DETECTING HATE SPEECH IN TURKISH TEXTS

by

Zehra Melce Hüsünbeyi

B.S., Computer Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my master thesis supervisor Assoc. Prof. Arzucan Özgür for her valuable guidance, support, and kindness. I have learned and experienced countless things since I became her student. I am also quite grateful to Assoc. Prof. Didar Akar who shared her knowledge and brought together different disciplines that led to make a connection between ideas and produce a creative study.

I thank Assist. Prof. Yusuf Yaslan for participating in my thesis committee and for his valuable revisions on the thesis. Special thanks to Assist. Prof. Canan Yıldız for her supportive attitude and providing me an opportunity that helped me develop a broader perspective to my thesis.

I have great pleasure in acknowledging my gratitude to my dearest friends Gamze Hallı, Gülşah Öztürk, Berçem Eren, Berna Erden, and Seçil Doğan. They have assisted me as per their abilities, in whatever manner possible for ensuring that good times keep flowing.

I wish to extend my special thanks to Hrant Dink Foundation for their collaboration and sharing their valuable studies which are a quite significant part of my thesis.

Finally and above all, I thank with love to my mother Melek, my father Ahmet, my brother Memduh and my aunt Dilek for their encouragement and motivation to go further, and their endless patience. It was their love that raised me up again when I got weary. My beloved brother has made a tremendous contribution by helping me reach this stage in my life and get through this agonizing period in the most positive way.

ABSTRACT

DETECTING HATE SPEECH IN TURKISH TEXTS

It is well known that prejudiced and discriminatory language is being widely used and spread through several channels such as printed or social media. The discriminatory language, in particular hate speech as its more aggressive, degrading and openly targeting form, which poses a threat to the values of democracy and human rights is a global problem that needs an immediate solution. Since we find the detection of hate speech important in the fight against hate speech, we have developed a model to detect it. For this purpose, we created a dataset by retrieving printed media news that the Hrant Dink Foundation systematically annotated in the context of hate speech from the website of the PRNet media monitoring company. To the best of our knowledge, with this study, the first model developed for Turkish language that runs on a labeled dataset is produced. In particular, the fact that most of the hate speech in printed media is based on context and implications requires a system that can detect changing discursive cues and understand the context around these discourses. With different word representations, we have examined the Hierarchical Attention Network (HAN) model, which aims to capture the changing meanings of expressions by using the hierarchical structure of the text. We studied the compatibility of our model with the problem by comparing it with Convolution Neural Network (CNN), which provided important results in text processing, and with machine learning models. In order to improve our study, we developed linguistic features for the problem based on critical discourse analysis techniques. We enhanced the HAN model using these features. Our results show that performance increases with a set of features that point out the use of 'othering language'. We believe that these feature sets created for the Turkish language will encourage new studies in the quantitative analysis of hate speech.

ÖZET

TÜRKÇE METİNLERDE NEFRET SÖYLEMİ TESPİTİ

Yazılı basın ve sosyal medya gibi birçok farklı mecrada önyargılı ve ayrımcı bir dilin kullanıldığı ve yaygınlaştığı görülmektedir. Demokrasi ve insan hakları değerlerine karşı tehdit oluşturan ayrımcı dil ve onun daha saldırgan ve aşağılayıcı, açıkça hedef gösterici şekliyle nefret söylemi acilen çözülmesi gereken küresel bir sorun teşkil etmektedir. Biz de nefret söylemiyle mücadelede önemli olan nefret söylemi tespiti için bir model geliştirdik. Bu amaçla, Hrant Dink Vakfı'nın sistematik bir şekilde nefret söylemi bağlamında annotate ettiği yazılı basın haberlerini PRNet medya takip şirketi web sitesinden çekerek bir dataset oluşturduk. Bildiğimiz kadarıyla, bu çalışmayla, etiketlenmiş bir dataset üzerinde çalışan Türkçe için geliştirilmiş ilk model üretilir. Özellikle yazılı basın haberlerindeki nefret söyleminin büyük kısmının bağlam ve imalara dayanması değişen söylemsel ipuçlarını tespit edebilen ve bu söylemlerin etrafında oluşan bağlamı anlayabilen bir sistem gerektirir. Biz de metnin hiyerarşik yapısını kullanarak ifadelerin değişen anlamlarını yakalamayı hedefleyen Hiyerarşik İlgi Ağları (HİA) modelini farklı kelime temsilleriyle inceledik. Modelimizi metin işlemede önemli sonuçlar veren Konvolüsyonel Sinir Ağları ve makine öğrenmesi modelleriyle kıyaslayarak probleme uygunluğunu tespit ettik. Çalışmamızı geliştirmek için eleştirel söylem analizi tekniklerini temel alarak probleme yönelik dilbilimsel özellikler geliştirdik. HİA modelini bu özelliklerle birlikte zenginleştirdik. Sonuçlarımız 'diğerleri dili' kullanımına işaret eden özellik kümesiyle performansın geliştiğini gösterir. Türkçe dili için oluşturulan bu özellik kümelerinin nefret söyleminin nicel analizinde yeni çalışmaları teşvik edeceğine inanıyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a_{it}	Normalized importance weight
$\overrightarrow{h}_{it}$	The forward hidden state
$\overleftarrow{h}_{it}$	The backward hidden state
r_t	The reset gate vector
u_s	The sentence level context vector
u_w	The word level context vector
w_{it}	Given a sentence with word sequence
x_t	The input sequence vector
z_t	The update gate vector
σ	Sigmoid function
$tanh$	Hyperbolic tangent function

LIST OF ACRONYMS/ABBREVIATIONS

CNN	Convolutional Neural Network
GRU	Gated Recurrent Units
HAN	Hierarchical Attention Network
IMST-UD	The ITU-METU-Sabancı UD Turkish Treebank
LSTM	Long Short Term Memory
POS	Part of Speech Tags
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
UDPipe	The Universal Dependencies Pipe

1. INTRODUCTION

Turkey is one of the members of the Council of Europe, and according to its Recommendation No. R(97) 20, "hate speech is any statement including racist hate, ethnocentrism [...] religion intolerance against minorities, immigrants or originally-immigrant groups [...] and any expressions spreading, provoking or legitimating hate." [4] Hate speech is a global problem, and it requires to be solved as soon as possible, since it harms social welfare, democracy and human rights. The use of prejudiced and discriminatory language has become popular on several channels such as printed press and social media in Turkey, and all around the world. As a result of this situation, not only national and local agents but also international institutions and large-scaled companies like Facebook and Twitter have started to pay attention to this concept. Council of Europe has signed a protocol regarding the detection of illegal hate discourse with Facebook, Microsoft, Twitter and YouTube in 2016 [5]. Also at the end of 2018, the scope of protocol has extended to include widely-used application producers such as Instagram, Google+, Snapchat, Dailymotion.

It is pivotal to detect hate speech in the fight against hate speech; however, since language is a dynamic and social system, detecting and defining discourse is a crucial problem. Manual detection of hate speech, which requires a lot of workforce, is not sustainable. Had there been a tool developed for the Turkish language that would use the data labeled by annotators the dependency on human effort would decrease. While automatizing the identification of statements which may include hate speech, one of the biggest obstacles is their context and implicit meanings which are peculiar to the groups using the language. It means that distinguishing what counts as hate speech and what does not is quite challenging while defining hate speech and managing data set, because it basically depends on context. In a defined context, discourse depends not only on socio-cultural, political and historical background, but also on ironic and implicit meanings of a language, and on identities of the target groups and their positions in society. For instance, while a sentence could be seen neutral-impartial

by a group in society, the same sentence would harm another group in that society. Therefore, to eliminate hate speech, it's necessary to define any clues in a changing discourse and to develop a systematic understanding of the context in discourse.

Since 2009, Hrant Dink Foundation has been conducting a systematic study to detect hate speech on the printed press [6]. Within the scope of the study, all national newspapers and almost 500 local newspapers have been examined. As a result of this study, a labeled ten-year data masthead has been prepared. Based on this study, we have developed a dataset for the news of 2016, 2017 and 2018, which doesn't include hate speech, by using web scraping in addition to that of the foundation's. We have developed a model making the detection of hate speech automatic through the dataset that we have produced. In our study, we have improved a model based on the Hierarchical Attention Network [7], which aims to detect changes in the meaning by using the hierarchical structure of texts, with problem-specific linguistic features. These novel linguistic methods seperate news contenting hate speech from the one not including hate speech by holding patterns of usage of othering language and objectivity/subjectivity of the news. These pretrained features utilizing paragraph embedding have been evaluated according to different feature sets by means of concatenating to Hierarchical Attention Network model.

This study aims to contribute to the promotion of respect for social, racial differences, and for human rights in the media, and to develop tools to increase active participation of civil society in the fight against hate speech and discrimination. In order to manage it, we have collaborated with Hrant Dink Foundation and collected answers to certain questions such as what the concept of hate speech, international legal paradigms, national legislation, and the relationship between hate speech and the Media are. This model which is firstly developed by annotated data for Turkish language, would promote new studies with the potential of gathering different agents and disciplines.

2. RELATED WORK

Domain specific and traditional linguistic features have a significant role in the detection of hate speech problems. Part of speech tags (POS) and typed dependency relations are some of the methods commonly used in the literature. According to Xu et al. [8], combining n-grams with POS does not significantly affect performance, while Typed Dependency Relationships greatly improve performance with the information of relationship between non-consecutive words in the detection of hate speech problem [9–11]. Burnap and Williams [12] have utilized Bayesian Logistic Regression to represent a sentence with a certain type of dependency relations.

Lexical sources have been used in several studies despite their lack of generalization abilities. Hate-related profanity and insulting expressions are used by Xiang et al. [13], Burnap and Williams [12], and Nobata et al. [11] on the assumption that hate speech contains certain negative words. In addition, Burnap and Williams [10] use a specialized lexicon that includes insults aimed at ethnic origins, slang expressions aimed at gender, and negative connotations aimed at disabilities as corresponding to specific hate speech subcategories. Razavi et al. [14] manually generated an Insulting and Abusive Language Dictionary containing hate words and sentences that manifest to different degrees. The weight of each lexical input determined by adaptive learning represents the potential level of influence in detecting hate speech. Also Gitari et al. [9] provide a lexical source of hate verbs which prod acts of violence.

Additionally, relation between sentiment analysis and hate speech have been examined and subjective language has been considered as more related with hate speech. By using rule based approach, objective sentences have been separated from subjective ones [15]. Authors also state that identifying stereotype language for several groups by utilising precise tokens like money and banking could be associated with anti-semitism, and is quite useful to develop a hate speech detection model. Another approach is regarding founding relation between user characteristics such as gender, geographic

localization and their language [16]. Also Zhong et. al. [17] have implemented offensiveness score which depends on the frequency of the usage of offensive terms and the user identifying words in the identical dependency relation.

As most promising approach, the theory of othering language has been used as a framework to identify hate speech for contents on social media [10, 18]. By using Stanford Lexical Parser, Burnap et al. [10] have presented syntactic grammatical relationships in a tweet to extract the opposition, for instance typed dependency nsubj(home, them) in "send them back home" sentence identifies relational sense between tokens and emphasizes differences between 'us' and 'them'. They have also stated that statistically significant results have been achieved especially detecting hate speech related to religious beliefs with othering feature set. According to Alorainy et al. [18], othering language theory, based on linguistics approaches such as set of in group (us) / out group (they) separation as othering terms, typed dependency relations (nsubj, dobj, nmod, det, advmod, and compound) and part of speech tags (nouns, adjectives, adverbs, verbs) in hate speech samples that include 'two-sided' pronoun (us vs them), outperforms state of the art in the detection of hate speech literature.

As well as Linguistics related features as mentioned in above, surface features such as bag-of-words (BOW), n-grams, local features e.g., TF-IDF weights of tokens, and rule-based approaches such as errors in spelling, number of punctuation have been performed in the detection of hate speech problems to train classifiers. According to survey [19], mostly used machine learning algorithm is SVM in the literature. Also Random Forests, Decision Trees, and Naive Bayes other commonly performed approaches.

SVM trained on word n-gram features get higher performance than the token-list baseline applying a Levenshtein distance-based heuristic [20]. For detecting surface-level BOW features, authors performed Brown clustering against data sparsity problem on a dataset annotated for antisemitism [15] and used SVMs. In other studies [21–23], word n-grams have been used by combining with other features such as typed dependency relations, lexicon based approaches, sentimental analysis for training SVM.

Additionally Mehdad and Tetreault [24] proved character n-grams to be more predictive than word n-grams and combination of token, and word grams enhances the classification performance [11].

In several studies, meta-information of users have been utilized for getting user profile of abusive language. Age of users, time of publication, geo-position, the gender of users of Twitter users have been combined character n-grams and lexicon-based features to enhance the performance of detection [16,25,26]. Also, social network-based features such as the number of followers and friends, the number of status updates and favorites have been performed by Unsvag and Gamback [27] and improved the performance along with Logistic Regression.

With the development of deep learning-based approaches, the architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), word and paragraph embeddings have been successfully performed in Natural Language Processing problems. It is stated that the best results were achieved with deep learning models in hate-speech detection problems in a survey [19].

LSTM models with word embeddings give high performances with training a gradient boosted decision tree (GBDT) classifier on the average of these tuned embeddings in the Twitter dataset. By using this two-step approach, Badjatiya [28] proved that randomly initialized word embeddings outperform GLOVE. Character and word-level CNN have been performed in several works. Park and Fung [29] stated that the combination of these models achieves higher performance than a character n-gram Logistic Regression. Taking advantage of CNN and RNN by sequentially combining has provided to capture long range dependencies on social media data [30, 31].

Lee et al. [32] utilized word-level Bidirectional Gated Recurrent Unit (bi-GRU) through latent topic clustering with getting topic information from hidden states of GRU. Pavlopoulos et al. [33] extend their study by using an attention mechanism on tokens alongside a word-level GRU. Also, transfer learning via machine translation has

been performed in deep learning models [34, 35] to address the challenge of automatically detecting aggression on social media posts, and researchers obtained the highest scores with their approach.

3. THEORY

3.1. Hate Discourse on Media

According to Van Dijk, discourse is a complex representation of communication, based on linguistic structure, meaning and performance. In any gatherings, from a meeting to a court scene, speakers and listeners have a bunch of different features, both individually and socially, and this affects the act of communication which is also applicable to the discourse in the news. Understanding news reports is achieved through social and cognitive processes, perception of journalists and interpretations of readers [36]. In this section, we examine hate speech in the media through the methods of structural and critical discourse analysis.

3.1.1. Critical Discourse Analysis

Discourse is a term used to express production of social negotiations and change of meanings. As Foucault states, it is a knowledge area, addressing societal practices and powers of an authority. Discourse is more than imagining meanings or producing them. It also creates body politics, individuals‘ emotions and conscious [37].

Critical Discourse Analysis is an interdisciplinary field, based on the methodologies of linguistics while explaining linguistic codes with political ideologies. According to Fairclough and Wodak [38], critical discourse analysis refers to social problems, discourse is diachronic, relationship between texts and society is oblique, and discourse is a kind of social practice. They state that discourse analysis is interpreting and explanatory.

Van Dijk asserts that discourse aspect of ideologies explains how those ideologies affect the content of our daily speech. It also shapes our understanding of ideological discourse, and reproduction of social ideologies. Therefore, discourse is a pivotal ele-

ment of reproduction of ideologies and the content of daily speech. Any component of discourse such as metaphors, syntactic structure, lexicon can contribute to the reproduction of ideologies. Conflict in discourse is a result of historical, political and social backgrounds, and that the analyses of context, conflicts, power relationships and groups are required. It is also necessary to define positive and negative assumptions against "Us and Them", to be clear about presumptions and implied knowledge, and to scrutinize word choice and syntactic structures referring to conflicting groups [39].

3.1.2. Discourse Analysis on Media

News is one of the most important sources giving information about everyday life beyond our experience [40]. So, it would be logical to claim that news is a fundamental information source on groups outcasted from 'majority' groups, regarded as the core identity of a society. The visibility of minority groups in a society has decreased due to historical oppressions. Therefore, assumptions against minorities are the result of implicit statements and limitations in the news as validated and genuine information. [41].

According to Van Dijk, the production of the news is neither a direct nor a passive process, but actually socially and ideologically constructed through controlled and structured strategies. News is written under the effect of social representations referring to the dominant culture, ethnicity, gender, national and political ideology, and definitive goals [42]. Discourse analysis on media requires the examination of morphology, syntax, and semantics, rhetoric, semiotics and pragmatic structures, and strategies. Such an analysis explains the relationship between textual structures and speech, and cognitive, social, cultural or historical backgrounds of these structures [43].

Hate speech as an aggressive, insulting and targeting way of discrimination, can be seen on media within different contexts. In Turkey, several important studies have been conducted to fight against hate speech in media via analysis and disclosure. In the studies on nationalist and sexist rhetoric articulation style of the media and its ne-

gotiation forms at national and local level in Turkey, ethnic/religious/gender-based discrimination and the existence of a construction of an "us and them" format is detected. Studies such as "Child-Focused Journalism" [44–46], "Human Rights Reporting" and "Women-Focused Journalism" [47–49] published by the Independent Communication Network focus on the violations of the rights of unprivileged sections of society.Media Watch on Hate Speech project [6] aims to explain in which ways hate speech in media takes place based on the findings of these studies. It is vital to use media as a tool to fight against discrimination and to stop it, instead of reproducing, promoting and increasing it.

3.2. Deep Learning Architectures

3.2.1. Gated Recurrent Units

Gated Recurrent Units have been presented By Cho et al. [50]. For capturing the vanishing gradient and exploding gradient problems, they proposed a simpler version of Long Short-Term Memory (LSTM) [51]. The GRU combines two gates, i.e reset and update that summarizes the past information. The update gate is defined as;

$$\mathbf{z}_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

Where x_t is input vector and h_t is state vector. W_t and U_z are weight matrices and σ is an activation function that is mostly element-wise logistic sigmoid function. The reset gate is defined as;

$$\mathbf{r}_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

Where W_z and U_z are weight matrices. The activation of state h_t is defined as;

$$\mathbf{h}_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t$$

This is linear interpolation between h_{t-1} and h_t. The candidate state ht is defined as;

$$\widetilde{h_t} = tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h)$$

Where tanh is the hyperbolic tangent. The GRU forgets the past information when $h_t = 0$ that is inital value of ht.

3.2.2. Hierarchical Attention Network

Hierarchical Attention Network is a deep learning model for text classification which has been introduced by Yang et al [7]. This model utilizes the knowledge of hierarchical structure of texts and contextual meaning of words and sentences. By using attention mechanisms, it aims to provide answers to the problems of the meaning of the word that changes with the content, considering that words and sentences are not equally important in giving the main idea. The architecture of the model consists of word encoder, word attention, sentence encoder and sentence attention layers.

3.2.2.1. Word encoder. Words of delivered sentence have been embedded through an embedding matrix, W_e.

$$w_{it}, t \in [0, T]$$
$$W_e, x_{ij} = W_e w_{ij}$$

Relevant context of each sentence which is called annotations of words have been extracted by using Bidirectional GRU [52]. Annotations of words are summarized information from forward direction, sentence read s_i from W_{i1} to W_{iT} and backward direction sentence read s_i from W_{iT} to W_{i1}. By this way, network figures the context

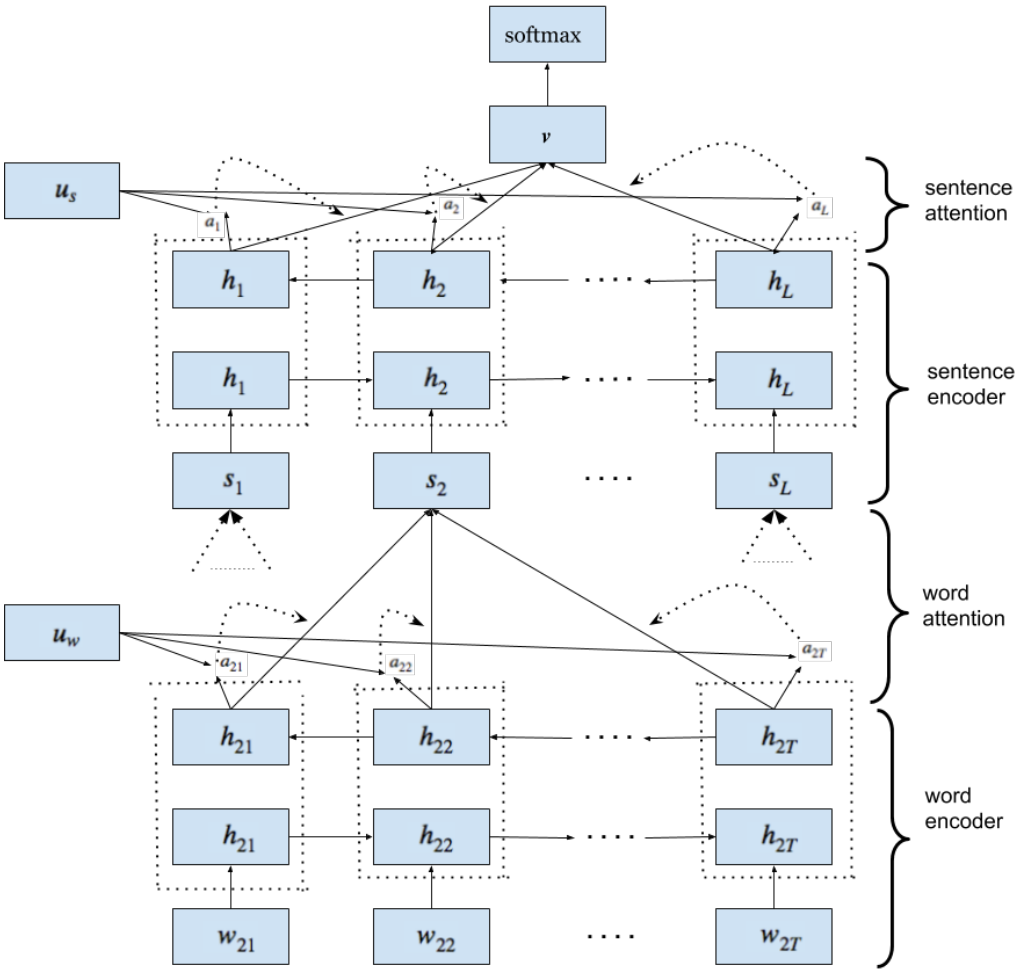

Figure 3.1. Hierarchical Attention Network

out in both directions while handling words.

$$x_{it} = W_e w_{it}, t \in [1, T]$$
$$\overrightarrow{h}_{it} = \overrightarrow{GRU}(x_{it}), t \in [1, T]$$
$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [T, 1]$$

Summarized information h_{it} obtained by merging forward hidden state h_{it} and backward hidden state h_{it} for a sentence with centered word w_{it}.

3.2.2.2. Word attention. Word attention mechanism depends on the fact that words are not equally relevant for sentence content. To emphasize connotation words for representing sentence meaning, this layer gets output of encoder layer and produces a sentence vector with indicative words. Firstly, a Single Layer Perceptron has been used to get hidden representation u_{it} by providing annotations h_{it} from previous layer. The affinity of u_{it} with a trainable context vector u_w has been calculated by dot product and then normalized to an importance weight a_{it} over a softmax function. The concatenation of the sum of normalized importance weights and word annotations h_i gives sentence vector s_i.

$$u_{it} = tanh(W_w h_{it} + b_w)$$
$$a_{it} = \frac{exp(u_{it}^\top u_w)}{\sum_t exp(u_{it}^\top u_w)}$$
$$s_i = \sum_t a_{it} h_{it}$$

3.2.2.3. Sentence encoder. Likewise the word level calculations, document vector has been obtained by feeding the sentence vector si to the network. At encoder layer, Bidirectional GRU has been utilized and concatenated in both directions to get sum-

marization of sentence contexts.

$$\overrightarrow{h}_i = \overrightarrow{GRU}(s_i), i \in [1, L]$$
$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i), t \in [L, 1]$$
$$h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$$

3.2.2.4. Sentence attention. Sentence attention mechanism gives importance to the indicative sentences as word level attention does, and gets the document vector v which summarizes the whole information of sentences.

$$\overrightarrow{h}_i = \overrightarrow{GRU}(s_i), i \in [1, L]$$
$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i), t \in [L, 1]$$
$$h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$$

4. METHODOLOGY

4.1. Dataset

4.1.1. Data analysis and annotation

In 2009, Hrant Dink Foundation started a project [6] to detect hate speech in the media. Within the scope of this project, each text of 16 newspapers chosen according to daily circulation was examined until 2011 and determinations were made based on Recommendation No. R(97) 20 of the Committee of Ministers of the Council of Europe [4]. Thanks to the obtained data within that period, the prominent words for producing hate speech were specified. With a 'keywords' list in Table 4.1. generated via these prominent words, the analysis and annotation methodology was developed. By the year of 2012, Hrant Dink Foundation used this predetermined 'keywords' list for scanning a huge amount of newspapers. The Foundation was able to monitor all national and approximately 500 local newspapers in Turkey systematically via the media monitoring company 'PRnet'. Every news that included the keywords were analyzed and annotated according to content.

Table 4.1. Pre-determined 'keywords' for media monitoring [1–3]

ETHNIC-NATIONAL IDENTITY		
Kürt, 'Kurd'	*Mülteci*, 'Refugee'	*Çingene*, 'Gypsy'
Ermeni, 'Armenian'	*Suriyeli*, 'Syrian'	*Yunan*, 'Greek'
Rum, 'Rûm'	*kaçak*, 'runaway'	*soykırım*, 'genocide'
Türk düşman, 'Turcophobe'	*göçmen*, 'immigrant'	*batılı*, 'western'
Roman, 'Romani'	*hain*, 'traitor'	
RELIGIOUS IDENTITY		
Alevi, 'Alevi'	*baş örtüsü*, 'head scarf'	*gerici*, 'backwards-looking'
Hıristiyan, 'Christian'	*türban*, 'hijab'	*sıkma baş*, 'jilbab'
Hristiyan, 'Christian'	*Musevi*, 'Juadaist'	*yobaz*, 'fanatic'
Müslüman, 'Muslim'	*Ezidi*, 'Êzîdî'	*İslamcı terör*, 'Islamist terrorism'
Yezid, 'Yezid'	*gavur*, 'infidel'	*İslami terör*, 'Islamic terrorism'
Yahudi, 'Jew'	*haçlı*, 'crusader'	*misyoner*, 'missionary'
Ateist, 'Atheist'	*kafir*, 'unbeliever'	*kripto*, 'crypto'

The analysis of annotated data uses the critical discourse analysis method applied in media research. While quantitative scaling is first applied in order to create certain indicators about media content and discourse, it is revealed in which newspapers, in which way the content of hate speech is most frequently targeted and who it targets. The categories of the analyzed data by years are given in Table 4.2. It is divided into three headings as columns, news articles and other (press archive pages, files, articles in readers' pages, book presentation / evaluation articles and similar texts). The fact that the column represents a larger proportion of these categories indicates that hate speech can be found more easily in the texts that give relative autonomy to the author [41].

4.1.2. Data collection

Hrant Dink Foundation has been provided annotated data which includes hate speech news masthead with the information 'publishing name', 'headline' and 'news

Table 4.2. Distribution of content per category [1–3]

	Columns	News Articles	Other
2016	61%	36%	3%
2017	57%	40%	3%
2018	54%	41%	5%

date'. Local and national printed media news have been captured daily based on the keywords via PRnet. News contenting hate speech from the one not including hate speech has been separated by using these annotated data masthead.

To collect data from website, The HTML content of the PRnet website is parsed. The lists of local and national news of weekdays, including tokens in the keyword list, viewed by the foundation for annotation are accessed. Those with hate speech are separated from those which do not include the information provided by the foundation. News that are determined by random sampling from news lists that do not contain hate speech are classified by information of the publishing date, publishing name, author, publishing type and content of the news. Those which do not include hate speech are also taken to the local in the same format using web scraping.

Our created dataset consists of 18318 annotated news articles with two classes such as 9311 not containing hate speech and 9007 containing hate speech. The words regarding ethnic - national or religious identity in table A. ensure that similar subjects are covered by the 18318 news articles. This situation makes our task quite challenging.

4.1.3. Preprocessing

Our dataset has been taken from PRNet's written press news archive collected with optical character recognition (OCR). This dataset scanned by OCR containing non-Turkish character strings and distorted news texts, has been quite noisy. To enhance the performance of the developed model, the preprocessing of dataset is a crucial step. During the preprocessing phase, all tokens were lower-cased. Non-Turkish char-

acters and numbers, URL links and stopwords were removed. Finally, the dataset is divided into 60% train, 20% validation, and 20% test splits for model development for the sake of consistency and comparability.

4.2. Traditional Machine Learning Models

The most common algorithms building a machine learning (ML) approach in the hate speech detection problems are Support Vector Machine (SVM) [11, 53, 54], Logistic regression [28, 53] and Naive Bayes [55, 56]. According to previous studies, Logistic Regression, SVMs, and Naive Bayes using term frequency-inverse document frequency (TF-IDF) weighted n-grams achieve high performance on the classification of hate speech in texts [56–58]. N-grams have been frequently preferred features by researchers thanks to being highly predictive. Also, it is stated that the combination of char and token n-grams with other features enhances the performance, in a survey [59].

To compare the performance between the deep learning-based models and machine learning approaches, we have performed SVM with Linear Kernel, Logistic regression and Naive Bayes with TF-IDF weighted character and word n-grams. Weighted representation taking into account sequences of words and chars has been obtained by using token n-grams range between 1-2 and character 2-grams along with TF-IDF. The hyperparameters of SVM, Logistic regression and Naive Bayes models are tuned by using grid search along with the validation split mentioned in section Dataset. According to tuned settings, SVM linear kernel with $C = 0.1$ and Logistic regression with $C = 5$ and $penalty = l2$ have been performed.

In Figure 4.1. TF-IDF values for 30 tokens with the highest score per class are represented. Tokens in 'Hate Speech' class have been mostly related to targeted groups that are reported by the foundation in Hate Speech and Discriminatory Discourse in Media reports [1, 2] and covered in the keyword list. According to weighted term list, *kendi* 'self', *bizim* 'our' pronouns fall into the 'Hate Speech' class, while several reported speech tags are in the contrary class. This may indicate the importance of

our proposed approaches as mentioned in section 3.4.

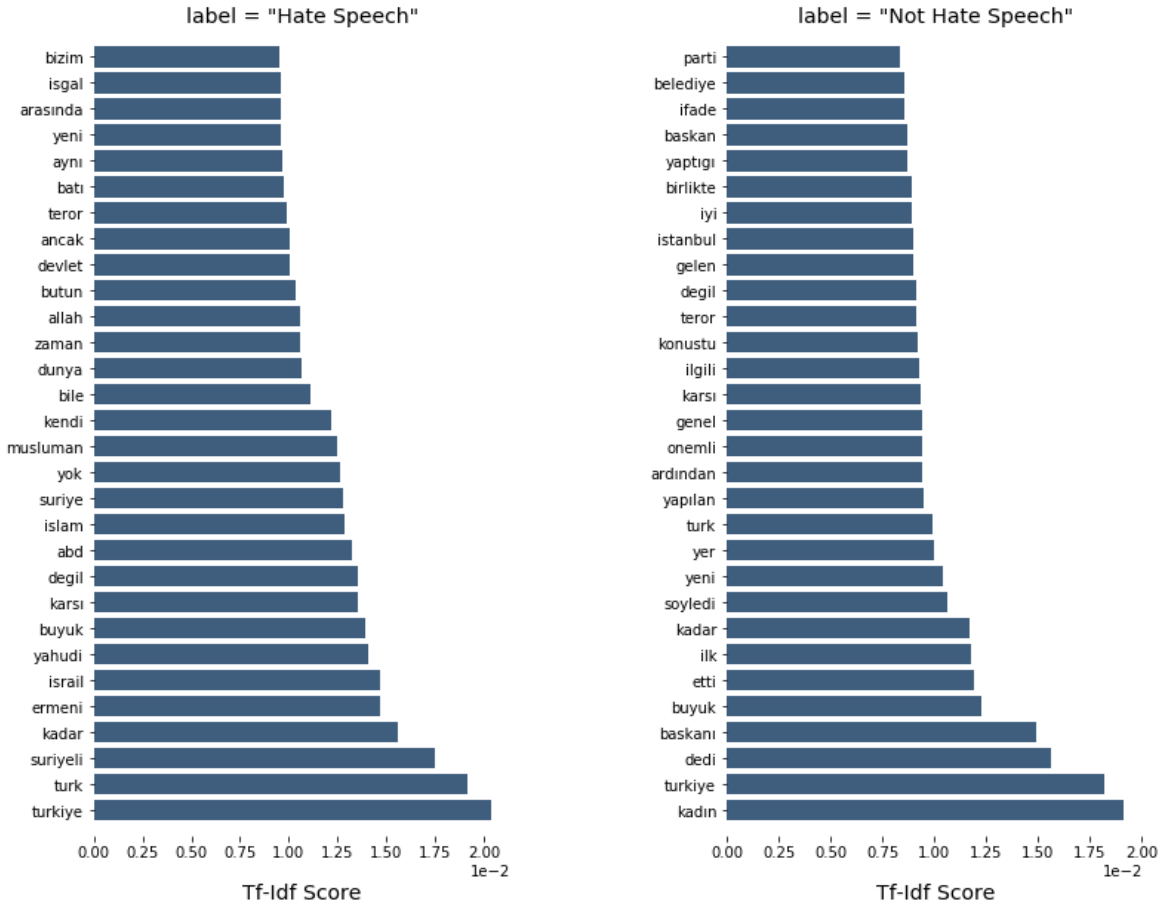

Figure 4.1. Tokens with the highest $n = 30$ TF-IDF scores per class

4.3. Hierarchical Attention Network for hate-speech detection

In this section, we have presented hate speech detection on national and local press as a document classification problem. Unlike social media texts, hate speech reflected in media texts is implicit and representative. While the explicit hate speech language usually includes sexist or racial slur words, these are mostly not used in the implicit media language. Instead, abusive language is concealed by the use of vague terms, ridicule, profanity, and other means. According to Van Dijk, discourse that controls semantic markers, such as media, can only be examined through its context [60]

.

CNN and LSTM being a type of RNN are deep learning architectures commonly used in natural language processing problems such as hate speech detection. While CNN is an effective feature extractor neural network, LSTM is capable of learning long-term dependencies. It is seen in the literature that these architectures have been used separately [61] and together [30] on social media data and good performances have been obtained. CNN extracts word and character combinations and LSTM learns word level and character level representations in the problem of hate speech detection.

However, addressing the contexts of words and sentences and their changing meanings in different texts is a new approach for text classification problems and finds the first solution with the Hierarchical Attention Network (HAN) [7]. HAN uses knowledge of the hierarchical structure of texts and includes the attention mechanism that can find the most significant words and phrases in the text content. The model depends on the fact that every word and every sentence in the text is not equally important in giving the main idea and that the meaning of words varies depending on the content. We have implemented the HAN which can measure the interest weights of words depending on the context that we consider is compatible with the feature of our dataset in the problem of hate speech detection. For model development, the preprocessed dataset as mentioned in the section Dataset has been used as %20 test, %20 validation and %60 training.

4.3.1. Word representations

We have utilized two common word vector models, word2vec [62] and fastText [63] in the embedding layer. While fastText considers a word to be built by n-gram characters, word2vec takes every single word into account as the smallest unit. As well as performing with these two different pre-trained word embedding models with dimension 200 which are published by Kyubyong Park [64]. We have also trained domain-specific word embeddings on training and validation splits with specified vector

shape 200 and window size 5. These several embedding models have been tested to demonstrate whether the performance can be enhanced compared to the model with 200 dimensional word2vec embeddings suggested by Yang et al [7].

4.3.2. Optimization

While training a deep learning architecture, there are various parameters that directly control the behavior of the model. Despite the hyperparameter tuning, the growing parameter space and lack of memory restrict the tuning with limited hyperparameters such as the number of hidden units in GRU layers, number of hidden units in attention layers, optimizer function, learning rate, type and size of word embeddings. We have tuned hyperparameters along with controlled experiments as stated in table 4.3.

Table 4.3. Analyzing hyperparameters of HAN model for obtaining optimized settings

Hyperparameters	**Search space**	**Optimal settings**
Type of word embedding	(Word2vec, FastText)	FastText
# of hidden units in GRU layer	(50 - 200)	100
# of hidden units in attention layer	(50 - 200)	200
Optimization function	(Adam, RMSprop, SGD)	RMSprop
Learning rate	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}	10^{-3}

4.4. Linguistic Processing of Hate-Speech in Turkish Language

We have developed several linguistic features in consideration of the qualitative analysis of hate discourse in the Turkish language. During this section, the novel methods to develop task-specific features have been examined.

4.4.1. Othering language

According to Hodge and Kress [65] , use of the qualitative relational syntax results in being much more judgmental. This means that the author can make an argument and evaluation in favor of the action he/she supports. The authors who support the action believe that they have a social position in being an active supporter or judge and reflect this in their discourse by using the sentences of relational syntax intensively.

Bias is reflected in the operational syntax with the confrontation of 'we' and 'you', and while evaluations about 'we' are often presented with an affirmation, 'you' is negated and the unapproved situation of 'you' is exaggerated. A highlighted 'we' and 'you' confrontation, or discrimination in a way, is intensified by a blame and mockery against people referred by the personal pronoun 'you' and adverb of place. Thus, when we consider that while confronting 'we' and 'they', ie 'you', 'they' ie 'you' are conceptualized as 'the cause of action' in terms of both discourses; a detailed and intense 'they', ie 'you' argument legitimizes 'us', on the one hand, and on the other hand, it provides the formation of ideological discourse against the forces that cause the action [66, 67] .

To employ contextual method while detecting positive self-representation, the negative representation of the other and the superiority of 'us' to 'them', we have defined a bunch of rules by taking advantage of pos tags and typed dependency relations in the sentences. The Universal Dependencies Pipe (UDPipe) [68] has been used with the UD Turkish Treebank (IMST-UD) model [69] to get part of speech (POS) tags and typed dependency relations.

(i) In Turkish, verb is obligatorily marked with person agreement morphemes, hence, we can drop the subject pronoun. Having an overt subject pronoun, however, serves discourse functions such as foregrounding the personal information. Sentences from news including first person pronoun (*ben* 'I', *biz* 'we') with person agreement marker on the verb (-Im, -m, -Iz, -k, -lIm) especially conjoined with

in another sentence second person subject pronoun (*sen* 'you sg.' and *siz* 'you-plural') with second person person agreement marker on verb (-sIn, -n, -Ø, -sInIz, -nIz, -In, -InIz) or third person pronoun (*o* 'he/she/it', *onlar* 'they') with person agreement marker on the verb (-Ø, -lAr(<-Ø+lAr)), have been extracted to develop a feature set. For instance, *"Ben söyledim. Siz dinlemediniz."* has been parsed by using UD pipe and typed dependency relations presented as, nsubj(*ben* 'I', *söyledim* 'said') and nsubj(*siz* 'you', *dinlemediniz* 'did not listen').

(ii) As has been in the case of the previous rule for identification of opposition, sentences from the news including a sentence with an overt first person subject pronoun (*ben* 'I', *biz* 'we') (footnote: sadece fiil cumleleri, verbal predicatelar tarandi) together with another sentence with first person pronoun in the complement position of a verb marked with second or third person agreement morphemes, and vice versa, have been extracted to develop a feature set. For instance, *Ben söyledim. Beni dinlemediniz. Siz beni dinlemediniz.* 'I told. You did not listen. You did not listen to me.' has been parsed by using UD pipe and typed dependency relations presented such as nsubj(*ben* 'I', *söyledim* 'said'), obj(*beni* 'me', *dinlemediniz* 'did not listen') and nsubj(*siz* 'you', *dinlemediniz* 'did not listen').

(iii) In Turkish, the possesive suffix attached to a noun indicates the possessor of the named entity. Using genitive pronoun together with the noun underlines the possessive relation. Sentences from news including sentences with genitive pronoun and noun with possessive suffix (-Im, -ImIz, -In, -InIz, -(s)I) together have been extracted to develop a feature set. For instance, *Onların benim ülkemde hayranları bir yerlerine kına yaksınlar.* 'Their supporters in my country may walk on air.' has been parsed by using UD pipe and typed dependency relations have been presented as nmod:poss(*onların* 'their', *hayranları* 'supporters'), nmod:poss(*benim* 'my', *ülkemde* 'country').

There are sentences from hate-speech labeled news as related to 'othering language' features described above. In these sentences, it can be detected that there is a clear

stress on 'you' and 'we' and such an expression marginalizes and labels a group of people in the society as the enemy. The expression obviously results in polarization among societal group by directing people to hateful emotions with the help of associations of the words in the collective conscious and historical background.

- *Biz her daim bu millet ile savaşan güçler olduğunu bilerek yaşıyoruz. Düşmanlarımızın olduğunu, onların bu mücadeleyi asla bırakmayacağını bilerek yaşıyoruz! Siz ise ne tarihi göz önüne alıyor, ne zamane şartlarını göz önüne alıyor, ne de zerre kadar vicdan gösteriyorsunuz!*, 'We are always aware of the existence of some forces against our nation. We are always aware of our enemies, who won't give up. Yet, you don't care about the history, conditions of today or a bit conscience.'
- *Kapılarınızı bugün ve dün olduğu gibi, sizin oyunlarınızla, entrikalarınızla ülkelerinden, topraklarından olan insanlara kapattığınız gibi kapatır mıydınız? [...] Sizin acımasız; dejenere olmuş ruh dengesizliğiniz neticesinde bugün dünya neredeyse bir ateş çemberinin içine girmiş durumda. Sanılmasın ki, siz, bu ateş çemberinin dışında kalacaksınız.*, 'Would you close your doors, as you do today and did in the past, for people running away from their countries because of your dirty games and intrigues? [...] The word is at the edge of annihilation due to your cruelty and your distorted, unstable state of mind. Do not assume that you'll be out of it.'

4.4.2. Imperative endings

Oktar et al. [66] also emphasize that while the authors highlight the contrast between "we" and "you", they prefer imperatives in the clauses addressed to "you". Phrasing such imperative sentences, which contain ridicule, scorn and accusation as far as pragmatics is concerned, can be considered as a way that the author aims an implicit representation of the authority and power of "us" which includes the author as well. Because imperative sentences imply that the language user has the power to give orders on their behalf [65]. Additionally, sentences containing imperative mood mediates more personal and subjective exposition of topics.

We have used the UDPipe to obtain imperative information present in the verbs. For instance, *Gavur gavurluğunu bil edebinle otur.* 'Infidel, be aware of your infidelity. Do not chase rainbows.' has been parsed and related dependency tags are presented as, 'otur' is imperative verb_root. As stated in the sentence above, the word infidel is associated with a negative and insulting expression against non-muslim individuals under collective conscious, and it functions as a political tool targeting 'Western' and European countries. Using imperative expressions aims to emphasize power, authority and consequently the superiority of 'us' on 'you'.

4.4.3. Reported speech forms

According to media discourse analysis reports, subjective media language tends to include hate discourse [41]. To detect objectivity/subjectivity we have considered the existence of reporting verbs in news. A list with 25 tokens has been created to detect news covering reported speech forms. Mostly, these tokens reflect objectivity in the news language such as *açıklamak* 'explain', *dile getirmek* 'state', and *aktarmak* 'report', while others include the interpretation of the journalist such as *suçlamak* 'accuse', *şikayet etmek* 'complain', and *uyarmak* 'warn'. Our lexicon includes objective tokens as well as subjective ones for evaluating discrimination in meaning. Also in the list of 30 tokens with the highest TF-IDF scores as showed at Figure 4.1, there are reported speech tags belonging to the Not Hate Speech class. It is proved how important these tokens are in our dataset.

In Figure 4.2, the proportion of documents per label can be seen according to the existence of reported speech forms. The tokens *işaret et(mek)* 'to point', *karşılık ver(mek)* 'to respond', *suçla(mak)* 'to blame', *uyar(mak)* 'to warn', *şikayet et(mek)* 'to complain' have higher ratios in the Hate Speech labeled news which may reflect the judgment of journalist as we assumed. There is a sentence from the dataset to emphasize changing narrative with the usage of reporting speech; *Bizi hep "İsrail'i ilk siz tanıdınız" diye suçladılar.*, 'We have always been accused of being the first country legitimizing Israel.'

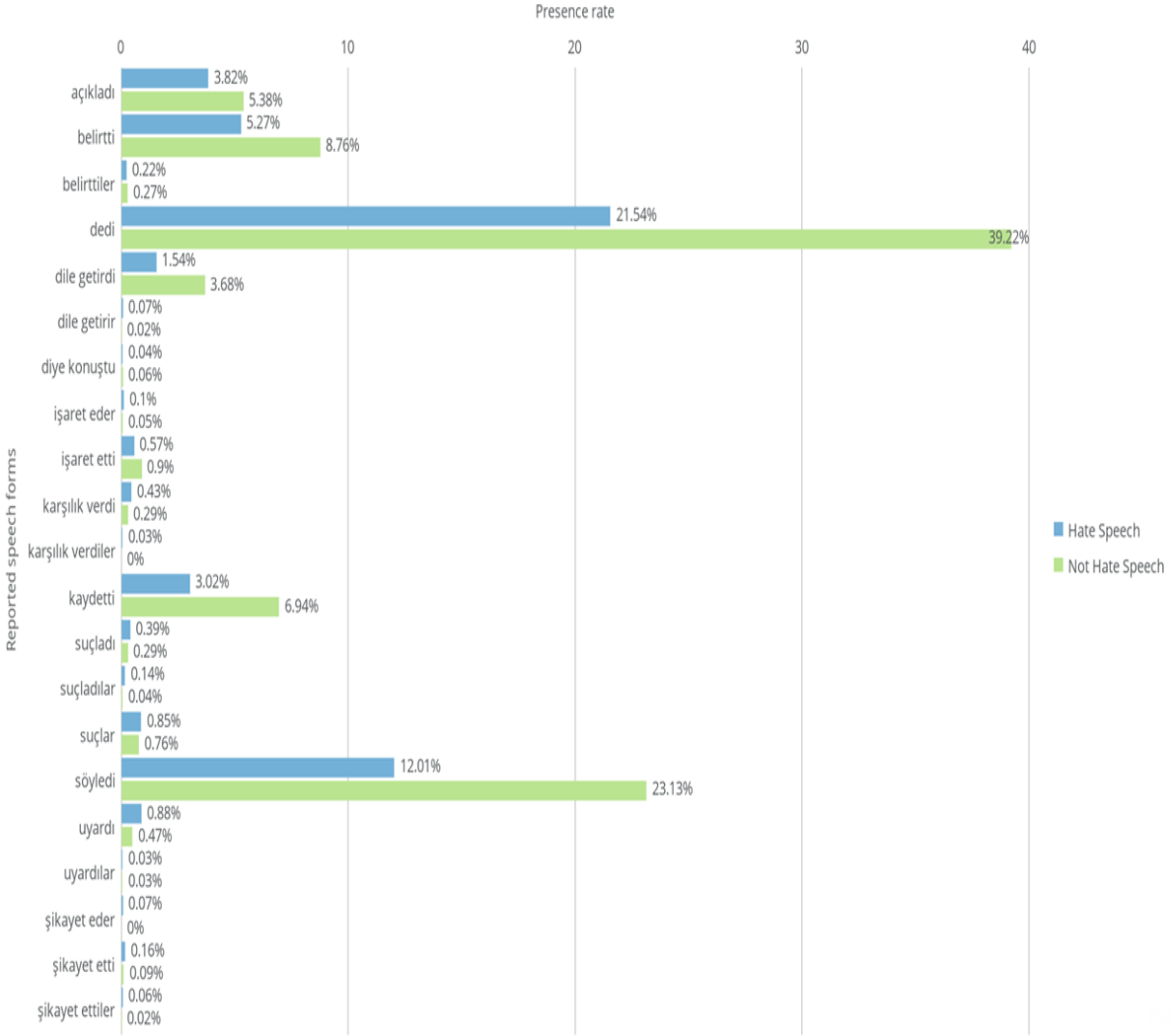

Figure 4.2. Presence rate of Reported Speech Forms among our term list per class

4.4.4. Numerical information

We have extracted certain numerical features which are imperative clause, indirect speech and othering language ratio of texts. The ratio has been calculated for each news by dividing the count of sentences containing extracted features to the number of sentences in a news. Feature vectors have been formed with these computed ratios. In Figure 4.3, the existence of imperative endings, othering language and reported speech ratio per label have been shown. The biggest difference between ratios per

label is 19.48% belonging to reported speech. As we expected, the existence of ratio is higher for Not Hate Speech labeled documents. The second-highest difference is in the existence of othering language ratios with 15.2% and lastly, the difference in the presence of imperative endings is 12.84%. These two ratios are higher in the Not Hate Speech labeled documents, which supports our proposed feature extraction methods.

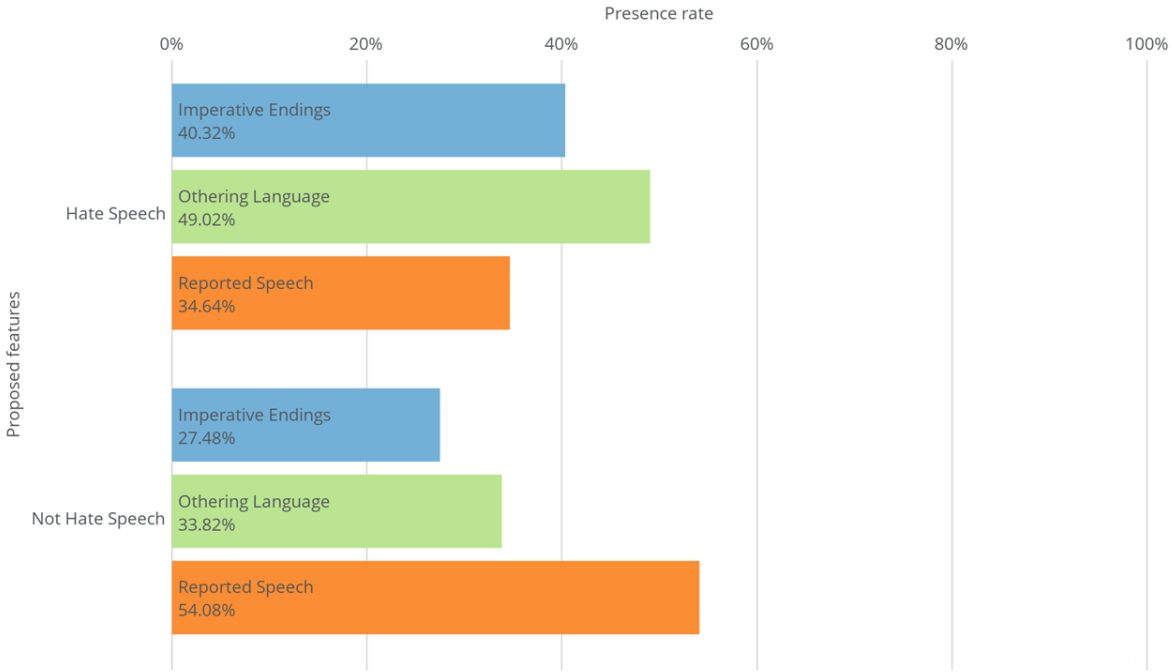

Figure 4.3. Proposed othering language, imperative endings and reported speech features existence rate per class

The numerical feature vectors have been examined as to whether they should be rescaled to adjust observed values mean to 0 and the standard deviation to 1, to decrease training time and enhancing performance. They have been subjected to the Shapiro-Wilk test [70] and the D'Agostino's K-squared test [71] to check if they are Gaussian distributions. Both tests coherently showed that they were not Gaussian distributions and so, standardization couldn't be performed on this dataset. We can also see the numerical feature vector distributions per label in Figure 4.4, Figure 4.5, Figure 4.6.

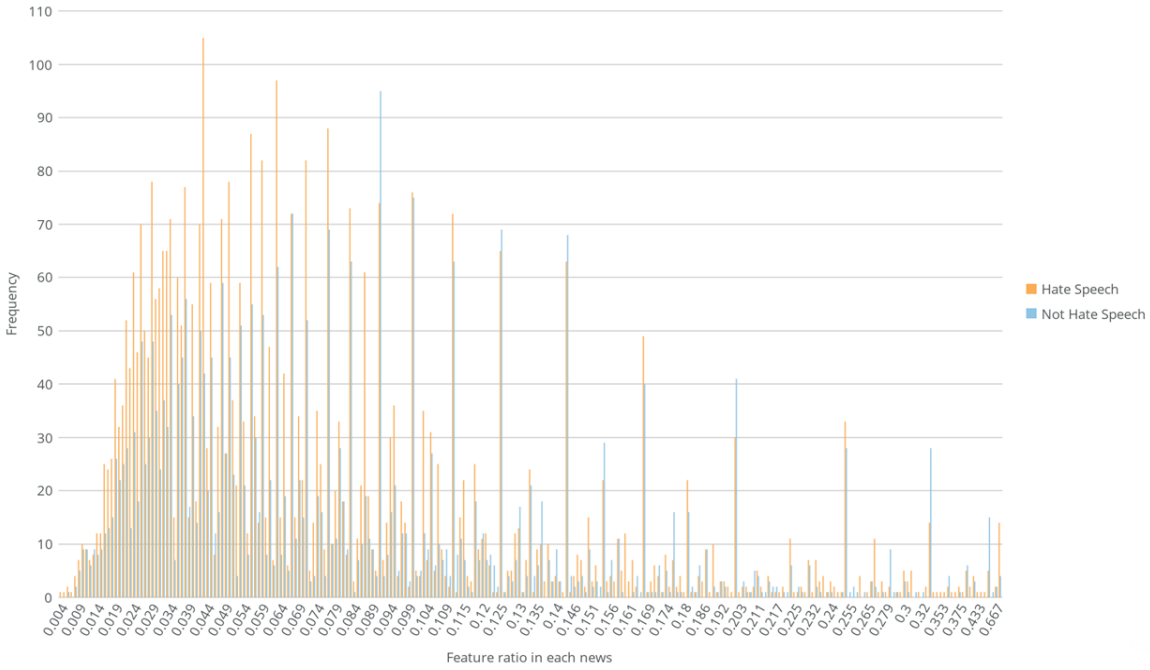

Figure 4.4. Distribution of othering language feature ratio in each news

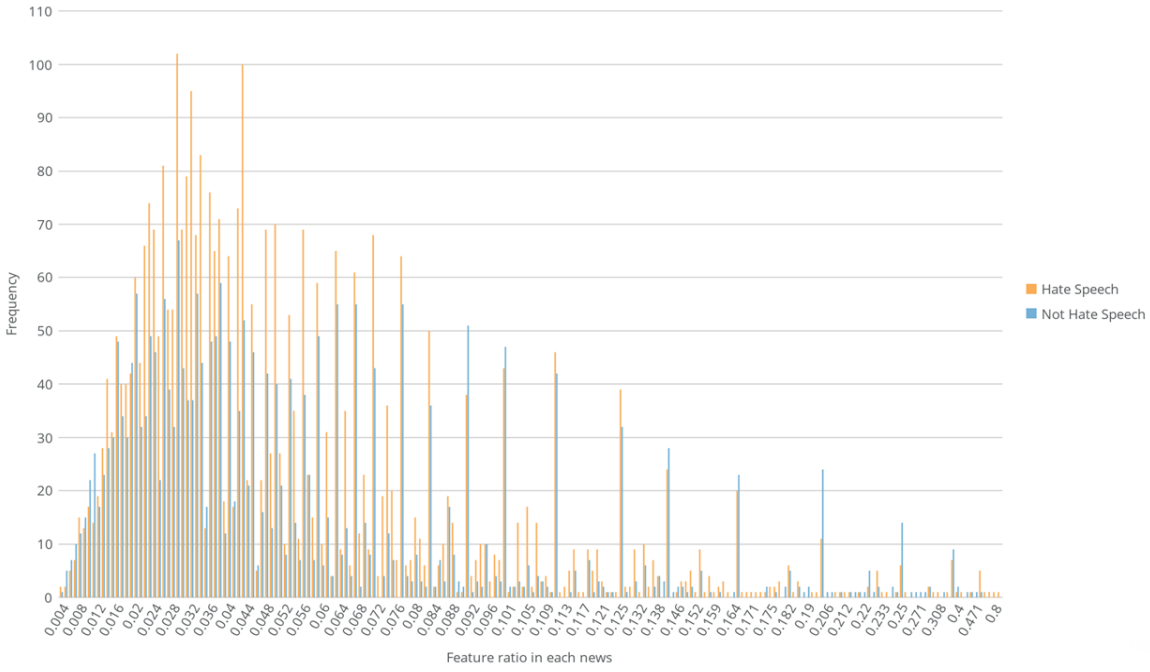

Figure 4.5. Distribution of imperative endings feature ratio in each news

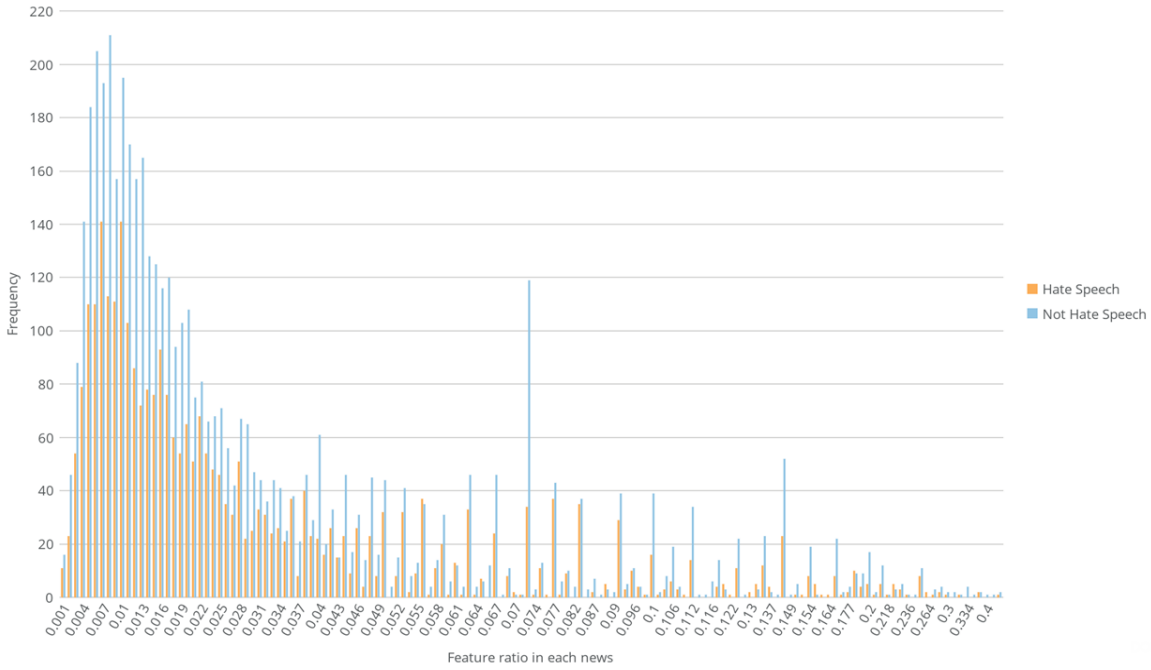

Figure 4.6. Distribution of reported speech feature ratio in each news

These linguistic patterns have been used to constitute linguistic feature sets for our task. We have developed two different feature sets. $Feature\ set_1$ captures the othering language and imperative endings. New document sequences have been created by extracting sentences of news including implemented patterns. These revised documents and unchanged documents not including patterns constitute $feature\ set_1$. Our second feature set, $feature\ set_2$ holds the information of the existence of reported speech based on our lexicon by one-hot encoded. And we combined the numerical features such as ratios of othering language, imperative ending and reported speech patterns to form $feature\ set_2$.

4.5. Hierarchical Attention Network with linguistic features

During this section, we examine document embedding merged with linguistic feature and concatenation of distributional representation of features and metadata features to HAN model as a novel architecture.

4.5.1. Document embedding

Considering that previous studies obtained effective results [11, 18] in the detection of hate speech by using document embeddings, which enabled to capture the semantics of texts, we have performed document embedding for our task. It has been proved that the embedding representation of documents with similar semantics of context belongs to a related part of space [72]. Therefore, we have expected to obtain similar vector representations of documents in the same class to get a good semantic representation of the news.

Each news has been analyzed to extract the feature of othering language and imperative endings. Hence, $feature\,set_1$ has been formed as mentioned in the Othering Language section. We have processed $feature\ set_1$ through paragraph embedding to represent low dimensional vectors of the documents [72] . In the document embedding model, as well as words, sentences are also mapped to unique vectors . The Distributed Memory Model of Paragraph Vectors (PV-DM) takes the context words as a random sample composed of consecutive words and the paragraph id from a paragraph, and predicts a center word. According to Le and Mikolov [72], the PV-DM takes into account the semantics of words and word orders which outperform high dimensional, generalized, poor representation of documents along with bagofngrams. Also they have stated that PV-DM shows consistently higher performance than the Distributed Bag of Words version of Paragraph Vector (PV-DBOW) which predicts words sampled randomly from the document.

We have utilized Doc2Vec which outperforms competitor document embedding models such as skip-thought [73] and paragram-phrase [74], to train document embeddings [75]. Hyperparameters were tuned through controlled experiments in consideration of settings showed at Table 4.4. Due to growing parameter space, it has been limited with vector size, window size, min count and the number of epochs. Finally, optimal hyperparameter settings for $feature\,set_1$ have been decided as vector size = 200, window size = 5, minimum frequency threshold = 1, and the number of training epochs

Table 4.4. Analyzing hyperparameters of document embedding for obtaining optimized settings

Hyperparameters	**Search space**	**Optimal settings**
Vector size	100 - 200 - 300	200
Window size	2 - 5 - 10	5
Min count	1 - 5	1
# of epoch	20 - 50	20

= 20. These learned lowdimensional vector representations have been concatenated to Hierarchical Attention Network.

Additionally, we investigated whether document embeddings have captured the othering language context, by computing distance between the semantically similar vectors. Qualitative analysis of Documents learned by utilizing 2D principal component analysis is represented in Figure 4.7 and 4.8. The relative distances between word vectors have been calculated with cosine similarity. Tagged dots in enlarged figures show 250 closest words to *bizim* 'our' and *onların* 'their'. In figure 4.7, the words "*bizim* 'our', *bizlerin* 'we all', *kendileri* 'ourselves', *kendimize* 'for us', *gitsinler* 'they may go', *düşmanlıkları* 'their hostility', *şerrinden* 'of their malice', *bölücüler* 'separatists', *cahil* 'ignorant', *kafirler* 'infidels', *batılılar* 'western people', *zalimlerin* 'tyrants', *avrupalıların* 'European people', *hain* 'traitors', *gaflet* 'hamartia' " have relatively smallest distance from the word onların 'their'. In figure 4.8, the relatively closest words to the word *bizim* 'our' are "*onların* 'their', *sizin* 'yours', *sizler* 'you all', *başkalarının* 'others' ', *düşmanımız* 'our enemies', *dostumuz* 'our allies', *silahlarımızı* 'our forces', *topraklarımız* 'our territories', *namusumuz* 'our honor', *inancımız* 'our creed' ". Their co-occurrences may indicate the usage of discriminatory language, that's to say othering language with the oppositon of 'us' and 'them'; and superiority of 'us' to 'them', which supports the approach we utilized.

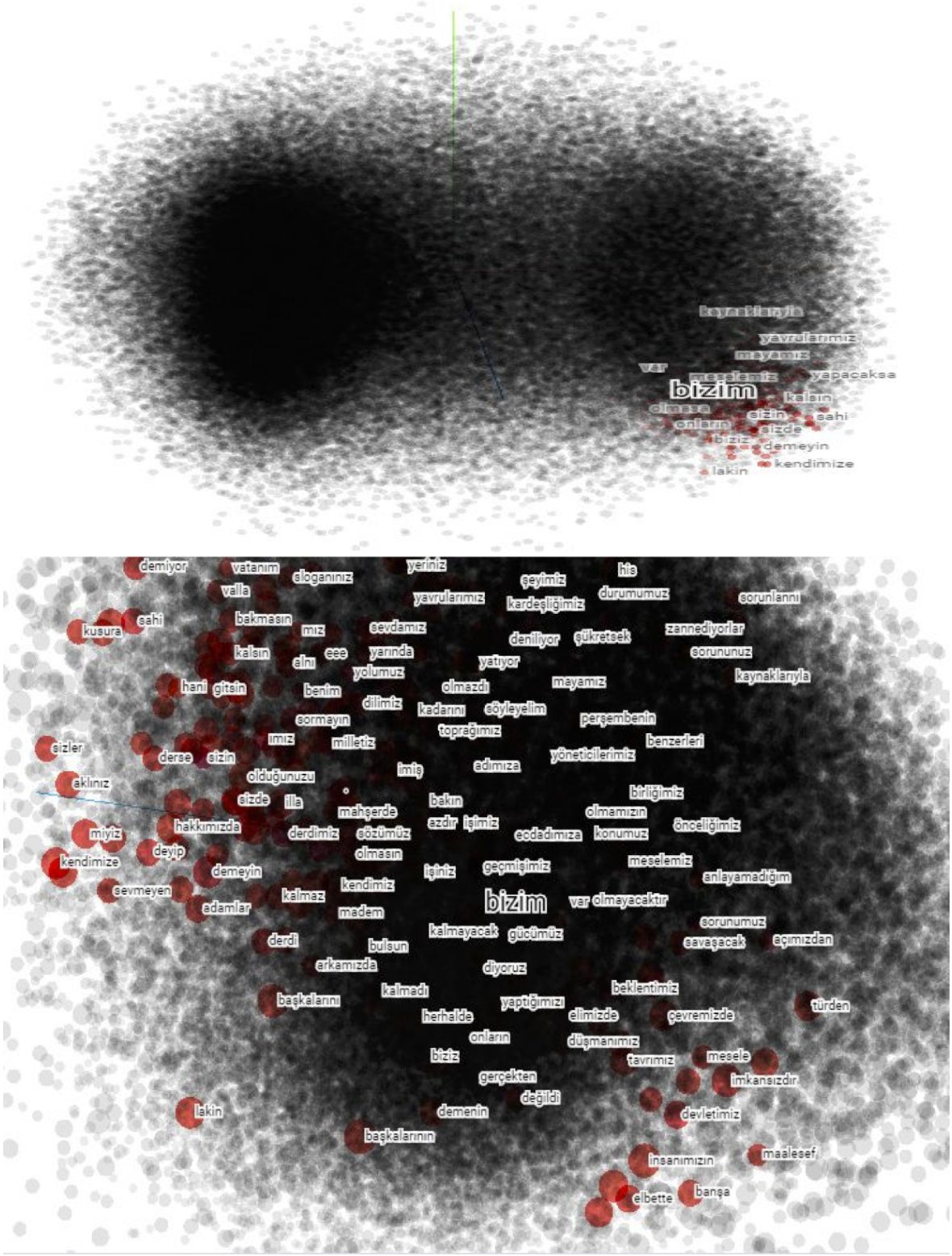

Figure 4.7. Othering language feature vector 'bizim / our' with $n = 250$ closest vectors through cosine distance

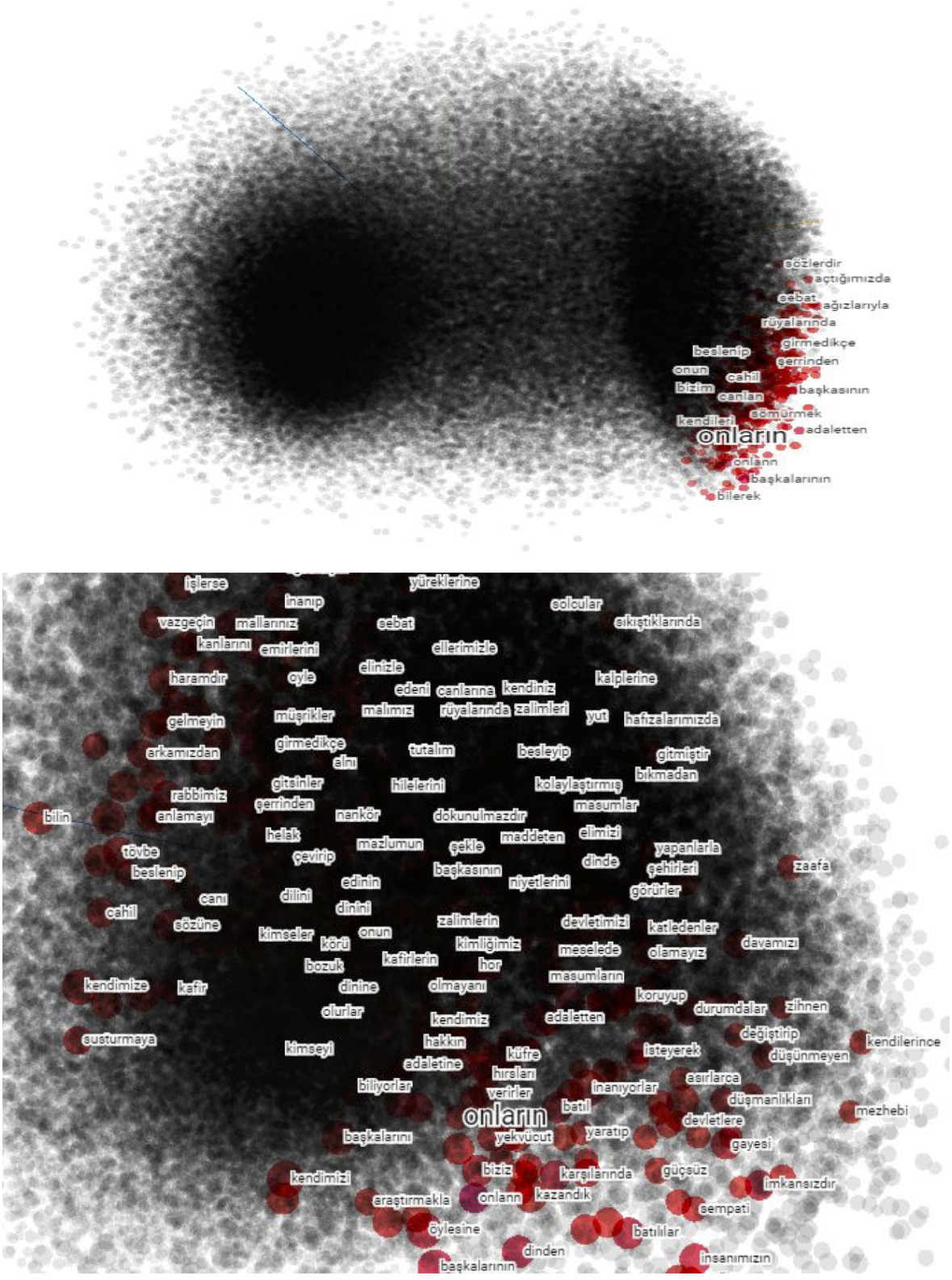

Figure 4.8. Othering language feature vector 'onların / their' with $n = 250$ closest vectors through cosine distance

4.5.2. Concatenation to HAN

In addition to capturing the semantic content of documents with Hierarchical Attention Network, we have also considered the patterns of hate discourse on the news to enhance our model. For this purpose, feature sets have been formed and $feature set_1$ has been pretrained along with paragraph embedding. $Feature\ set_1$ and $feature\ set_2$ concatenated to HAN separately and together, and their performance in detecting hate speech has been examined.

In the first case, we have concatenated $feature\ set_1$ pretrained by doc2vec. Our proposed architecture has been presented in Figure 4.9. Pretrained paragraph embedding is processed through two fully connected layers with 200 hidden units and the Rectified Linear Unit (ReLU) activation function before concatenating with the output of the attention layer of HAN. Concatenated vectors have been fed into a fully connected layer with 200 hidden units and the ReLU activation function. We have implemented dropout regularization with a rate of 0.2 to the hidden layer. Lastly, the sigmoid activation function has been performed to generate predictions.

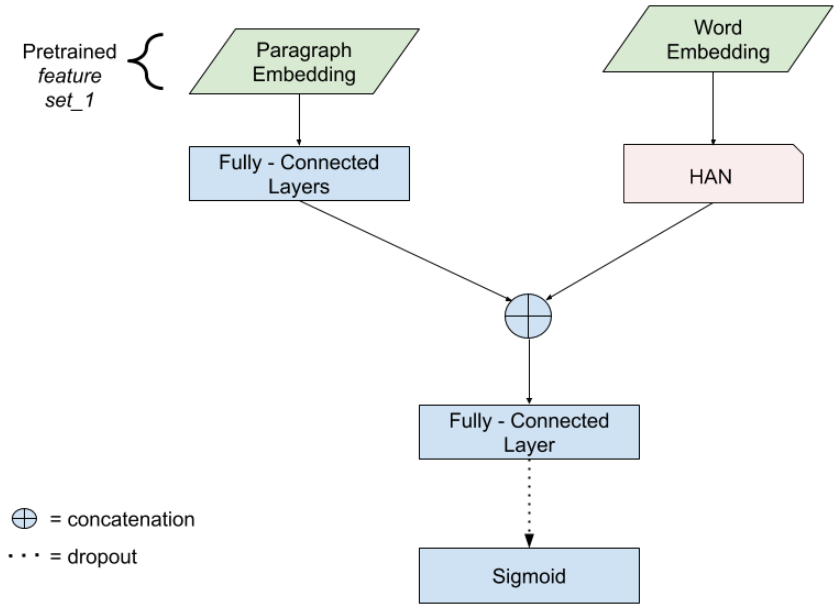

Figure 4.9. Architecture of HAN with $feature\ set_1$

In the second case, we have concatenated $feature\ set_2$. Our proposed architecture has been presented in Figure 4.10. External features concatenated with the output of the attention layer of HAN. Concatenated vectors are processed along with a fully connected layer with 200 hidden units and the ReLU activation function. We have implemented dropout regularization with a rate of 0.2 to the hidden layer. Finally, to create predictions, the sigmoid activation function has been performed.

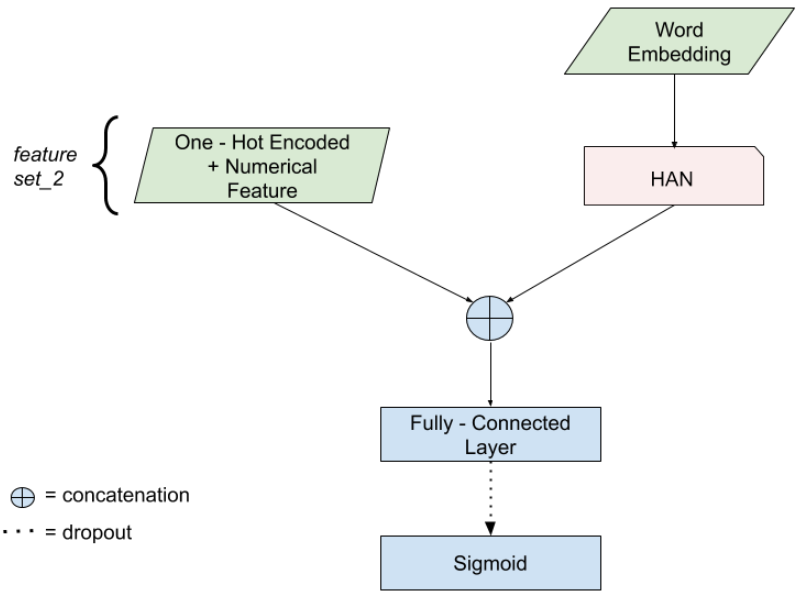

Figure 4.10. Architecture of HAN with $feature\ set_2$

In the third case, we have concatenated pre-trained $feature\ set_1$ as well as $feature\ set_2$. Our proposed architecture has been presented in Figure 4.11. We have basically merged the previous two models. $Feature\ set_1$ and output of the attention layer have been concatenated and passed to a fully connected layer with 200 hidden units and the ReLU activation function. $Feature\ set_2$ are concatenated to the document vectors and processed with a fully connected layer with 200 hidden units and the ReLU activation function. Dropout regularization with a rate of 0.2 have been performed to the hidden layer. Lastly, predictions are generated by performing the sigmoid activation function.

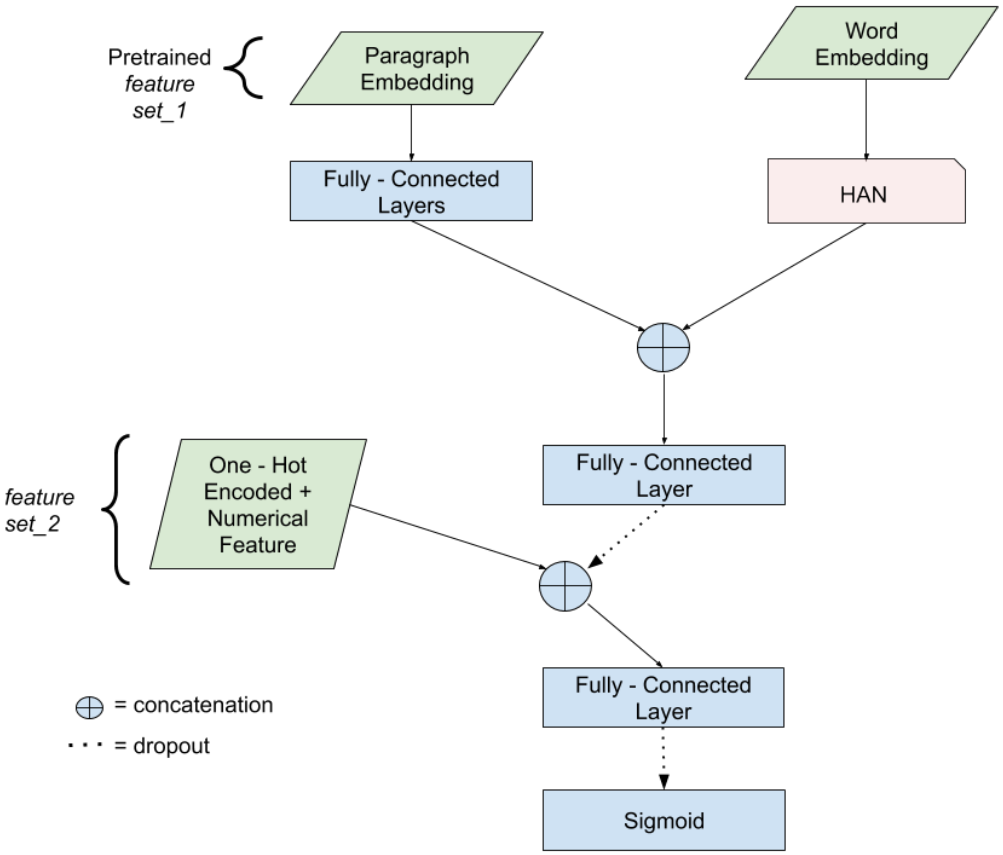

Figure 4.11. Architecture of HAN with $feature\ set_1 + feature\ set_2$

During model configurations, controlled experiments have been performed and the bestfitting model has been decided for our case by considering the number of hidden layers, number of hidden units, and dropout rate. Also, the same optimized hyperparameters of HAN model have been used.

5. EXPERIMENTS AND RESULTS

5.1. Comparative study of hate-speech detection on Media

To the best of our knowledge, Çoban et al. [58] have developed the hate speech detection system on Turkish Media. 1036 news annotated as including hate speech have been gathered from the Nefret Soylemi (hate speech) [76] web site by crawling. They have constituted non-hate speech class by collecting CNN [77] and BBC [78] news assumed to not include hate-inciting content. They have used keywords stated by Nefret Soylemi project report as related to hate content and extracted new keywords list from the retrieved BBC news to obtain the news not including hate speech. Thus, 1038 BBC and 948 CNN texts have been retrieved and labeled as not-hate speech news. To develop the system, they have proposed three classifiers, SVM with linear kernel. Naive Bayes and Multilayer Perceptron with TF-IDF weighted character and word n-grams. According to the experiment results presented by Çoban et al., MLP with word n-grams achieves the highest recall and f-score in within-corpus (BBC news as train and test) as well as highest recall score in cross-corpus (BBC news as train and CNN news as test). The highest precision score in within-corpus has been obtained by SVM with linear kernel and Naive Bayes with word n-gram. Besides overall performance drops as compared to the within-corpus, they have got the highest f-score for cross-corpus by using MLP with char n-gram [58].

We have compared the performance of the HAN and other proposed methods with the methods performed in this study [58] by using a different dataset which is identified by annotators. For an accurate comparison, we have followed similar training and feature extraction steps. Additionally, we have implemented Logistic regression and performed a grid search by using the validation set to get optimized parameters of classifiers. According to our results in the table below, overall scores with char n-gram are lower than the ones obtained with word n-grams. Logistic regression reaches the highest scores in all metrics and second-highest scores obtained by using SVM with

Table 5.1. Evaluation scores for comparative methods

Classifier with char 2-gram + tf-idf	**Accuracy**	**Precision**	**Recall**	**F-score**
SVM with linear kernel	78.87%	0.77	0.80	0.78
Logistic Regression	78.05%	0.77	0.78	0.78
Naïve Bayes	71.91%	0.73	0.67	0.70
Multilayer Perceptron	79.06%	0.78	0.79	0.79
Classifier with word (1,2)-gram + tf-idf	**Accuracy**	**Precision**	**Recall**	**F-score**
SVM with linear kernel	86.21%	0.85	0.87	0.86
Logistic Regression	86.84%	0.85	0.88	0.87
Naïve Bayes	81.85%	0.80	0.83	0.82
Multilayer Perceptron	84.82%	0.84	0.85	0.84

linear kernel.

5.2. Deep learning models

In this section, we present the performance of HAN and our proposed models. To maintain consistency, these models have been evaluated on the test set with 3644 documents, 20% of the overall dataset. The tuned hyperparameters during model configurations have been used and all models trained for 10 epochs with early stopping and batch size 32. We have compared the performance of HAN with traditional ML classifiers and CNN as well as with proposed models. Our evaluation metrics are accuracy, precision, recall, and f-score. For the sake of reliability of results, we have computed an average of evaluation scores of five experimental runs in each case with a fixed seed.

Word representations have been trained on our training dataset along with word2-vec and fastText in the first phase. Using trained word embeddings we have developed the HAN model and changing performances have been explored through different word embeddings. Table 5.2 shows that the HAN model with the word embedding trained by fastText achieves higher performance than the pre-trained embeddings and obtained

Table 5.2. Evaluation scores of HAN with several word embeddings and CNN

Deep learning models	**Accuracy**	**Precision**	**Recall**	**F-score**
CNN	86.68%	0.85	0.89	0.87
HAN with pre-trained word2vec	85.95%	0.91	0.82	0.86
HAN with pre-trained fastText	86.83%	0.88	0.85	0.86
HAN with word2vec	86.99%	0.89	0.85	0.87
HAN with fastText	89.43%	0.90	0.89	0.89

word embedding by using word2vec on our Turkish dataset. This result has proved that considering the internal structure of words, as fastText does, is quite useful for morphologically rich languages such as Turkish and domain-specific word embeddings trained on our train dataset outperform pre-trained embeddings for this task.

We have also observed that HAN with fastText outperforms the evaluation scores of traditional ML-based approaches showed in table 5.1 in all metrics with 0.89 accuracy, 0.90 precision, 0.89 recall, and 0.89 F-score. For the comparison with DL models, the word embedding vectors trained by fastText have been used in CNN and HAN with feature sets models. In the survey [19] stated that word-level CNN is one of the recent state of the art approaches in hate speech detection problems. Rusert et al. [79] stated that combination of multiple CNNs slightly worse than the CNN proposed by Yoon [80] while classifying offensive tweets. Thus we have performed CNN model that is based on the Yoon's model [80] with 3 parallel convolution layer and windows size 3, 4, 5 of words with filter size 100 of each for feature extraction. The performance of CNN architecture achieves the second-highest F-score. The result prove that the linguistic context of long sequences cannot be captured by CNN.

Lastly, we have compared the performance of HAN with proposed feature sets. Also the way feature sets change the performance of the HAN as the baseline has been examined. Thus, proposed models have been performed along with trained fastText embedding. As stated in table 5.3, HAN with $feature\ set_1$ achieves slightly higher

Table 5.3. Evaluation scores of HAN by combining with proposed feature sets

Models with feature sets	**Accuracy**	**Precision**	**Recall**	**F-score**
HAN with $feature\ set_1$	90.22%	0.90	0.90	0.91
HAN with $feature\ set_2$	90.01%	0.91	0.88	0.89
HAN with $feature\ set_1$ + $feature\ set_2$	90.59%	0.91	0.90	0.91

performance than the HAN in all metrics and HAN with $feature\ set_2$ in accuracy, precision and Fscore. HAN with $feature\ set_2$ obtains better evaluation scores in accuracy and precision metrics than the HAN model. The combination of $feature\ set_1$ and $feature\ set_2$ within HAN outperforms other architectures.

These results prove that $feature\ set_1$ is more predictive than $feature\ set_2$ with combining HAN. $Feature\ set_2$ that includes feature vectors related to ratios and one hot encoded of reported speech tags, cannot capture semantic contexts from the news sufficiently. There could be several possible reasons. $Feature\ set_2$ might hold too sparse information regarding our problem to be useful and might not be able to become predictive significantly along with sequential data processed by HAN model. On the other hand, $feature\ set_1$ assigning othering language features to similar vectors spaces has been slightly effective within abstract linguistic structures captured by HAN. While $feature\ set_1$ increases the performance of the HAN model, combined with $feature\ set_2$, it reaches the highest evaluation score in all metrics. Othering language features might leverage the predictive effect of reported speech forms and numerical features.

5.3. Qualitative analysis

By considering the natural segmentation of documents, HAN model captures changing attention weights of words and sentences. Along with attention weights in sentence level and word level we have visualized texts as in figure 5.1 and in figure 5.2. The highlighted words/sentences are the most significant ones while shades of colors represent degree of significance. The highlighted sentence in darker orange indicates higher importance, similarly darker green represents more significant words. In figure

5.1 and in figure 5.2, there are sample news articles annotated as hate speech. It can be seen that same words have different importance in particular sentences. We can also obtain words which are mostly associated with the detection of hate speech in news articles.

Figure 5.1. Hate speech labeled sample news 1 on media data

Figure 5.2. Hate speech labeled sample news 2 on media data

6. CONCLUSION AND FUTURE WORK

In this thesis, we have created a dataset consisting of 18318 national and local news labeled as hate speech or not hate speech by the Hrant Dink Foundation [6]. Detection of hate speech in printed media is a challenging problem as most of the hate speech in the news is based on context and implications, and and also as the dataset is formed with similar topics by keywords filtered in the table. Thus, a system that can detect changing discursive cues and understand the context around these discourse is required. To the best of our knowledge, this is the first model developed in the Turkish language as hate speech detection system using an annotated dataset. With our method, we aim to reduce dependency on human effort for the detection of hate speech which is crucial for the elimination of discrimination.

For the detection of hate speech is reflected implicitly and in a representative manner on media, we have performed the HAN model by analyzing different word representations. In consideration of the qualitative analysis of hate discourse in the Turkish language, several linguistic features have been developed. The HAN has been leveraged with these novel features and performances of the new models have been examined. Our results suggested that the HAN model is able to address changing interest weights of words depending on the context by considering the natural segmentation of documents and work more effectively than comparable machine learning and CNN models in the hate speech detection problems. Also, combining with pre-trained othering language based features enhances performance.

As future work, with qualitative analysis of the dataset, we are planning to investigate other linguistic approaches to hate speech detection in Turkish Language by collaborating with linguists. We will investigate two-step approaches to tune word embeddings and then train a GBDT classifier on the average of these tuned embeddings in news articles. We will also investigate user profiling by creating user embeddings and combining them with our deep learning architecture. Lastly, contextual models

such as BERT [81] and ELMO [82] will be integrated to our architecture that may be appropriate for our problem.

REFERENCES

1. Foundation, H. D., *Hate Speech and Discriminatory Discourse in Media 2017 Report*, HDV Publication, 2017.

2. Foundation, H. D., *Hate Speech and Discriminatory Discourse in Media 2018 Report*, HDV Publication, 2018.

3. Foundation, H. D., *Hate Speech and Discriminatory Discourse in Media 2016 Report*, HDV Publication, 2016.

4. N., R. R., "of the Committee of Ministers to Member States on 'Hate Speech"', *Adopted On*, Vol. 30, 1997.

5. Jourová, V., "Code of Conduct on countering illegal hate speech online: First results on implementation", *European Commission.[cit. 8. březen 2018]*, 2016.

6. Foundation, H. D., *Medyada Nefret Söyleminin İzlenmesi Projesi*, 2009-2019, `https://hrantdink.org/tr/asulis/faaliyetler/projeler/medyada-nefret-soylemi/`.

7. Yang, Z., D. Yang, C. Dyer, X. He, A. Smola and E. Hovy, "Hierarchical Attention Networks for Document Classification", *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, Association for Computational Linguistics, San Diego, California, Jun. 2016, `https://www.aclweb.org/anthology/N16-1174`.

8. Xu, J. M., K. Jun, X. Zhu and A. Bellymore, "Learning from bullying traces in social media.", *Association for Computational Linguistics.*, pp. 656–666, 2012.

9. Gitari, N. D., Z. Zuping, H. Damien and J. Long, "A lexicon-based approach

for hate speech detection", *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10, No. 4, pp. 215–230, 2015.

10. Burnap, P. and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics", *EPJ Data Science*, Vol. 5, No. 1, p. 11, 2016.

11. Nobata, C., J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content", *Proceedings of the 25th international conference on world wide web*, pp. 145–153, International World Wide Web Conferences Steering Committee, 2016.

12. Williams, M. L. and P. Burnap, "Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data", *British Journal of Criminology*, Vol. 56, No. 2, pp. 211–238, 2015.

13. Xiang, G., B. Fan, L. Wang, J. Hong and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus", *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1980–1984, ACM, 2012.

14. Razavi, A. H., D. Inkpen, S. Uritsky and S. Matwin, "Offensive language detection using multi-level classification", *Canadian Conference on Artificial Intelligence*, pp. 16–27, Springer, 2010.

15. Warner, W. and J. Hirschberg, "Detecting hate speech on the world wide web", *Proceedings of the second workshop on language in social media*, pp. 19–26, Association for Computational Linguistics, 2012.

16. Waseem, Z. and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter", *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

17. Zhong, H., H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller and C. Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network.", *IJCAI*, pp. 3952–3958, 2016.

18. Alorainy, W., P. Burnap, H. Liu and M. Williams, "The Enemy Among Us: Detecting Hate Speech with Threats Based'Othering'Language Embeddings", *arXiv preprint arXiv:1801.07495*, 2018.

19. Mishra, P., H. Yannakoudakis and E. Shutova, "Tackling Online Abuse: A Survey of Automated Abuse Detection Methods", *arXiv preprint arXiv:1908.06024*, 2019.

20. Sood, S. O., J. Antin and E. Churchill, "Using crowdsourcing to improve profanity detection", *2012 AAAI Spring Symposium Series*, 2012.

21. Burnap, P. and M. L. Williams, "Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making", *In Internet, Policy Politics, Oxford, UK*, 2014.

22. Dinakar, K., R. Reichart and H. Lieberman, "Modeling the detection of textual cyberbullying", *fifth international AAAI conference on weblogs and social media*, 2011.

23. Van Hee, C., E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans and V. Hoste, "Detection and fine-grained classification of cyberbullying events", *Proceedings of the international conference recent advances in natural language processing*, pp. 672–680, 2015.

24. Mehdad, Y. and J. Tetreault, "Do characters abuse more than words?", *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 299–303, 2016.

25. Dadvar, M., D. Trieschnigg, R. Ordelman and F. de Jong, "Improving cyberbullying detection with user context", *European Conference on Information Retrieval*,

pp. 693–696, Springer, 2013.

26. Galán-García, P., J. G. d. l. Puerta, C. L. Gómez, I. Santos and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying", *Logic Journal of the IGPL*, Vol. 24, No. 1, pp. 42–53, 2016.

27. Unsvåg, E. F. and B. Gambäck, "The effects of user features on twitter hate speech detection", *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pp. 75–85, 2018.

28. Badjatiya, P., S. Gupta, M. Gupta and V. Varma, "Deep learning for hate speech detection in tweets", *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, International World Wide Web Conferences Steering Committee, 2017.

29. Park, J. H. and P. Fung, "One-step and two-step classification for abusive language detection on twitter", *arXiv preprint arXiv:1706.01206*, 2017.

30. Zhang, Z., D. Robinson and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network", *European Semantic Web Conference*, pp. 745–760, Springer, 2018.

31. Wang, C., "Interpreting neural network hate speech classifiers", *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 86–92, 2018.

32. Lee, Y., S. Yoon and K. Jung, "Comparative studies of detecting abusive language on twitter", *arXiv preprint arXiv:1808.10245*, 2018.

33. Pavlopoulos, J., P. Malakasiotis and I. Androutsopoulos, "Deep learning for user comment moderation", *arXiv preprint arXiv:1705.09993*, 2017.

34. Samghabadi, N. S., D. Mave, S. Kar and T. Solorio, "RiTUAL-UH at TRAC 2018

shared task: Aggression identification", *arXiv preprint arXiv:1807.11712*, 2018.

35. Risch, J., E. Krebs, A. Löser, A. Riese and R. Krestel, "Fine-grained classification of offensive language", *14th Conference on Natural Language Processing KONVENS 2018*, p. 38, 2018.

36. Van Dijk, T. A., *News as Discourse*, Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, NJ, USA, 1988.

37. Weedon, C., *Feminist practice and poststructuralist theory (2nd ed.)*, Blackwell, Oxford, 1997.

38. Van Dijk, T. A., *Critical Discourse Analysis, The Handbook of Discourse Analysis, Second Edition. Edited by Deborah Tannen, Heidi E. Hamilton, and Deborah Schiffrin*, John Wiley Sons, Inc., 2015.

39. Van Dijk, T. A., *Söylem ve İdeoloji çok Alanlı bir Yaklaşım Çev: N. Ateş, Söylem ve İdeoloji, Haz: Barıs Çoban, Zeynep Özarslan*, Su Yayınları, İstanbul, 2003.

40. Stephens, M., *A History of News: From the Drum to the Satellite*, Viking Penguin Inc., New York, 1988.

41. Çınar, M., "Habercilik ve Nefret Söylemi", *Medya ve Nefret Söylemi: Kavramlar, Mecralar, Tartışmalar. Ed. Mahmut ÇINAR.*, Hrant Dink Vakfı Yayınları, İstanbul, 2013.

42. Van Dijk, T. A., *News as Discourse*, Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, NJ, USA, 1988.

43. Van Dijk, T. A., "Ideological discourse analysis, u: New Courant (English Dept, University of Helsinki), No: 4,(ur.) Eija Ventola and Anna Solin", *Special issue Interdisciplinary approaches to Discourse Analysis*, 1995.

44. Alankuş-Kural, S., "Türkiye'de medya, hegemonya ve ötekinin temsili", *Toplum ve Bilim*, Vol. 67, pp. 76–110, 1995.

45. Erdoğan Tosun, G., "Çocuklar ve çocuk haklarının medyada temsili", *İçinde S. Alankuş (Hazırlayan), Çocuk Odaklı Habercilik (ss. 172-195). İstanbul: IPS İletişim Vakfı Yayınları*, 2007.

46. Duran, R., "Çocuk Hakları Odaklı Habercilik ve Röportaj", *içinde Sevda Alankuş (Der.), Hak Haberciliği Dizisi*, Vol. 3, 2007.

47. Kerestecioğlu, F., "Hukuk-Basın İlişkisi ve Kadınlara İlişkin Yasal Değişiklikler", *içinde Sevda Alankuş (Der.), Hak Haberciliği Dizisi*, Vol. 2, 2007.

48. Gülbahar, H., "Kadına Yönelik Şiddet Genelgesi ve Medyanın Sorumluluğu", *içinde Sevda Alankuş (Der.), Hak Haberciliği Dizisi*, Vol. 2, 2007.

49. Köker, E. and Ü. Doğanay, "Televizyonda protesto görüntüleri: Egemen haber söylemlerinde toplumsal eylemler", *Kültür ve İletişim*, Vol. 7, No. 2, pp. 43–72, 2004.

50. Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *arXiv preprint arXiv:1406.1078*, 2014.

51. Hochreiter, S. and J. Schmidhuber, "Long short-term memory", *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

52. Bahdanau, D., K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv:1409.0473*, 2014.

53. Davidson, T., D. Warmsley, M. Macy and I. Weber, "Automated hate speech detection and the problem of offensive language", *Eleventh international aaai conference on web and social media*, 2017.

54. Del Vigna, F., A. Cimino, F. Dell'Orletta, M. Petrocchi and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook", *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017.

55. Agarwal, S. and A. Sureka, "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website", *arXiv preprint arXiv:1701.04931*, 2017.

56. Liu, S. and T. Forss, "Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification.", *KDIR*, pp. 530–537, 2014.

57. Salminen, J., H. Almerekhi, M. Milenković, S.-g. Jung, J. An, H. Kwak and B. J. Jansen, "Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media", *Twelfth International AAAI Conference on Web and Social Media*, 2018.

58. Coban, E. B. and E. Filatova, "Incendiary News Detection", *Association for the Advancement of Artificial Intelligence*, 2019.

59. Schmidt, A. and M. Wiegand, "A survey on hate speech detection using natural language processing", *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10, 2017.

60. Van Dijk, T. A., *Discourse, knowledge, power and politics*, pp. 27–64, 2011.

61. Castelle, M., "The linguistic ideologies of deep abusive language classification", *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 160–170, 2018.

62. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, pp. 3111–3119, 2013.

63. Bojanowski, P., E. Grave, A. Joulin and T. Mikolov, "Enriching word vectors with subword information", *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

64. Park, K., "Pre-trained word vectors of 30+ languages", *URL https://github.com/Kyubyong/wordvectors*, 2018.

65. Gage, W., G. Kress and R. Hodge, "Language as Ideology", *The Modern Language Journal*, Vol. 64, p. 512, 1997.

66. Oktar, L. and A. C. Değer, "Gazete Söyleminde Kiplik ve İşlevleri", *Dilbilim Araştırmaları Dergisi*, pp. 45–53, 1999.

67. Oktar, L., "The Ideological Organization of Representational Processes in the Presentation of us and them", *Discourse Society - DISCOURSE SOCIETY*, Vol. 12, pp. 313–346, 2001.

68. Straka, M. and J. Straková, "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe", *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, 2017.

69. Sulubacak, U., M. Gokirmak, F. Tyers, Ç. Çöltekin, J. Nivre and G. Eryiğit, "Universal dependencies for Turkish", *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3444–3454, 2016.

70. SHAPIRO, S. S. and M. B. WILK, "An analysis of variance test for normality (complete samples)", *Biometrika*, Vol. 52, No. 3-4, pp. 591–611, 1965.

71. D'AGOSTINO, R. B., "Transformation to normality of the null distribution of g1", *Biometrika*, Vol. 57, No. 3, pp. 679–681, 1970.

72. Le, Q. and T. Mikolov, "Distributed representations of sentences and documents",

International conference on machine learning, pp. 1188–1196, 2014.

73. Kiros, R., Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba and S. Fidler, "Skip-thought vectors", *Advances in neural information processing systems*, pp. 3294–3302, 2015.

74. Wieting, J., M. Bansal, K. Gimpel and K. Livescu, "Towards universal paraphrastic sentence embeddings", *arXiv preprint arXiv:1511.08198*, 2015.

75. Lau, J. H. and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation", *arXiv preprint arXiv:1607.05368*, 2016.

76. *Nefret Söylemi*, `http://www.nefretsoylemi.org/`.

77. CNN, `https://www.cnnturk.com/`.

78. BBC, `https://www.bbc.com/turkce`.

79. Rusert, J. and P. Srinivasan, "NLP@ UIOWA at SemEval-2019 Task 6: Classifying the Crass using Multi-windowed CNNs", *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 704–711, 2019.

80. Kim, Y., "Convolutional neural networks for sentence classification", *arXiv preprint arXiv:1408.5882*, 2014.

81. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

82. Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep contextualized word representations", *arXiv preprint arXiv:1802.05365*, 2018.

APPENDIX A: REPORTED SPEECH FORMS LIST

Table A.1. The list with 25 tokens in Turkish/English to detect news covering reported speech forms

dedi	s/he said
söyledi	s/he told
açıkladı	s/he explained
açıklar	s/he/it explains
belirtti	s/he stated that...
belirttiler	they stated that...
diye konuştu	s/he stated as
kaydetti	s/he noted
dile getirdi	s/he mentioned
dile getirir	s/he mentions
uyardı	s/he warned
uyardılar	they warned
uyarır	s/he/it warns
işaret etti	s/he/it pointed out
işaret eder	s/he/it points out
suçladı	s/he blamed
suçlar	s/he/it blames
suçladılar	they blamed
tepkilere yol açtı	it caused reactions
tepkilere yol açtılar	they caused reactions
şikayet etti	s/he reported
şikayet eder	s/he reports
şikayet ettiler	they reported
karşılık verdi	s/he responded
karşılık verdiler	they responded