ANKARA YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES



AUTOMATIC HATE SPEECH DETECTION ON SOCIAL MEDIA: TURKISH TWEETS AS AN EXAMPLE

M.Sc. Thesis by

Tuğba DAĞAŞAN

Department of Computer Engineering

September, 2019

ANKARA

AUTOMATIC HATE SPEECH DETECTION ON SOCIAL MEDIA: TURKISH TWEETS AS AN EXAMPLE

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Tuğba DAĞAŞAN

September, 2019

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"AUTOMATIC HATE SPEECH DETECTION ON SOCIAL MEDIA: TURKISH TWEETS AS AN EXAMPLE"** completed by **TUĞBA DAĞAŞAN** under the supervision of **ASSIST. PROF. DR. RAHİME BELEN SAĞLAM** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Rahime BELEN SAĞLAM

Supervisor

Assist. Prof. Dr. Hilal KAYA

Jury Member

Assist. Prof. Dr. Nurcan ALKIŞ

Jury Member

Prof. Dr. Ergün ERASLAN

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: **Signature: ……..**

Name & Surname:…….

ACKNOWLEDGMENTS

Firstly, I would like to express my gratitude to two women. To my supervisor, Assist. Prof. Dr. Rahime BELEN SAĞLAM for her support and motivation. And to my mother-in-law for taking care of my children in the whole thesis process. Also, I would like to express my gratitude to my spouse for his infinite support and faith in me. Thank you all...

20.08.2019 **Tuğba DAĞAŞAN**

AUTOMATIC HATE SPEECH DETECTION ON SOCIAL MEDIA: TURKISH TWEETS AS AN EXAMPLE

ABSTRACT

Massive rise of user-generated web content, in particular on social media networks, caused research in social media to grow substantially in the last decade. Researches show that people may use popular micro blogging websites, especially Twitter, to make offensive comments that cover hate speech and such kind of contents can lead to a negative impact on the society. In this thesis we focus on Turkish context and investigate methods to detect hate crime incidents against ethnic/religious minorities, the LGBT population and women in social media automatically. We also evaluate the cross-domain performance of our approach and try to identify common discourses towards those groups aiming to establish a lexical baseline. Applying several supervised classification methods, we obtain out best results of 93% accuracy in detecting sexist tweets. Other results demonstrate that the main challenge lies in discriminating racist tweets.

Keywords: Hate speech, machine learning, text mining, classification, social media analysis, Twitter

SOSYAL MEDYADA NEFRET SÖYLEMİNİN OTOMATİK OLARAK TANIMLANMASI: TÜRKÇE ÖRNEĞİ

ÖZ

Kullanıcı tarafından oluşturulan web içeriğinin, özellikle sosyal medya ağlarındaki büyük yükselişi, son on yılda sosyal medyadaki araştırmaların büyük ölçüde artmasına neden oldu. Araştırmalar, insanların nefret içerikli, rahatsız edici yorumlar yapmak için popüler mikroblog web sitelerini, özellikle Twitter'ı kullanabildiğini ve bu tür içeriklerin toplum üzerinde olumsuz etkilere yol açabildiğini göstermektedir.

Bu tez çalışmasında Türkçe içeriğe odaklanıp, sosyal medyadaki etnik/dini azınlıklara, LGBT topluluğuna ve kadınlara yönelik nefret suçu olaylarını otomatik olarak tespit etmeye yarayacak yöntemleri araştırdık.

Ayrıca yaklaşımımızın başlıklar arası performansını değerlendirdik ve sözcüksel bir temel oluşturmayı amaçlayan gruplara yönelik ortak söylemleri belirlemeye çalıştık. Birkaç denetimli sınıflandırma yöntemi uygulayarak, en iyi sonucu cinsiyetçi tweet'leri saptamada %93 doğrulukla elde ettik. Diğer sonuçlar, asıl zorluğun ırkçı tweet'leri ayırt etmekten geçtiğini göstermektedir.

Anahtar Kelimeler: Nefret söylemi, makine öğrenmesi, metin madenciliği, sınıflandırma, sosyal ortam analizi, Twitter

.

NOMENCLATURE

Acronyms

NB	Naïve Bayes Classification
RF	Random Forest Classification
LLR	Linear Logistic Regression
LR	Logistic Regression
SVM	Support Vector Machines
PoS	Part of Speech
TF-IDF	Term Frequency-Inverse Document Frequency

LIST OF TABLES

Table 2.1 Example for word and character level n-gram 8

Table 2.2. Literature comparison table 13

Table 3.1. Distribution of tweets 18

Table 4.1. Word-level n-gram test results with highest 50 features 21

Table 4.2. Word-level n-gram results with highest 100 features 22

Table 4.3. Cross test results: Created models with Armenians dataset are applied on the other four datasets with highest 100 features 23

Table 4.4. Common n-gram counts between datasets (respectively in highest 50 and highest 100 TF-IDF-valued n-grams) 23

Table 4.5. Test results with PoS tagging features 25

Table 4.6. Character 5-gram results with top 50 features. 25

Table 4.7. Some of common words between datasets. 26

Table A.1. Cross test results: Created models with Armenians dataset are applied on the other four datasets with highest 50 features.. 32

Table A.2. Cross test results: Created models with Kurds dataset are applied on the other four datasets with highest 50 features. 32

Table A.3. Cross test results: Created models with Kurds dataset are applied on the other four datasets with highest 100 features. 33

Table A.4 Cross test results: Created models with Syrians dataset are applied on the other four datasets with highest 50 features 33

Table A.5. Cross test results: Created models with Syrians dataset are applied on the other four datasets with highest 100 features. 34

Table A.6. Cross test results: Created models with Women dataset are applied on the other four datasets with highest 50 features. 34

Table A.7. Cross test results: Created models with Women dataset are applied on the other four datasets with highest 100 features. 34

Table A.8. Cross test results: Created models with LGBTI dataset are applied on the other four datasets with highest 50 features. 35

Table A.9. Cross test results: Created models with LGBTI dataset are applied on the other four datasets with highest 100 features. 35

Table A.10. For all topics character 5-gram results with top 100 features. Values are precision, recall. 35

Table A.11. Test results with character-level n-gram features using highest 50 features in Armenians dataset 36

Table A.12. Test results with character-level n-gram features using highest 100 features in Armenians dataset. 36

Table A.13. Test results with character-level n-gram features using highest 50 features in Kurds dataset. 36
Table A.14. Test results with character-level n-gram features using highest 100 features in Kurds dataset. 37
Table A.15. Test results with character-level n-gram features using highest 50 features in Syrians dataset. 37
Table A.16. Test results with character-level n-gram features using highest 100 features in Syrians dataset. 37
Table A.17. Test results with character-level n-gram features using highest 50 features in Women dataset. 38
Table A.18. Test results with character-level n-gram features using highest 100 features in Women dataset. 38
Table A.19. Test results with character-level n-gram features using highest 50 features in LGBTI dataset. 38
Table A.20. Test results with character-level n-gram features using highest 100 features in LGBTI dataset. 39

LIST OF FIGURES

Figure 3.1 Data Preprocessing Phases..... 19

CONTENTS

M.Sc. THESIS EXAMINATION RESULT FORM ii
ETHICAL DECLARATION iii
ACKNOWLEDGMENTS iv
ABSTRACT v
ÖZ vi
NOMENCLATURE vii
LIST OF TABLES viii
LIST OF FIGURES x
CHAPTER 1 - INTRODUCTION 1
1.1 Problem Statement 1
1.2 Research Questions 3
1.3 Significance of the Thesis 3
1.4 Organization of the Thesis 3
CHAPTER 2 - BACKGROUND AND RELATED STUDIES 5
2.1 Background 5
2.1.1 Definition of Machine Learning 5
2.1.1.1 Unsupervised Machine Learning Algorithms 5
2.1.1.2 Supervised Machine Learning Algorithms 5
2.1.2. Definition of TF-IDF? 7
2.1.3 Definition of N-gram 8
2.2 Related Studies 9
CHAPTER 3 - METHODOLOGY 16
3.1 Data Collection 16
3.2 Data Preprocessing 18
3.3 Feature Construction 20
3.4 Model Construction 20
CHAPTER 4 - RESULTS AND DISCUSSION 21
CHAPTER 5 - CONCLUSION AND FUTURE WORK 27
REFERENCES 28
APPENDICES 32
Appendix A – Experiment Results 32
CURRICULUM VITAE 41

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Freedom of expressions on social media provides leverage to an individual to share their opinions and beliefs about anything. However, it is a known fact that people may misuse freedom of expression to make offensive comments that have the potential to cause harm and suffering on an individual basis. Such contents may also lead to social tension and disorder beyond cyber space. However, lack of policy and consistent enforcement and difficulty to detect such online contents lead to increase in hate on social media.

There is no formal definition for “hate speech”; but according to decision of Council of Europe in 1997, hate speech covers all forms of expressions that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance [1]. Hate speech and hate crime are closely related notions. There are two elements that make up a hate crime. The first element is the existence of an act which constitutes a criminal offense under ordinary criminal laws. Although there are some differences in the laws of countries, what is similar in these actions is usually that these acts involve violence. The second element of hate crime is prejudices. The delinquent selects the target intentionally according to his/her race, gender, religion, language, nation or ethnicity. If the existence of the second element is proved, it is called hate crime [2].

In many countries like Holland, Denmark, Island, Norway or United Kingdom, hate speech is a crime prohibited by laws. It can lead large fines and even imprisonment [3]. Inevitably, these laws have been expanded to cover online content leading many sites to seek to actively combat hate speech. In Turkey, although there is no law directly for hate crime, there is discrimination crime notion since 2005 and “hate crime” notion since 2014 in the Turkish Criminal Law [4]. It is a known fact that, in hate crimes in Turkey, victims are actually members of vulnerable and disadvantaged

groups [5]. However, data collected by Ministry of Justice is dominated by specific hate crimes such as “grave desecration” or “attacks on places on worship” [6] which excludes those protected characteristics. Consequently, it is possible to conclude that crimes attacking religion, sexual orientation or ethnicity are not reported officially in Turkey which reveals the importance of our study.

According to the report of We Are Social, 72% of the population is internet users in Turkey. 58% of these are Twitter users. In Turkey, Twitter is the most commonly used social media platform, after YouTube, Instagram and Facebook. As of July 2019, Turkey is the fifth country on Twitter country rankings [7].In the direction of this information, this study focused on the detecting hate speech on Turkish tweets. With the aim of detecting racism and sexism following groups, which are frequently targeted by hate speech in Turkey, were covered: Armenians, Kurds, Syrians, women and LGBTI. In this context, Turkish tweets with different hashtags were collected via Twitter Search API.

For all target groups, completely irrelevant tweets (for example, tweets for news), tweets contain just symbols, just hashtags and URLs, retweets, tweets written in another language were removed. After this elimination, the rest 4081 tweets were annotated as hateful or non-hateful by three labelers.

In preprocessing phase, the tweets were lowercased, tokenized,corrected if misspelled, and stemmed. Three feature setswere constructed;word level n-grams, part of speech tags and character level n-grams. For word level n-grams, unigrams, bigrams and trigrams were generated. Character n-grams (of order 2–5) were generated. Finally, number of pronouns, adjectives and nouns were counted. Following machine learning algorithms were apllied on these feature sets for each target group : Naive Bayes, linear logistic regression, logistic regression, Random Forests, Support Vector Machine.

We also cross-validated across all classes on an individual basis - training on one and testing on another to inspect similarity in context of hate speech among different target groups.

1.2 Research Questions

This thesis presents a model to detect hate speech in Turkish tweets. Five datasets which are possibly containing racist and sexist speech are used. It is aimed to answer following questions:

-How can hate speech in Turkish tweets be detected automatically? Which feature sets can be the best for this aim?

-Is there any similarity between languages used in hate speech towards different target groups?

1.3 Significance of the Thesis

Hate speech is a new concept in Turkey. Consequently, the number of studies that focus on hate speech in Turkey is quite limited. Most of the existing studies are social studies that focus on understanding the hate speech and its effects on public. In addition, even though there are studies which also used social media data including Twitter, Youtube and newspapers as data source, they focused on a single target group like Kurds, Armenians, or LGBTIs [8-11] and did not attempt to identify potential commenalities among the different groups.

To the best of our knowledge this is the first study that covers *detecting* hate speech toward different target groups on Turkish tweets.

With this thesis;

- The best feature set is identified to detect hate speech in Turkish tweets.

- The similarities between race and sex oriented hate speech are presented.

1.4 Organization of the Thesis

The thesis is divided into five chapters. Next chapter, Chapter 2 includes background information and related studies. Chapter 3 introduces the methodology employed in this study. This chapter explains the data collection and preprocessing, feature construction and model construction. Chapter 4 presents the results obtained and discussion on results. Finally, Chapter 5 includes

conclusions and thoughts about future work.

CHAPTER 2

BACKGROUND AND RELATED STUDIES

2.1 Background

2.1.1 Definition of Machine Learning

Machine learning definition is put forward in 1959 by a computer pioneer named Arthur Samuel. According to him, it is a field of study that gives computers the ability to learn without being explicitly programmed.

As the data and program together gives an output in traditional programming, in machine learning approach, data and output is given to the computer and a program is produced [12]. In other words, machine learning which a subdivision of artificial intelligence lead systems can learn from data.

Machine learning algorithms are classified into two main headings in terms of learning: Supervised and unsupervised.

2.1.1.1 Unsupervised Machine Learning Algorithms

Unsupervised learning algorithms which use no train data are used to understand the structure or distribution of the input data. These are mainly used in clustering and association problems. K-means clustering and Apriori algorithms can be given as two commonly used algorithms that split data into clusters.

2.1.1.2 Supervised Machine Learning Algorithms

In supervised machine learning models, the algorithms learn from a labeled dataset namely training data. The label indicates class each record actually belongs to. The created models are evaluated on unseened dataset, namely test data. Supervised learning algorithms are mainly used for classification and regression problems. Support Vector Machines (SVM), discriminant analysis, Naïve Bayes (NB), k-nearest neighbor and neural networks are some of classification algorithms. Linear (LLR)and logistic(LR)regression are some of regression algorithms.

For the classification problem studied in this thesis, following supervised algorithms have been utilized: NB, SVM, LLR, LR and Random Forest(RF).

Naive Bayes

Naive Bayes is a classification algorithm for binary and multi-class classification problems. It is a probabilistic learning method.

The probability of a document *d* being in class *c* is computed as

$$P(c|d) = P(c)\prod_{1\le k\le n_d} P(t_k|c) \quad \textbf{(1.1)}$$

where $P(t_k|c)$ is the conditional probability of term $t_k$occurring in a document of class *c*. $P(t_k|c)$ is interpreted as a measure of how much evidence $t_k$contributes that *c* is the correct class. $P(c)$ is the prior probability of a document occurring in class *c*. If a document's terms do not provide clear evidence for one class versus another, higher prior probability is chosen. (t_1, t_2,..., t_{n_d}) are the tokens in *d*. n_d is the number of these tokens in *d*.. The best prediction about class in NB is the most likely or *maximum a posteriori* (*MAP*) class c_{map}[13]:

$$c_{map} = \arg\max P(c|d) \quad \textbf{(1.2)}$$

Linear Regression

. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [14]. Given an input vector $x \in R^m$, where $x_1, \ldots, x_m$ represent features (also called independent variables or predictors), we find a prediction $y' \in R$ for the class of a document $y \in R$ using a linear regression model:

$$y' = \beta 0 + x^T\beta \quad \textbf{(1.3)}$$

where β0 and β are the parameters to estimate. Usually, the parameters are learned by minimizing the sum of squared errors [15].

Logistic Regression

In this method, input values are blended linearly using weights to predict an output value *(y)* which is the tweet class in this study. A key difference from linear regression is that the output value being modeled is binary values (hateful or non-hateful) rather than a numeric value.

Random Forest

Random forest (RF), consists of a large number of individual decision trees that operate as collaboration. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Support Vector Machine

The purpose of this algorithm is to find a hyperplane in an N-dimensional space(N is the number of features) that distinctly classifies the data points. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM [16].

2.1.2 Definition of TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a concept for statistical analysis of texts in text mining. It is used for weighting terms in a document collection.

The term frequency (TF) stands for the number of occurrences of term *t* in document *d*. It is denoted as $tf_{t,d}$.

While computing TF, all terms are considered equally important. In fact certain terms like "the", "of", "bu ('this' in Turkish)", "bir ('one in Turkish')" or which occurs many times in the document actually does not have much distinctive effect in determining class. So, the frequencies of these terms are weighted down by computing the inverse document frequency(IDF). *IDF* of a term *t* is computed as follows [13]:

$$idf_t = \log \frac{N}{df_t} \quad \textbf{(1.4)}$$

The document frequency df_t, stands for the number of documents in the collection that contain a term t . N in the formula stands for the total number of documents in a collection.

TF-IDF is calculated as the product of TF and IDF values and computed as follows:

$$tf\text{-}idf_{t,d} = tf_{t,d} \, x \, idf_t \quad \textbf{(1.5)}$$

2.1.3 Definition of N-Gram

N-gram is a subsequence of n items from a given sequence. One-sized-n-gram is called unigram, 2-sized n-gram is called bigram and 3-sized n-gram is called trigram. Items phrase in the definition could be phonemes, characters or words. N-gram can be better understood with a sample.

Word and character n-gram examples of the "This study is about hate speech detection" sentence are shown in the following table. Underscore characters in the table represent spaces in the sentence.

Table 2.1 Example for word and character level n-gram

Level	Size	Grams
Word	1 (unigram)	This, study, is, about, hate, speech, detection
	2 (bigram)	This study, study is, is about, about hate, hate speech, speech detection
	3 (trigram)	This study is, study is about, is about hate, about hate speech, hate speech detection
Character	2 (bigram)	Th, hi, is, s_, _s, st, tu, ud, dy, y_, _i, is,….
	3 (trigram)	Thi, his, is_,s_s, _st,stu,tud,udy,dy_, y_i, _is, is_,..
	4 (4-gram)	This, his_, is_s, s_st, _stu, stud, tudy, udy_, dy_i, y_is,..
	5 (5-gram)	This_,his_s, is_st, s_stu, _stud, study, tudy_, udy_i,…

2.2 Related Studies

In the literature, several approaches for different languages have been proposed for different target groups.

In 2013 Burnap et al. has argued that offline events can trigger 'online tension' which enables to detect offline events monitoring peaks in tension in social media, affording a form of 'anticipatory governance'. The researchers investigated the possibility of forecasting spikes in social tension. They focused on a single hashtag and scaled the English tweets with this hashtag from 0 to 3 (from irrelevant to high tension), then applied three different techniques: tension analysis engine, sentiment analysis and machine learning algorithms, namely naïve Bayes, Logistic Regression and Support Vector Machine (SVM). The researchers work with unigram (single word), bigram (two words) and trigram (three words), and combined (uni-, bi- and tri-gram) features, with separate experiments for n-gram presence and frequency. Naïve Bayes is reported to perform best overall out of three machine learning [17]. In our study, we have applied 5 machine learning algorithms including these three and weighted the features with tf-idf values in addition to presence and frequency. Working on different hashtags for different target groups and applying cross validation among the different groups are some other differences.

In another work, Burnap and Williams tried to capture 'othering' language which focused on especially identifying anti-Muslim sentiment. Othering is an established construct in rhetorical narrative surrounding hate speech. Researchers created several features including 1-gram to 5-grams, predefined hate terms and typed dependencies. They applied SVM with linear kernel and Random Forests algorithms to detect othering language on race, disability and sexual orientation where the best result was obtained with SVM. The researchers also made cross validations of different types of cyber hate. Results of cross validations show that the hate structure of disability is pretty different than the others and there are similarities between race and sexual orientation. Best cross validation results were achieved when the model was trained using homophobic cyber hate instances and tested using the race dataset [18]. In our study, we have expanded the feature set with character grams and followed a similar approach for different target groups in Turkey. However, we did not cover

predetermined hateful terms which make our study to be applicable across different languages.

In 2017, Davidson et al. conducted a study to separate the tweets into three groups: hate speech, offensive language and none. They applied stemmer on the tweets which are annotated using Crowdflower, computed the word-grams and their TF-IDF values. They also used binary and count indicators for many properties like hashtags, mentions, retweets, number of characters and words for each tweet. Logistic Regression, Naïve Bayes, decision trees, Random Forests and linear SVM techniques were applied and it has been reported that Logistic Regression and Linear SVM tended to perform significantly better than other models. The best performing model had overall precision 0.91, recall of 0.90 [3].

Malmasi and Zampieri used the same dataset introduced in [3] in their study to detect hate speech in social media. However, they also used surface n-grams which are 2-8-character n-grams and 1-3-word n-grams, word skip-grams and Brown clusters. The researchers applied linear SVM classifier with 10-fold cross validation. Best accuracy (%78) was achieved with character 4-gram and best accuracy in word n-grams was reported as %77.5 with unigrams [19]. These findings lead us to cover character-grams and evaluate their performance in Turkish dataset.

In 2017, Karamollaoglu et.al. ,Turkish researchers, tried to detect offensive language in English tweets by using machine learning algorithms. In this study that did not consider any target group, they produced balanced dataset consisting 250 hateful and 250 non-hateful tweets. 80% of them were used as training data. In preprocessing phase https, hashtags, unnecessary punctuations and stop words were removed from tweets. Also, words are stemmed via Lucene API. They reported 81% accuracy with Naive Bayes and vector space model [20].

In 2018, Pitsilis et.al. introduced a deep learning method, they employed an ensemble of LSTM-based classifiers to improve classification. They used users' characteristics, such as the tendency to post hateful messages. They chose to vectorize the tweets using word-based frequency vectorization. The dataset they used contained approximately 16k Tweets which 3166 of them labeled as Sexism, 1943 of them labeled as Racism and rest of them labeled as neutral namely non-

hateful. They reported the small number of tweets and the imbalanced dataset made the task more challenging. Nonetheless, they reported best results in literature which are precision of 0.9305, recall of 0.9334 and f-score of 0.9320 [21].

In 2018, Alorainy et.al proposed another deep learning method. They proposed a novel 'othering' feature set which contains three parts: dependency relationship labels which probably represent othering, parts of speech associated with othering (e.g. send them home) and a list of English pronouns. They preprocessed the feature set with Paragraph2Vec and used Multilayer Perceptron (MLP) as classifiers for the detection. They experimented the model on approximately 24k tweets which 5% of them were labeled as hateful. They obtained F-measures of 0.81, 0.60, 0.89 and 0.71 for religion, disability, race and sexual oriented hate speech respectively. They compared the results of their novel method with state of the art baseline models from the literature. As they produced 0.99 F-measure, they reported their approach outperformed all existing methods [22].

In 2019, Basile et. al. studied multilingual detection of the hate speech towards women and immigrants. This team actually created an event called HatEval. In this event, the created unique dataset is studied to detect hate speech by different teams via different methods like Recurrent Neural Network, LSTM models and SVM. They investigated hate speech in English and Spanish tweets. They also investigated further features in hateful contents such as the aggressiveness attitude and the target range which indicates whether the target is a generic group of people or a specific individual. They reported that the best result is achieved via SVM model with RBF kernel only [23].

In 2019, Ayan et.al. tried to detect Islamofobic tweets on English tweets using sentiment analysis. They splitted the dataset which consisted of 162k tweets according to some keywords. Tweets containing words used to insult Muslims like "muzrad" were added to a table which holds highly likely to be Islamophobic content. Tweets containing the words in the Muslim jargon like "mashallah" were added to another table which holds likely to be non-Islamophobic. Words in the tweets were weighted with TF-IDF. Then they utilized Ridge regression and Naive

Bayes. They reached 96.9% accuracy rate for classification via Ridge Regression in F1 criterion [24].

These studies are listed in Table 2.2 for better comparison. The letters P, R, A and F in this table stand for precision, recall, accuracy and F-score respectively.

Table 2.2 Literature comparison table

Studies	Language	Hate Types	Preprocessing	Labelling tweets	Techniques	Algorithms	Feature set	Best Performer	Best Result
Burnap et.al[17]	English	Racism		0-3 (irrelevant to high tension)	Tension analysis engine, Sentiment analysis, Machine learning	Naïve Bayes, Logistic Regression, Support Vector Machine (SVM)	-Unigram -Bigram -Trigram -Combined (uni-,bi-,tri-)	Naïve Bayes	P: 0.83 R: 0.91 F: 0.87
Burnap &Williams [18]	English	Race Disability Sexual	Stemmed Lowercased	0-1 (offensive, not offensive)	Machine learning	SVM with linear kernel, Random Forests	-1 to 5-grams -predefined hate terms -typed dependencies	SVM	For disability, P:0.91 R: 0.61 F: 0.73
Davidson et.al [3]	English	Race Sexual	Stemmed Lowercased	0-2 (Hate speech, offensive, none)	Machine learning	Logistic Regression Naïve Bayes decision trees Random Forests Linear SVM	-1-3 Word-grams -Binary and count indicators for hashtag, mention, retweet	Logistic Regression Linear SVM	P: 0.91 R: 0.90
Malmasi &Zampieri [19]	English	No specific target group	Lowercased Removed URL Removed emoji	0-2 (Hate speech, offensive, none)	Machine learning	Linear SVM	-2-8 Char-gram -1-3 word-gram -Word skip-gram -Brown clusters	Linear SVM	A: 78% (char-4-gram) A: 77.5% (unigram)
Karamollaoglu et.al [20]	English	No specific target group	Stemmed Removed http, hashtag, unnecessary punctuations and stop words	0-1 (hate, non-hate)	Machine learning	Naïve Bayes and Vector Space Model	Word-grams	Naïve Bayes and Vector Space Model	A: %81

Pitsilis et.al [21]	English	Sexual Race	Word-based frequency vectorization	0-1 (hate, non-hate)	Deep learning	Ensemble of LSTM-based classifiers	Users' tendency to post hateful messages	Ensemble of LSTM-based classifiers	P: 0.9305, R: 0.9334 F: 0.9320
Alorainy et.al. [22]	English	Religion Disability Race Sexual	Paragraph2Vec	0-2 (Hate speech, offensive, none)	Deep learning	Multilayer Perceptron	Othering feature set (dependency relationship, PoS, list of English pronouns)	Multilayer Perceptron	F: 0.99
Basile et.al. [23]	English and Spanish	Sexual Immigrant		0-1 (hate, non-hate)	Machine learning Deep Learning	SVM with RBF kernel, Recurrent Neural Networks, LSTM model	Aggressiveness attitude, target range	SVM with RBF kernel	F: 0.7050 for Spanish, F: 0.57 for English
Ayan et.al. [24]	English	Religion		0-1 (hate, non-hate)	Sentiment Analysis	Ridge regression and Naive Bayes		Ridge Regression in F1g criterion	A: 96.9%

In the literature, classification algorithms have been used frequently until recently. Although some neural network studies have recently been reported to give better results, we wanted to start this study with more commonly used algorithms, since there has never been a search for hate speech in Turkish. In our study, we aimed to build a generalizable model and handled Turkish tweets about topics that can potentially include race- and sex-oriented hate speech. We created feature sets involving word-level and character-level n-grams and POS tagging features and applied some machine learning based classifiers to detect hate speech.

CHAPTER 3

METHODOLOGY

3.1 Data Collection

According to the Minority Right Group International Report of 2007, there are several ethnic and linguistic minorities in Turkey including Kurds, Caucasians, Laz, Roma and Arabs. Among those groups, the Kurdish population is one of the biggest ethnic groups within Turkey. Armed conflict between Turkish army authorities and Kurdistan Worker's party since the late 1980s have deteriorated the relationship between two communities with the impact of the nationalism movement all over the world [5]. One of the negative things said about Kurds who mostly living in east of Turkey is that they do not pay their electricity bills. It is believed there is a system that divides the cost of illegal use of electricity to all subscribers in Turkey. In the light of this information, "KürtünFaturasınıÖdemiyorum(IDontPayBillOfKurd)" hashtagged tweets are collected and included to this study.

In addition to ethnic and linguistic minorities in Turkey listed above, an emerging issue has been the migration flows from Syria as a result of the Syrian Civil War that broke out in 2011. Since then, Turkey has been a transit location and a place for permanent settlement where it is reported to be the most refugee-hosting country that receives about 3 million of Syrians. This rate is much higher than the in other neighbourhood countries especially EU countries, USA or Canada. While Turkey become home to foreigners than ever before, Turkish soldiers were sent to Syria to create safe zones in Turkey - Syria border. The fact that many Turkish soldiers were martyred in these operations increased hatred for Syrians especially for Syrian men.

The biggest religious minorities in Turkey are Armenians, Jews and Rums, however, due to limitations of our research we just focused on Armenians. Even though Armenians are one of the autochthonous populations of Anatolia, the population in this region has dramatically decreased during the First World War. Catalyst for public debates on hate crimes against Armenians in Turkey was the murder of

Armenian journalist and human rights activist Hrant Dink in 2007 where hate crimes is perceived as an issue which is in need of urgent attention. When the French National Assembly recognized the killing of Armenians in Ottoman Empire as genocide in 2001, the Armenian genocide came to the agenda and caused the discussion in the society. With the recognition of the genocide by the German parliament in 2016, the genocide debates came up again. In the light of this information , we collect data about Armenians in our study.

Within the last decade, hate speech and hate crime towards woman have also attracted more attention and have become part of public discussion in Turkey. Especially the incidents where two women were suffered from violence in public transport and one woman suffered from sexual assault due to their dressing style led to a demonstration organized by women's rights activists; called “KıyafetimeKarışma (Hands off my outfit)” [25]. The data for hate crime against women used in this study are Turkish tweets that were collected for a two-week window following the start of the demonstration with a specific hashtag used for the demonstration.

In order to inspect sexism, we have also covered tweets regarding LGBT community who faces several obstacles in social life and hate crime incidents. The tweets with the hashtag #orlando which is used to discuss the incident occurred in June 2016, in Orlando, Florida where an armed attack occurred in a gay nightclub have been collected. Whole list of hashtags can be seen in Table 3.1.

Briefly, in this study, we gathered tweets about Armenians, Kurds and Syrians to inspect racism; women and LGBTI to inspect sexism. We crawled the Turkish tweets via Twitter Search API [26] that were posted between April and August of 2016 and contain one of the hashtags listed in Table 2.1. For all topics, completely irrelevant tweets (for example, tweets for news or about completely different subject), tweets contain just symbols, hashtags and URLs, retweets, tweets written in another language were removed. After all, we have ended up with 4081 tweets. In each dataset, we have unequal instances for different classes. So, the datasets are imbalanced. Tweets numbers, hateful tweet numbers and collected hashtags according to topics are presented in Table 3.1.

Table 3.1 Distribution of tweets

Topics	Total tweets (#)	Hateful tweets (#)	Hashtags
Armenians	952	224	#ErmeniSoykırımı(ArmenianGenocide), #SoykırımYalanınaDURde(SaySTOPtheGenocideFalsehood),
Kurds	692	184	#KürtlerTehcirEdilmelidir(KurdsShouldBeMigrated), #KürtünFaturasınıÖdemiyorum(IDontPayBillOfKurd)
Syrians	956	244	#Suriye , #SuriyeliCocukOlmak(BeingSyrianChild), #SuriyelilerDinKardeşimizdir(SyriansAreOurCoreligionist) #SuriyedeSoykırımYapılıyor(ThereIsAGenocideInSyria), #Suriyelileriİstemiyorum(IDontWantSyrians)
Women	1285	200	#KıyafetimeKarışma(HandsOffMyOutfit)
LGBTI	196	32	#LGBT, #loveislove, #Orlando
Total	4081	884	

3.2 Data Preprocessing

Even though we collected the tweets with specific hashtags, it contained some number of noisy posts that do not belong to our context. Therefore, we first annotated the related posts and later label them as hateful or not. Three annotators were employed for labelling and the tweets with less than 66 percent agreement were removed.

During the pre-processing phase, we lowercased each tweet, corrected some of wrong written words in the tweets and stemmed them using the stemmer algorithm of Zemberek library that is an NLP library for Turkish [27].

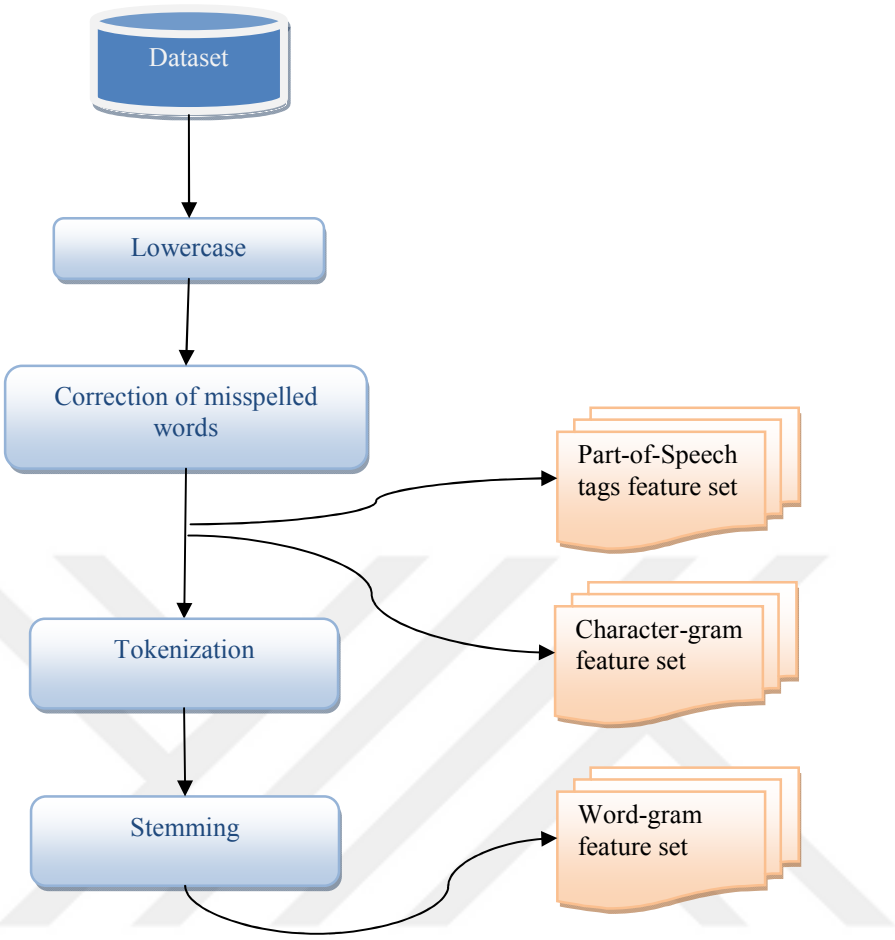


Figure 3.1 Data Preprocessing Phases

1. Lowercase: All tweets are lowercased via built-in toLowerCase() method of Java.

2. Correction of misspelled words: Misspelled words are as possible as corrected by Zemberek NLP library. After this step, part of speech tagging has done by Zemberek and character-grams are created by the method we wrote.

3. Tokenization: Tweets are tokenized by Zemberek.

4. Stemming: Tokens are stemmed by again Zemberek. At the end of this stage word-gram feature set is created. N-grams are produced by LingPipe which is a Java toolkit for text processing [28].

3.3 Feature Construction

Three feature set are used in this study: Word n-grams, part of speech tags and character n-grams. Character n-grams (of order 2–5) and word n-grams (of order 1–3) are extracted using LingPipe which is a suite of Java libraries for the linguistic analysis of human language. Word grams are created after stemming. All were weighted by their TF-IDF scores. For each target group, TF-IDF values of n-grams are calculated and the highest 50 and 100 TF-IDF-valued n-grams are used in the feature sets in different experiments.

The other feature set covers part-of-speech tags of the words. The adjective, adverb and noun counts for each tweet are computed via Zemberek NLP library. In our dataset, different hashtags yielded different number of instances are seen in Table 3.1. Each dataset is divided into train (3/4) and test (1/4) sets by protecting total hate ratio at %23 for Armenians, %26 for Kurds, %25 for Syrians, %15 for women and %16 for LGBTI in accordance with the natural hate speech ratio in the tweets with the hashtags covered in the study.

3.4 Model Construction

In this study, supervised machine learning based classification algorithms namely, SVM with linear kernel, Logistic Regression, linear Logistic Regression, Random Forests and Naive Bayes were applied. Experiments were performed using Java-based machine learning software WEKA [29].

Besides these, we tried to determine the effectiveness of each individual model on classifying cyber hate for other target groups with the aim of investigating possibility of developing a more generalizable model of cyber hate and identifying a possible lexical baseline for hate speech in Turkish. For this purpose, we cross-validated across all classes on an individual basis and reported the results.

CHAPTER 4

RESULTS AND DISCUSSIONS

The results of the classification experiments are provided using standard text classification measures of: precision, recall and accuracy. We have run different experiments with 3 main features sets; word n-grams, character n-grams and PoS tags.

While working with the first feature set, word n-grams, we worked with top 50 (Table 4.1) and top 100 TF-IDF valued features and obtained the highest accuracies with top 100 features (Table 4.2). The classifier that gives the most accurate result differs depending on the target groups we are working on. For Armenians, Kurds and LGBTI datasets, Linear Logistic Regression gives the best results; on the other hand, for Syrians dataset, it is SVM and for women Random Forest performs the best.

Table 4.1 Word-level n-gram test results with highest 50 features. Values are precision, recall and accuracy.

	Armenians	Kurds	Syrians	Women	LGBTI
NB	0.7420.752 %75.21	0.692 0.717 %71.68	0.6650.678 %67.78	0.762 0.711 %71.11	0.807 0.837 %83.67
RF	0.8160.824 %82.35	0.658 0.723 %72.25	0.669 0.720 %71.97	**0.813 0.848 %84.78**	**0.926 0.918 %91.84**
LLR	0.8240.828 %82.77	0 0.734 %73.41	0 0.745 %74.48	0.7900.835 %83.54	0.926 0.918 %91.84
LR	0.7790.798 %79.83	0.702 0.734 %73.41	0.674 0.728 %72.80	0.806 0.845 %84.47	0.837 0.837 %83.67
SVM	**0.846 0.840 %84.03**	**0.706 0.740 %73.99**	**0.724 0.753 %75.31**	0.799 0.845 %84.47	0.909 0.898 %89.80

Table 4.2 Word-level n-gram results with highest 100 features. Values are precision, recall and accuracy.

	Armenians	Kurds	Syrians	Women	LGBTI
NB	0.752 0.744 %74.37	0.684 0.711 %71.10	0.642 0.644 %64.44	0.766 0.702 %70.19	0.807 0.837 %83.67
RF	0.829 0.832 %83.19	0.658 0.723 %72.25	0.660 0.707 %70.71	**0.846 0.863 %86.34**	0.909 0.898 %89.80
LLR	**0.854 0.849 %84.87**	**0.722 0.751 %75.14**	0.679 0.720 %71.97	0.820 0.851 %85.10	**0.926 0.918 %91.84**
LR	0.806 0.815 %81.51	0.688 0.711 %71.10	0.660 0.669 %66.95	0.800 0.817 %81.68	0.807 0.816 %81.63
SVM	0.846 0.840 %84.03	0.680 0.734 %73.41	**0.715 0.753 %75.31**	0.825 0.854 %85.40	0.836 0.857 %85.71

We have continued our experiments with word level n-grams with cross validation and we cross-validated across all classes on an individual basis - training on one and testing on another. For all topics, both top 50 and 100 features were used. The best results were obtained in the experiments where the model was trained on Armenian datasets with highest 100 features. These results are given in Table 4.3. The other cross test results are given in appendix.

If we compare the results in Table 4.2 and Table 4.3, results for Kurds and Syrians are close in both; for women, latter gives better results with %87.58; for LGBTI, latter cannot exceeds the accuracy of %91.84 in the former. In other words, it is possible to obtain even better results for identifying hate speech towards women when the model was trained on Armenian dataset and the model trained on Armenian datasets can give similarly accurate results for other ethnic groups.

Table 4.3 Cross test results: Created models with Armenians dataset are applied on the other four datasets with highest 100 features. Values are precision, recall and accuracy.

Models belong to Armenians	Kurds	Syrians	Women	LGBTI
NB	0.644 0.618 %61.85	0.649 0.657 %65.69	0.748 0.730 %72.98	0.807 0.837 %83.67
RF	0.658 0.728 %72.83	0.706 0.749 %74.90	0.819 0.851 %85.09	**0.878 0.857 %85.71**
LLR	0 0.734 %73.41	**0.787 0.766 %76.57**	0.855 0.866 %86.65	0 0.837 %83.67
LR	0.667 0.682 %68.21	0.694 0.736 %73.64	0.815 0.845 %84.47	0.830 0.776 %77.55
SVM	**0.706 0.740 %73.99**	0.778 0.762 %76.15	**0.881 0.876 %87.58**	0.836 0.857 %85.71

Obtaining high results with cross validation experiments, we investigated the common words used in hate speeches toward different target groups and computed the number of overlapping n-grams in top weighted words (Table 4.4). It has been observed that there are several common words in hate speech towards Syrians, Kurds and Armenians which are made up of discriminative racist context. However, it is not the same for sexist tweets. N-grams used in hate speech towards LGBTI do not have much in common with the n-grams in hate speech towards women. In accordance with the cross validation results we obtained, the hate speech vocabulary of women dataset is most similar to the Armenians' dataset.

Table 4.4: Common n-gram counts between datasets (respectively in highest 50 and highest 100 TF-IDF-valued n-grams)

	Armenians	Kurds	Syrians	Women	LGBTI
Armenians		30 - 63	34 - 67	30 - 56	27 - 45
Kurds	30 - 63		32 - 61	27 - 52	24 - 42
Syrians	34 - 67	32 - 61		29 - 50	26 - 47
Women	30 - 56	27 - 52	29 - 50		31 - 53
LGBTI	27 – 45	24 - 42	26 - 47	31 - 53	

When the n-grams that exist in top 100 features of Armenian dataset are analyzed further, it is observed that there are several good hate indicators towards women in this set like hain(traitor), p*ç(b*stard), köpek(dog), or*spu (b*tch), s*kmek (f*ck). It reveals the usage of sexist contents in hate speech disregarding the target group.

We collected tweets with the hashtag for the "Kıyafetime Karışma (Hands off my outfit)" for women dataset, so in the top 100 features there are some words from clothing theme: giymek(wearing), şort(shorts), etek(skirt), başörtüsü(scarf), mini-etek(mini-skirt), çarşaf(chador). Because of this type of words, words that may be good hate indicators like çıplak(naked), g*t(a**), p*ç(bast**d), köpek(dog), or***u(b*tch) could not take place in the top 100 features. Thanks to the words like p*ç(bast**d), köpek(dog), or***u(b*tch) perceived as hate indicators in Armenian dataset too, better results on women dataset are obtained in cross validation. This result encourages us to enrich training datasets with the contents belonging to different target groups for the future studies.

Another finding in here is that the hate speech towards women who are criticized and attacked because of topics other than clothing can be better comprehended by adding different tweets to dataset with different hashtags. Diversity of the hashtags is the key point to generate a comprehensive model that can identify hate speech with different arguments.

In terms of content, it is expected that the Armenian set, which has very common hate speech content with women dataset, did not produce very successful results in the LGBTI set. Hate speech towards LGBTI is moving on the concept of "masculinity" in Turkey which has male-dominated society. It can be easily understood by these words taking place top 100 features: gibi (like), adam(man), erkek(male), bay(male), ç*k sallamak(shaking d**k), erkek-adam (real man).

Another finding can be stated as the challenge introduced by the hate speeches published towards the people who published some hateful contents towards target groups. Such contents lead neutral words to be learned as the indicator of hate speech and increases false positives.

The other challenge is about wrong written tweets. During the stemming the words via Zemberek NLP library, the library was able to partially correct the misspelled words. Since the stemming algorithms can not function properly for misspelled words, stemming leads those words to be lost. The amount of words lost during the preprocessing process lead us to question the effectiveness of stemming process for tweets where typos are quite common.

The second feature set that is based on part of speech did not give accurate results and did not improve the result when used with the other features (Table 4.5).

Table 4.5 Test results with PoS tagging features. Values are precision, recall and accuracy.

	Armenians	Kurds	Syrians	Women	LGBTI
NB	0 0.765 %76.47	0.538 0.728 %72.83	0.554 0.741 %74.06	0 0.845 %84.47	0.694 0.796 %79.59
RF	0.598 0.723 %72.27	0.658 0.723 %72.25	0.628 0.699 %69.88	0.764 0.832 %83.23	0.722 0.653 %65.31
LLR	0 0.765 %76.47	0.538 0.728 %72.83	0 0.745 %74.48	0 0.845 %84.47	0 0.837 %83.67
LR	0.823 0.769 %76.89	0.538 0.728 %72.83	0 0.745 %74.48	0 0.845 %84.47	0.691 0.776 %77.55
SVM	0 0.765 %76.47	0.538 0.728 %72.83	0 0.745 %74.48	0 0.845 %84.47	0.694 0.796 %79.59

However, it was possible to improve the results for three datasets when 5-character-gram features were extracted. The best results are achieved with highest 50 TF-IDF-valued character 5-grams in three datasets. SVM is the most successful one in 5-grams (Table 4.6). The other results of character-gram experiment are placed in appendix.

Table 4.6 Character 5-gram results with top 50 features. Values are precision, recall and accuracy.

	Kurds	Syrians	Armenians	Women	LGBTI
NB	0.663 0.711 %71.10	0.667 0.661 %66.11	0.745 0.748 %74.79	0.768 0.683 %68.32	0.817 0.735 %73.47
RF	0.731 0.757 %75.72	0.661 0.724 %72.39	0.734 0.769 %76.89	0.781 0.820 %81.99	0.891 0.898 %89.80
LLR.	**0.765 0.769 %76.88**	0.681 0.741 %74.06	0.774 0.786 %78.57	**0.793 0.845 %84.47**	0.894 0.898 %89.80
LR	0.734 0.757 %75.72	0.650 0.720 %71.97	0.736 0.769 %76.89	0.796 0.835 %83.54	0.854 0.837 %83.67
SVM	0.755 0.763 %76.30	**0.713 0.749 %74.90**	**0.792 0.790 %78.99**	0 0.845 %84.47	**0.943 0.939 %93.88**

Results especially belonged to cross-validations show that, racist speeches about selected topics in Turkey are pretty similar which makes it possible to establish a lexical baseline for racism. In addition to this, models belong to the Armenian dataset

which includes both racial and sexual hatred together gives better accuracies in cross validations.

Table 4.7 Some of common words between datasets

Datasets	Common word examples
Armenians and Kurds	Irk (race) , ülke (country), or*spu (b*tch), Kürtler (Kurds), faşist (fascist), s*kmek (f*ck), hain (traitor), şeref (honour), Türk (Turk), vatan (homeland), p*ç (b*stard), köpek (dog), itler (dogs), devlet (government), Ermeni (Armenian)
Armenians and Syrians	Hain (traitor), ülke(country), kendi (own), Osmanlı (Ottoman), Türkiye (Turkey), millet (nation), vatan (homeland), Allah (God), Türkler(Turks), köpek (dog), soysuz (b*stard), s*kmek (f*ck), ırk(race), savunmak(defend)
Armenians and Women	İstemek (want), kim (who), şeref (honor), pis (dirty), ülke (country), zaman (time), bırakmak (give up), insan (human), görmek(see), bakmak(look)
Armenian and LGBTI	İnsan (human), katliam(massacre), Dünya (World), kınamak (condemn), katletmek(kill)
Kurds and Syrians	Hain (traitor), ülke (country), vatan (homeland), insan(human), çocuklar(children), Türk(Turk), yaşamak(live), Allah (God), köpek (dog), savaşmak (fight), savaş (war), s*kmek (f*ck), sevmek (love), kardeşler (siblings), ırk (race), Müslüman (Muslim), sahip (owner), gitmek (go)
Kurds and Women	Bakmak (look), istemek(want), şeref (honor), ülke(country), fitne (sedition), yaşamak(live), herkes(everyone)
Kurds and LGBTI	İnsan(human), sevmek(love), herkes(everyone), terör (terror)
Syrians and Women	Ülke(country), kendi(own), insan (human), yaşamak(live), gitmek(go), bakmak (look), sanmak(suppose)
Syrians and LGBTI	Dünya (World), yazık(shame), karşı (against), pislik (dirt), sevmek (love), sanmak(suppose), büyük (big)
Women and LGBTI	İğrenç(disgusting), erkek (man), beyin (brain), kendi (own), hayat (life), saygı (respect), özgür (free), karışmak (meddle), başka(other)

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this study, hate speech in means of racism and sexism and existence of similarities between hate speeches towards different target groups are investigated via some surface level features in Turkish tweets. We applied standard lexical features and several classifiers for this task. The best result was obtained by a character 5-gram model with SVM achieving 93% accuracy for LGBTI dataset.

The results show that hate speech characteristics towards women and LGBTI are different from each other which lead poor results in cross-validations. However, it is possible to obtain promising results when the model was trained on one of ethnic groups and tested on another which highlights existence of some common 'othering' terms for racism in Turkish tweets.

A limitation of the study is extent of the data. To obtain better results, scope of datasets should be increased by adding different tweets contain different hashtags about target groups, especially for women dataset holding tweets about clothing-theme-event. Another limitation is misspelled words. Zemberek NLP library has correction capability but this is not sufficient. Stemming the misspelled words produces "unknown" tokens and leads loss of information. Using these misspelled words efficiently, better results could be obtained.

Tweets contain profanity and slang phrases like p*ç(bastard), şerefsiz(bastard) are annotated as hateful by both labelers and algorithms. While this situation increases recall, on the other hand it causes misclassification of non-hateful tweets that contain those phrases revealing that distinguishing profanity from hate speech is a very challenging task. In future, it is aimed to cover typed dependencies within the tweet text to overcome this problem. In addition, we would like to build a dictionary automatically for terms used in hate speeches for different target groups that can facilitate the identification of hateful contents and improve the results of classification task.

REFERENCES

[1] Council of Europe. CoE Hate Speech Retrieved December 17, 2018, from https://www.coe.int/en/web/freedom-expression/hate-speech.

[2] Rayburn,N.R.,Mendoza,M & Davison, G.C. Bystanders' Perception of Perpetrators and Victims of Hate Crimes, Journal Of Interpersonal Violence, 18, 1055-1074 . 2003

[3] Davidson, T, Warmsley, D, Macy, M, Weber, I, Automated Hate Speech Detection and the Problem of Offensive Language, In the Proceedings of ICWSM 2017, 2017.

[4] Grand National Assembly of Turkey . Law Amending Various Laws for the Development of Fundamental Rights and Freedoms. (Law No 6529)Article 15. 2014. Retrieved from https://www.tbmm.gov.tr/kanunlar/k6529.html.

[5] Göktan D.Ü.,"Hate Crime in Turkey: Implications of Collective Action, Media Representations and Policy Making", Cambridge Scholars Publishing, 2017.

[6] İnceoğlu Y, "Nefret Söylemi ve Nefret Suçları", Ayrıntı Publishing, 2012.

[7] Global Social Media Users Pass 3.5 Billion (n.d.). Retrieved August 20,2019 from.https://wearesocial.com/blog/2019/07/global-social-media-users-pass-3-5-billion

[8] Gümüş, B., & Dural, A. B. Othering through hate speech: the Turkish-Islamist (V) Akit newspaper as a case study. *Turkish Studies*, *13*(3), 489-507. 2012

[9] Arcan, H. E. Ethnic Conflicts and the Role of the Media: The Case of Turkish Media. *Mediterranean Journal of Social Sciences*, *4*(10), 338. 2013

[10] Göl, A. Imagining the Turkish nation through ‘othering' Armenians. *Nations and Nationalism*, *11*(1), 121-139. 2005

.

[11] Dondurucu, Z. B., & Uluçay, A. P. Yeni medya ortamlarında nefret söylemi: Eşcinsellere yönelik nefret söylemi içeren videoların Youtube üzerinden incelenmesi. *International Journal of Social Sciences and Education Research*, *1*(3), 875-902. 2015

[12] Introduction to Machine Learning. Retrieved March 2019 from https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/lecture-videos/lecture-11-introduction-to-machine-learning/

[13] Manning C.D.,Raghavan P.,Schütze H. "Introduction to Information Retrieval", Cambridge University Press. 2008.

[14] Linear Regression.(n.d.). Retrieved August 20,2019 from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[15] Nguyen, D., Smith, N. A., & Rosé, C. P. Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 115-123). Association for Computational Linguistics. 2011

[16] Support Vector Machine- Introduction to Machine Learning Algorithms (n.d.). Retrieved August 20,2019 from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[17] Burnap P.,Rana O.F., Avis N., Williams M., Housley W.,Edwards A., and others. Detecting tension in online communities with computational Twitter analysis. Technol. Forecast. Soc. Change, 2013, (95), pp. 96-108, July 2015.

[18] Burnap P.& Williams M.L., Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci., (5),(1), p.11, December,2016.

[19] Malmasi,S. and Zampieri,M., Detecting Hate Speech in Social Media, In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), 2017.

[20] Karamollaoğlu, H., Doğru, İ. A., & Utku, A. (2017, May). Identification of shares containing offensive charge in social media. In *2017 25th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[21] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, *48*(12), 4730-4742. 2018

[22] Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. "The Enemy Among Us": Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web (TWEB)*, *13*(3), 14. 2019

[23] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., and others.. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 54-63). (2019)

[24] Ayan, B., Kuyumcu, B., & Ceylan, B. Twitter Üzerindeki İslamofobik Twitlerin Duygu Analizi ile Tespiti. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, *7*(2), 495-502.

[25] Hands off my outfit. Sputnik News. Retrieved January 21,2019 from https://sputniknews.com/middleeast/201707301056011737-istanbul-march-womens-rights-clothing/

[26] Twitter Search API.(n.d.) Retrieved January 21,2019 from https://developer.twitter.com/en/docs/tweets/search/api-reference.

[27] Akın A. A. & Dündar M., Zemberek, an Open Source NLP Framework for Turkic languages. Structure, (10), pp. 1-5, 2007.

[28] Ling-Pipe: Text Processing Toolkit. Retrieved August 10, 2019 from http://alias-i.com/

[29] WEKA machine learning toolkit.(n.d.) Retrieved January 12,2019 from http://www.cs.waikato.ac.nz/ml/weka.

APPENDICES

Appendix A – Experiment Results

Table A.1 Cross test results: Created models with Armenians dataset are applied on the other four datasets with highest 50 features. Values are precision, recall and accuracy.

Models belong to Armenians	Kurds	Syrians	Women	LGBTI
Naïve Bayes	0.606 0.595 %59.54	0.657 0.682 %68.20	0.762 0.770 %77.02	0.788 0.816 %81.63
Random Forest	0.609 0.682 %68.21	0.697 0.745 %74.47	0.770 0.835 %83.54	0.794 0.837 %83.67
Linear Logistic Reg.	0.680 0.734 %73.41	0 0.745 %74.47	0.810 0.848 %84.78	0 0.837 %83.67
Logistic Reg.	0.640 0.705 %70.52	0.572 0.707 %70.71	0.781 0.835 %83.54	0.747 0.796 %79.59
SVM	0.674 0.734 %73.41	0.554 0.741 %74.06	0.880 0.860 %86.02	0.794 0.837 %83.67

Table A.2 Cross test results: Created models with Kurds dataset are applied on the other four datasets with highest 50 features. Values are precision, recall and accuracy.

Models belong to Kurds	Armenians	Syrians	Women	LGBTI
NB	0.702 0.723 %72.27	0.685 0.707 %70.71	0.738 0.764 %76.40	0.755 0.755 %75.51
RF	0.708 0.744 %74.37	0.664 0.711 %71.13	0.806 0.842 %84.16	0.794 0.837 %83.67
LLR	0.736 0.773 %77.31	0.715 0.753 %75.31	0.889 0.873 %87.27	0.878 0.857 %85.71
LR	0.694 0.744 %74.37	0.703 0.745 %74.48	0.860 0.866 %86.65	0.736 0.776 %77.55
SVM	0.774 0.786 %78.57	0.704 0.749 %74.90	0.889 0.873 %87.27	0.878 0.857 %85.71

Table A.3 Cross test results: Created models with Kurds dataset are applied on the other four datasets with highest 100 features. Values are precision, recall and accuracy.

Models belong to Kurds	Armenians	Syrians	Women	LGBTI
NB	0.7130.718 %71.85	0.668 0.682 %68.20	0.741 0.733 %73.29	0.755 0.755 %75.51
RF	0.783 0.803 %80.25	0.672 0.720 %71.97	0.824 0.851 %85.09	0.794 0.837 %83.67
LLR	0.818 0.819 %81.93	0.741 0.762 %76.15	0.864 0.873 %87.27	0.694 0.796 %79.59
LR	0.727 0.752 %75.21	0.694 0.736 %73.64	0.817 0.845 %84.47	0.714 0.714 %71.43
SVM	0.794 0.807 %80.67	0.7480.762 %76.15	0.889 0.873 %87.27	0.794 0.837 %83.67

Table A.4 Cross test results: Created models with Syrians dataset are applied on the other four datasets with highest 50 features. Values are precision, recall and accuracy.

Models belong to Syrians	Armenians	Kurds	Women	LGBTI
NB	0.666 0.706 %70.59	0.598 0.590 %58.96	0.722 0.727 %72.67	0.764 0.816 %81.63
RF	0.666 0.731 %73.11	0.622 0.688 %68.79	0.750 0.829 %82.92	0 0.837 %83.67
LLR	0.641 0.718 %71.85	0.808 0.740 %73.99	0 0.845 %84.48	0 0.837 %83.67
LR	0.674 0.723 %72.27	0.646 0.694 %69.36	0.767 0.842 %84.16	0.788 0.816 %81.63
SVM	0 0.765 %76.47	0.808 0.740 %73.99	0.871 0.848 %84.78	0 0.837 %83.67

Table A.5 Cross test results: Created models with Syrians dataset are applied on the other four datasets with highest 100 features. Values are precision, recall and accuracy.

Models belong to Syrians	Armenians	Kurds	Women	LGBTI
NB	0.705 0.710 %71.01	0.645 0.630 %63.01	0.731 0.727 %72.67	0.788 0.816 %81.63
RF	0.772 0.794 %79.41	0.612 0.705 %70.52	0.819 0.851 %85.09	0.893 0.878 %87.76
LLR	0.749 0.777 %77.73	0.808 0.740 %73.99	0.713 0.839 %83.85	0.878 0.857 %85.71
LR	0.731 0.756 %75.63	0.715 0.723 %72.25	0.811 0.842 %84.16	0.764 0.714 %71.43
SVM	0.792 0.790 %78.99	0.818 0.757 %75.72	0.810 0.848 %84.78	0.836 0.857 %85.71

Table A.6 Cross test results: Created models with Women dataset are applied on the other four datasets with highest 50 features. Values are precision, recall and accuracy.

Models belong to Women	Armenians	Kurds	Syrians	LGBTI
NB	0.691 0.723 %72.27	0.620 0.630 %63.01	0.646 0.678 %67.78	0.755 0.755 %75.51
RF	0.686 0.731 %73.11	0.631 0.682 %68.21	0.625 0.695 %69.46	0.794 0.837 %83.67
LLR	0.695 0.748 %74.79	0.718 0.746 %74.57	0 0.745 %74.48	0.878 0.857 %85.71
LR	0.697 0.739 %73.95	0.646 0.705 %70.52	0.633 0.728 %72.80	0.736 0.776 %77.55
SVM	0.801 0.786 %78.57	0.648 0.728 %72.83	0 0.745 %74.48	0.878 0.857 %85.71

Table A.7 Cross test results: Created models with Women dataset are applied on the other four datasets with highest 100 features. Values are precision, recall and accuracy.

Models belong to Women	Armenians	Kurds	Syrians	LGBTI
NB	0.655 0.664 %66.39	0.618 0.618 %61.85	0.666 0.690 %69.04	0.691 0.776 %77.55
RF	0.677 0.731 %73.11	0.689 0.734 %73.41	0.631 0.695 %69.46	0.794 0.837 %83.67
LLR	0.736 0.733 %77.31	0.628 0.728 %72.83	0 0.745 %74.48	0 0.837 %83.67
LR.	0.688 0.718 %71.85	0.657 0.688 %68.79	0.665 0.720 %71.97	0.787 0.776 %77.55
SVM	0.792 0.790%78.99	0.605 0.723 %72.25	0.812 0.749 %74.90	0.697 0.816 %81.63

Table A.8 Cross test results: Created models with LGBTI dataset are applied on the other four datasets with highest 50 features. Values are precision, recall and accuracy.

Models belong to LGBTI	Armenians	Kurds	Syrians	Women
NB	0.664 0.681 %68.07	0.626 0.642 %64.16	0.648 0.686 %68.62	0.770 0.770 %77.02
RF	0.654 0.710 %71.01	0.611 0.665 %66.47	0.621 0.690 %69.04	0.796 0.842 %84.16
LLR	0.584 0.761 %76.05	0.718 0.746 %74.57	0.628 0.732 %73.22	0.825 0.851 %85.09
LR	0.638 0.714 %71.43	0.658 0.711 %71.10	0.625 0.724 %72.39	0.801 0.845 %84.47
SVM	0 0.765 %76.47	0.674 0.734 %73.41	0.729 0.749 %74.90	0.810 0.848 %84.78

Table A.9 Cross test results: Created models with LGBTI dataset are applied on the other four datasets with highest 100 features. Values are precision, recall and accuracy.

Models belong to LGBTI	Armenians	Kurds	Syrians	Women
NB	0.662 0.676 %67.65	0.609 0.601 %60.12	0.655 0.653 %65.27	0.755 0.739 %73.91
RF	0.677 0.731 %73.11	0.640 0.699 %69.94	0.615 0.690 %69.04	0.790 0.842 %84.16
LLR.	0.584 0.761 %76.05	0.669 0.728 %72.83	0.812 0.749 %74.90	0.779 0.842 %84.16
LR.	0.655 0.718 %71.85	0.652 0.688 %68.79	0.634 0.699 %69.88	0.763 0.820 %81.99
SVM	0.582 0.752 %75.21	0.676 0.734 %73.41	0.729 0.749 %74.90	0.825 0.851 %85.09

Table A.10 For all topics character 5-gram results with top 100 features. Values are precision, recall and accuracy.

	Kurds	Syrians	Armenians	Women	LGBTI
Naïve Bayes	0.647 0.688 %68.79	0.673 0.623 %62.34	0.759 0.761 %76.05	0.798 0.711 %71.11	0.845 0.816 %81.63
Random Forest	0.716 0.740 %73.99	0.652 0.707 %70.71	0.759 0.786 %78.57	0.820 0.848 %84.78	0.891 0.898 %89.80
Linear Logistic Reg.	0.700 0.734 %73.41	0.676 0.736 %73.64	0.776 0.794 %79.41	0.810 0.845 %84.47	0.918 0.918 %91.84
Logistic Reg.	0.681 0.676 %67.63	0.679 0.699 %69.88	0.777 0.794 %79.41	0.821 0.839 %83.85	0.892 0.878 %87.76
SVM	0.690 0.723 %72.25	0.684 0.741 %74.06	0.800 0.811 %81.09	0.849 0.860 %86.03	0.868 0.878 %87.76

Table A.11 Test results with character-level n-gram features using highest 50 features in Armenians dataset. Values are precision, recall and accuracy.

Armenians Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.740 0.769 %76.89	0.743 0.765 %76.47	0.724 0.744 %74.37	0.745 0.748 %74.79
Random Forest	0.830 0.782 %78.15	0.799 0.794 %79.41	0.723 0.765 %76.47	0.734 0.769 %76.89
Linear Logistic Reg.	0.584 0.761 %76.05	0.704 0.765 %76.47	0.706 0.765 %76.47	0.774 0.786 %78.57
Logistic Reg.	0.752 0.782 %78.15	0.772 0.794 %79.41	0.753 0.782 %78.15	0.736 0.769 %76.89
SVM	0 0.765 %76.47	0.782 0.777 %77.73	0.835 0.790 %78.99	0.792 0.790 %78.99

Table A.12 Test results with character-level n-gram features using highest 100 features in Armenians dataset..Values are precision, recall and accuracy.

Armenians Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.745 0.761 %76.05	0.762 0.765 %76.47	0.757 0.752 %75.21	0.759 0.761 %76.05
Random Forest	0.782 0.777 %77.73	0.682 0.761 %76.05	0.786 0.803 %80.25	0.759 0.786 %78.57
Linear Logistic Reg.	0.682 0.761 %76.05	0.704 0.765 %76.47	0.816 0.807 %80.67	0.776 0.794 %79.41
Logistic Reg.	0.740 0.769 %76.89	0.730 0.752 %75.21	0.745 0.769 %76.89	0.777 0.794 %79.41
SVM	0.828 0.807 %80.67	0.772 0.794 %79.41	0.766 0.790 %78.99	0.800 0.811 %81.09

Table A.13 Test results with character-level n-gram features using highest 50 features in Kurds dataset. Values are precision, recall and accuracy.

Kurds Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.604 0.601 %60.12	0.650 0.601 %60.12	0.679 0.694 %69.36	0.663 0.711 %71.10
Random Forest	0.808 0.740 %73.99	0.629 .723 %72.25	0.626 0.676 %67.63	0.731 0.757 %75.72
Linear Logistic Reg.	0.658 0.723 %72.25	0.615 0.717 %71.68	0.749 0.769 %76.88	0.765 0.769 %76.88
Logistic Reg.	0.643 0.676 %67.63	0.611 0.665 %66.47	0.699 0.728 %72.83	0.734 0.757 %75.72
SVM	0 0.734 %73.41	0.628 0.728 %72.83	0.756 0.757 %75.72	0.755 0.763 %76.30

Table A.14 Test results with character-level n-gram features using highest 100 features in Kurds dataset. Values are precision, recall and accuracy.

Kurds Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.625 0.561 %56.07	0.691 0.642 %64.16	0.662 0.682 %68.21	0.647 0.688 %68.79
Random Forest	0.721 0.740 %73.99	0.676 0.734 %73.41	0.782 0.786 %78.61	0.716 0.740 %73.99
Linear Logistic Reg.	0.634 0.694 %69.36	0.699 0.740 %73.99	0.758 0.769 %76.88	0.700 0.734 %73.41
Logistic Reg.	0.668 0.671 %67.05	0.668 0.671 %67.05	0.687 0.705 %70.52	0.681 0.676 %67.63
SVM	0.646 0.699 %69.94	0.636 0.688 %68.79	0.750 0.769 %76.88	0.690 0.723 %72.25

Table A.15 Test results with character-level n-gram features using highest 50 features in Syrians dataset. Values are precision, recall and accuracy.

Syrians Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.636 0.603 %60.25	0.612 0.435 %43.52	0.650 0.548 %54.82	0.667 0.661 %66.11
Random Forest	0.684 0.745 %74.48	0.656 0.724 %72.39	0.593 0.682 %68.20	0.661 0.724 %72.39
Linear Logistic Reg.	0 0.745 %74.48	0.554 0.741 %74.06	0.713 0.749 %74.90	0.681 0.741 %74.06
Logistic Reg.	0.634 0.699 %69.88	0.636 0.682 %68.20	0.698 0.724 %72.39	0.650 0.720 %71.97
SVM	0 0.745 %74.48	0 0.745 %74.48	0.685 0.745 %74.48	0.713 0.749 %74.90

Table A.16 Test results with character-level n-gram features using highest 100 features in Syrians dataset. Values are precision, recall and accuracy.

Syrians Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.633 0.548 %54.81	0.657 0.469 %46.86	0.644 0.561 %56.07	0.667 0.661 %66.11
Random Forest	0.552 0.732 %73.22	0.635 0.724 %72.39	0.650 0.724 %72.39	0.652 0.707 %70.71
Linear Logistic Reg.	0 0.745 %74.48	0 0.745 %74.48	0.674 0.741 %74.06	0.676 0.736 %73.64
Logistic Reg.	0.641 0.669 %66.95	0.641 0.649 %64.85	0.662 0.665 %66.53	0.679 0.699 %69.88
SVM	0.653 0.736 %73.64	0.653 0.736 %73.64	0.633 0.728 %72.80	0.684 0.741 %74.06

Table A.17 Test results with character-level n-gram features using highest 50 features in Women dataset. Values are precision, recall and accuracy.

Women Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.7788 0.730 %72.98	0.784 0.727 %72.67	0.791 0.724 %72.36	0.768 0.683 %68.32
Random Forest	0.813 0.848 %84.78	0.813 0.848 %84.78	0.786 0.835 %83.54	0.781 0.820 %81.99
Linear Logistic Reg.	0.871 0.848 %84.78	0.871 0.848 %84.787	0.793 0.845 %84.47	0.793 0.845 %84.47
Logistic Reg.	0.767 0.829 %82.92	0.771 0.832 %83.23	0.793 0.842 %84.16	0.796 0.835 %83.54
SVM	0 0.845 %84.47	0 0.845 %84.47	0 0.845 %84.47	0 0.845 %84.47

Table A.18 Test results with character-level n-gram features using highest 100 features in Women dataset. Values are precision, recall and accuracy.

Women Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.777 0.727 %72.67	0.790 0.708 %70.81	0.788 0.636 %68.63	0.798 0.711 %71.11
Random Forest	0.813 0.848 %84.78	0.825 0.851 %85.09	0.776 0.835 %83.54	0.820 0.848 %84.78
Linear Logistic Reg.	0.768 0.839 %83.85	0 0.845 %84.47	0.713 0.839 %83.85	0.810 0.845 %84.47
Logistic Reg.	0.762 0.804 %80.44	0.783 0.826 %82.61	0.792 0.820 %81.99	0.821 0.839 %83.85
SVM	0.793 0.845 %84.47	0.793 0.845 %84.47	0.825 0.851 %85.09	0.849 0.860 %86.03

Table A.19 Test results with character-level n-gram features using highest 50 features in LGBTI dataset. Values are precision, recall and accuracy.

LGBTI Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.821 0.837 %83.67	0.854 0.837 %83.67	0.882 0.857 %85.72	0.817 0.735 %73.47
Random Forest	0.697 0.816 %81.63	0.891 0.898 %89.90	0.840 0.857 %85.72	0.891 0.898 %89.80
Linear Logistic Reg.	0 0.837 %83.67	0.909 0.898 %89.80	0.909 0.898 %89.80	0.894 0.898 %89.80
Logistic Reg.	0.771 0.735 %73.47	0.807 0.816 %81.63	0.847 0.755 %75.51	0.854 0.837 %83.67
SVM	0.764 0.816 %81.63	0.891 0.898 %89.80	0.891 0.898 %89.80	0.943 0.939 %93.88

Table A.20 Test results with character-level n-gram features using highest 100 features in LGBTI dataset. Values are precision, recall and accuracy

LGBTI Dataset	Character bigram	Character trigram	Character 4-gram	Character 5-gram
Naïve Bayes	0.878 0.878 %87.76	0.904 0.898 %89.80	0.882 0.857 %85.71	0.845 0.816 %81.63
Random Forest	0 0.837 %83.67	0.866 0.878 %87.76	0.8910.898 %89.80	0.891 0.898 %89.80
Linear Logistic Reg.	0.868 0.878 %87.76	0.891 0.898 %89.80	0.807 0.837 %83.67	0.918 0.918 %91.84
Logistic Reg.	0.778 0.755 %75.51	0.894 0.898 %89.80	0.924 0.857 %85.71	0.892 0.878 %87.76
SVM	0.865 0.857 %85.72	0.878 0.878 %87.76	0.878 0.878 %87.76	0.868 0.878 %87.76

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname : Tuğba DAĞAŞAN
Date of Birth : 13.01.1987
Phone : 0506 958 74 59
E-mail : tugbadagasan@gmail.com

Photo

EDUCATION

High School : Etimesgut High School
Bachelor : Ankara University Engineering Faculty Computer Engineering

WORK EXPERIENCE

Computer Engineer: Turkish Coast Guard Command (Kasım 2011- Temmuz 2018)

Specialist Assistant : Family, Work and Social Services Ministry (Temmuz 2018-)

TOPICS OF INTEREST

- Data Mining
- Social sciences

September 2019
ANKARA

Department of Computer Engineering

Tuğba DAĞAŞAN

