



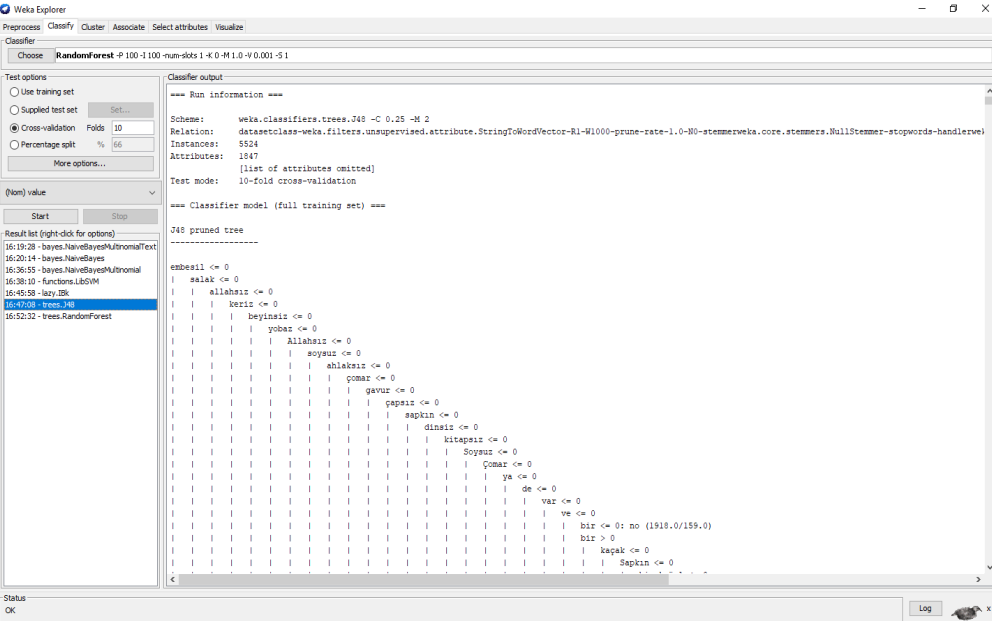
**Figure 4.1:** Weka GUI

After converting the dataset in the .arff format, StringToVector filter of Weka used to convert each word to vector in order to be able to classify.



**Figure 4.2:** Weka Explorer

After this process, Classify tab led to desired process. As shown in the Figure 4.3, hate speech words listed correctly by using decision trees.



**Figure 4.3:** Weka Classification

In this part, results obtained after our outcomes from Java program and dataset put into various classifications with Weka are presented. As classification methods, decision tree algorithms, Naïve Bayes algorithm and Linear Search algorithm used. Accuracy rates obtained from this experiments shared on Table 4.1. In addition, words that contains hate speech detected in the decision tree algorithms on Weka.

**Table 4.2:** Classification Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification Method** | **Correctly Classified Data** | **Incorrectly Classified Data** | **Accuracy Rate** |
| **Naïve Bayes** | 5039 | 485 | 91.2201% |
| **Naïve Bayes Multinomial** | 5276 | 248 | 95.5101% |
| **Linear Search** | 5169 | 355 | 93.5735% |
| **Libsvm** | 2975 | 2549 | 53.8559% |
| **J48 (Decision Trees)** | 5237 | 287 | 94.8045% |
| **Random Forest** | 5404 | 120 | 97.8277% |

According to results obtained, it was determined that a more accurate approach was found in the classifications made with decision trees.

# 5. CONCLUSIONS

This thesis study focused on detecting hate speech content and making classifications related to that. A program that filters hate speech content from tweets written in Java Programming Language. Also, classification made by Weka and accuracy of that classification tested. Using the decision trees, detection of hate speech content by Weka achieved.

In classifications, it is aimed to gain more accurate results and increase the accuracy of them by keep the numbers of tweets that are with and without hate speech equal in numbers.

The classifications we make clearly show that we are achieving a higher and more realistic rate in the random forest classification we made using the decision trees.

# REFERENCES

[1] Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

[2] Waseem, Z. (2016, November). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138-142)

[3] Zhang, Z., Robinson, D., & Tepper, J. (2018, June). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference* (pp. 745-760). Springer, Cham.

[4] Kwok, I., & Wang, Y. (2013, June). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence.*

[5] Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427.*

[6] Djuric, N., Zhou, J., Morrris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web (pp. 29-30).*