Proceedings of LATERAISSE Workshop @LREC2022, pages 32–41
Marseille, 25 June 2022

Identifying Hate Speech using Neural Networks and Discourse Analysis Techniques

Zehra Melce Hüsünbeyi, Didar Akar, Arzucan Özgür
Department of Computer Engineering, Department of Linguistics, Department of Computer Engineering
Bogaziçi University
melce.husunbeyi@gmail.com, akar@boun.edu.tr, arzucan.ozgur@boun.edu.tr

Abstract

Discriminatory language, in particular hate speech, is a global problem posing a grave threat to democracy and human rights. Yet, it is not always easy to identify, as it is rarely explicit. In order to detect hate speech, we developed Hierarchical Attention Network (HAN) based and Bidirectional Encoder Representations from Transformer (BERT) based deep learning models to capture the changing discursive cues and understand the context around the discourse. In addition, we designed linguistic features using critical discourse analysis techniques and integrated them to the these neural network models. We studied the compatibility of our model with the hate speech detection problem by comparing it with traditional machine learning models, as well as a Convolution Neural Network (CNN) based model, a Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) based model which reached significant performance results for hate speech detection. Our results on a manually annotated corpus of print media in Turkish show that the proposed approach is effective for hate speech detection. We believe that the feature sets created for the Turkish language will encourage new studies in the quantitative analysis of hate speech.

Keywords: deep learning, hierarchical attention network, bert, linguistic features

1. Introduction

Hate speech is defined by the European Council as "any statement including racist hate, ethnocentrism [...] religion intolerance against minorities, immigrants or originally-immigrant groups [...] and any expressions spreading, provoking or legitimating hate." [1]Hate speech has grown exponentially and become more visible around the world as various social media platforms and conventional media become more accessible to people. Turkey is no exception in this regard. Given the potential harm hate speech can cause in terms of human rights, social justice and democracy, it is not surprising that both national and international institutions and large-scale businesses are interested in monitoring this phenomenon. The protocol signed by Council of Europe and Facebook, Microsoft, Twitter, and YouTube in 2016 to detect illegal hate speech can be given as an example of this monitoring attempt (Jourová, 2016). The protocol has been later extended to cover more platforms such as Instagram, Google+, Snapchat, and Dailymotion in 2018.

The first step in the fight against hate speech is obviously to detect it. Manual detection of hate speech which has been the common practice in many institutions requires an enormous amount of time, effort and work force, and therefore, is not sustainable. Instead, automating the identification process would be highly advantageous. However, detection and defining discourse is not an easy task due to the dynamic and contextual nature of language. The same sentence or text can mean different things when used by different speakers belonging to different social groups or when uttered in different contexts. Even if we can define the context, irony and implicit or implied meanings can still create serious problems for detecting hate speech. Therefore, it is essential to find ways of examining various clues about discourse and its context.

This study is partially based on a master's thesis by (Hüsünbeyi, 2020). In the thesis, a model has been developed for the automatic detection of hate speech through the HAN (Yang et al., 2016), which aims to detect changes in the meaning by using the hierarchical structure of texts. Then task specific linguistic features were used to enhance this neural network model and the results showed that these novel linguistic methods were effective in distinguishing news texts with hate speech from the ones without it. These linguistics features include certain forms of othering language such as possessive pronouns and lexical choices indicating the subjectivity level of the news texts. To the best of our knowledge, this is the first study that utilizes manually annotated data for hate speech in the print media of Turkey and we believe it will promote new studies with the potential of gathering different agents and disciplines. Later on, in order to further improve the results we got for the thesis, we have also considered Transformer-based BERT(Devlin et al., 2019) model, which offers the latest state-of-the-art solutions to numerous NLP problems. We investigated whether the BERT model, which processes long sequences limited by input length constraint and does not use the knowledge of the hierarchical structure of documents, unlike the HAN model, would enhance the performance of

[1]Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech"

our task. We took into consideration BERTurk[2], pre-trained language model for Turkish, and examined how it would yield results with the proposed architecture and novel linguistic features.

2. Related Work

In the detection of hate speech, domain specific and traditional linguistic features have a significant role. There are some commonly used features in the literature (Xu et al., 2012; Gitari et al., 2015; Burnap and Williams, 2016a) such as part of speech tags (POS), typed dependency relations. As one of the most promising linguistic approaches, the othering language concept was utilized as a framework to determine hate speech for contents on social media (Burnap and Williams, 2016b; Alorainy et al., 2019). By using the Stanford Lexical Parser (De Marneffe et al., 2006). (Burnap and Williams, 2016b) presented syntactic grammatical relationships in a tweet to obtain opposition. For example, the typed dependency relation nsubj(home, them) in the "send them back home" sentence identifies the relational sense between the tokens and underlines the divergence between 'us' and 'them'. They also stated that statistically significantly better results were achieved with the othering feature set, especially for detecting hate speech related to religious beliefs. According to (Alorainy et al., 2019), othering language theory, based on the combination of linguistics approaches such as set of in group (us) / out group (they) separation in hate speech samples that include 'two-sided' pronoun (us vs them).

Besides linguistics related features, surface features e.g., n-grams, bag-of-words (BOW), local features e.g., TF-IDF weights of tokens, and rule-based approaches e.g., errors in spelling, and the count of punctuation marks were used with traditional machine learning algorithms. According to the recent survey in (Mishra et al., 2019), the most commonly used model in the detection of hate speech systems is Support Vector Machines (SVM), and other commonly utilized learning algorithms are Random Forests, Decision Trees, and Naive Bayes.

In recent times, deep learning-based approaches such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), word and paragraph embeddings have been successfully used in Natural Language Processing (NLP) problems. (Badjatiya et al., 2017) developed CNN and LSTM neural network models with random embeddings or GloVe embeddings using a dataset which contains 16K annotated tweets labeled as 'sexist', 'racist', and 'neither sexist nor racist' (Waseem and Hovy, 2016). These task specific learned embedding weights have also been used as features along with SVM and Gradient Boosting Decision Tree (GBDT) classifiers. The best evaluation score was obtained by tuning random embeddings with LSTM and then by using these weights to train a GBDT classifier. Character and word-level CNN have also been used in several works. The combination of these models achieves higher performance than a character n-gram Logistic Regression model as reported by (Park and Fung, 2017). In order to obtain long range dependencies on social media data, (Zhang et al., 2018; Wang, 2018) considered combining CNN and RNN sequentially. (Zhang et al., 2018; Wang, 2018). More recently, transformer based approaches such as the BERT model (Devlin et al., 2019) and its variants have gained importance with their power of learning large language models. The effectiveness of these models have also been demonstrated in the recent hate speech detection shared tasks at Semeval-2019 (Liu et al., 2019), Semeval-2020 (Wiedemann et al., 2020), and HASOC-2020 (Mishraa et al., 2020).

While deep learning models often do not contain manually designed linguistic features, it has become important to use linguistics to get a better idea of how the model works and to avoid generalization errors. We integrate and examine the contribution of hate speech signaling linguistic structures to the HAN model. We focus on hate speech detection in the Turkish language, which is a morphologically rich and agglutinative language. At the Semeval-2020 task, a Turkish Track was organized and a Twitter dataset which is an extended version of the dataset studied by (Çöltekin, 2020), was provided to the participants. The first two ranked teams utilized multi-lingual pre-trained Transformers based on XLM-RoBERTa (Wang et al., 2020) and ensemble of CNN-LSTM, BiLSTM-Attention, and BERT models as well as word embeddings (Ozdemir and Yeniterzi, 2020). To the best of our knowledge, there is only one prior study on the automatic detection of hate speech in Turkish news articles (Coban and Filatova, 2019), where traditional machine learning models with automatically annotated data were used. In this paper, we use manually annotated data from print media and develop a novel hybrid approach for Turkish hate speech detection by integrating linguistic features with a deep learning model.

3. Methodology

3.1. Dataset

In this study, we used a dataset of print media news articles, the manual annotations of which were obtained from Hrant Dink Foundation, who have been monitoring the media for hate speech since 2009 in the scope of the Media Watch Project (Hrant Dink Foundation, 2021). Within the scope of this project, the foundation monitors all national and approximately 500 local newspapers in Turkey methodically through the media monitoring company 'PRNet'. The news articles including a predetermined set of 'keywords' are examined along with the critical discourse analysis methods and annotations are made manually based on Recommendation No. R(97) 20 of the Committee of Ministers of the Council of Europe.

[2]https://huggingface.co/dbmdz/bert-base-turkish-cased

The dataset that we obtained from the foundation consists of 18316 annotated news articles published between 2016-2018, with two classes: 9309 news articles not containing hate speech and 9007 news articles containing hate speech. Both classes are composed of news articles that contain prominent words regarding ethnic or religious identity, which makes the task of distinguishing articles with hate speech from the ones without hate speech more challenging. This dataset, which was scanned by OCR, is quite noisy. It contains non-Turkish character strings and distorted news texts. To enhance the performance of the developed models, we lower-cased all tokens, and removed the non-Turkish characters and numbers as well as the URL links. Then, we divided the dataset into 60% train, 20% validation, and 20% test splits for model development.

3.2. Linguistic Processing of Hate-Speech

We developed several linguistic features taking into account the qualitative analysis of hate discourse in the Turkish language. The novel methods to generate task-specific features are examined in this section.

3.2.1. Othering Language

The opposition between 'we' and 'you' is typically used in biased texts, while 'we' has positive representations, and 'you' or sometimes 'they' receive negative representation. This opposition can lead to discrimination and hate speech by reinforcing blaming and mockery directed at 'you' (Oktar and Değer, 1999; Oktar, 2001).

In order to detect the opposition between the positive representation of "we" and the negative representation of "you", we made use of some discursively constrained morpho-syntactic properties of Turkish that are listed below. To this end, we got part-of- speech (POS) tags and typed dependencies in the sentences by using Universal Dependencies Pipe (UDPipe) (Straka and Straková, 2017) with the UD Turkish Treebank (IMST-UD) model (Sulubacak et al., 2016).

1. Turkish is a pro-drop language; in other words subject pronouns can be dropped because verbs are inflected with obligatory person agreement morphemes. When a subject pronoun is not dropped, it serves discourse functions such as contrastive focus and foregrounding person information. Based on this feature we extracted sentences with overt subject pronouns in first person conjoined with sentences with overt subject pronouns in second person.

2. Another case of opposition can be established when the subject of the first sentence is used as a complement in the following sentence.

3. The genitive construction in Turkish also follows the pro-drop principle. Since the noun is obligatorily inflected with the possessive agreement marker, the pronoun marking the possessor can be dropped. When the genitive pronoun is overtly present, it is also used for contrastive focus or foregrounding purposes. Based on this feature, we extracted sentences containing genitive pronouns followed by nouns marked with possessive person agreement (-Im, -ImIz, -In, -InIz, -(s)I).

In the training set, we found that hate-speech labeled news indeed include sentences with the 'othering language' features described above.

- The following extract from the dataset it can be seen that the use of overt subject pronouns sets up oppositions between biz 'we' and siz 'you'.
'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!. Biz devletimize güveniyoruz! Her ne olursa olsun devletimizin yanındayız, yanında olacağız! Biz bu toprakları vatan yapmak için yüzyıllardır can veririz, can alırız! [...]'[3]
'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. We trust in our state! No matter what happens, we stand by our state, and we will continue to do so! We have been dying and killing for centuries in order to make these lands our homeland!'

- The following extract illustrates the use of overt genitive pronouns to set up oppositions between 'you' and 'us' followed by an overt subject pronoun in second person with the same effect.
[...] Yani kendi ülkemizdeki sizin uşağınız Haçlı zihniyeti ile mücadele ettik. Bu bizim utancımız değil. Ama sizin büyük bir utancınız var. Almanlar, yani sizler Hitler gibi korkunç bir katili yarattınız. Ülkenizin sokaklarında hala gamalı haçlı Nazi artıkları dolaşıyor. İnsanları sabun fabrikalarında yakan bir Nazi despotu sizin eserinizdir. Genetiğinizde soykırımcılık var. Siz onların torunlarısınız. [...], '[...] In other words, we struggled with your servant Crusader mentality in our own country. This is not our shame. But you have a great shame. You, the Germans, created a terrible killer like Hitler. Nazi scraps with swastikas still roam the streets of your country. A Nazi despot who burns people in soap factories is your achievement. You have genocidalism in your genetics. You are their descendants.[...]'

[3]Right after each Turkish text shown in Italic, we provide its English translation

3.2.2. Use of Imperatives

In media texts, imperative structures are occasionally used and like the aforementioned structures, they, too, represent the opposition between "we" and "you" (Oktar and Değer, 1999). Imperative structures in these oppositional contexts typically display the authority and power of "we" over "you", because imperative sentences imply that the language user has the power to give orders (Kress and Hodge, 1997).

We have utilized UDPipe to obtain imperative morphemes on the verbs. For example, "*Gavur gavurluğunu bil edebinle otur.*" 'Infidel, know your infidelity and know your place.' has been parsed 'otur' has been identified as imperative verb_root. Here the word infidel is associated with non-Muslims and it is a derogatory term. It functions as a political tool targeting 'Western' and European countries. In this sentence the addressee (i.e. the infidel) is ordered to know their place and behave accordingly. Imperative expressions as in this example emphasize power, authority and consequently the superiority of 'us' on 'you'.

3.2.3. Reported Speech Forms

In general, subjective media language tends to include hate discourse (Çınar, 2013). To detect objectivity/subjectivity we have considered reported speech, in particular reporting verbs. A list of 30 reporting verbs has been created to detect texts covering reported speech. Some of 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' and iddia etmek 'to claim'. The changing narrative with the usage of reported speech form can be observed in a sample sentence from the dataset; *'Gavur gazeteleri kin kusmaya devam ediyor. Türkiye düşmanlarının hevesleri kursaklarında kalınca hazımsızlıkları gazetelerine de yansıdı. Gavur İngiltere'nin Independent gazetesi Orta Doğu muhabiri Cockburn, işgal girişimi sonrası hainlerin açığa alınmasının Türkiye'yi zayıflattığını iddia etti.'*, 'Infidel newspapers continue to throw up hatred. When the enemies of Turkey couldn't get what they wanted, their indigestion reflected on their newspapers. Infidel Cockburn, the Middle East correspondent of Britain's Independent newspaper, claimed that the suspension of traitors after the invasion attempt weakened Turkey.'

3.2.4. Encoding of Linguistic Features

The linguistic patterns described in Section 3.2 have been used to constitute novel linguistic feature sets for our task. We developed two separate feature sets. *ling_set1* captures the othering language and use of imperatives rules. If a news article includes these linguistic patterns, the portion of the document consisting of the sentences containing these patterns is extracted and used as *ling_set1*. Otherwise, if the news article doesn't includes any of these linguistic patterns, *ling_set1* consists of the entire document itself.

Our second feature set, *ling_set2* holds the information of the existence of reported speech expressions, which were encoded using the one-hot encoding scheme. In addition, three numerical features are calculated for each document, namely the ratio of sentences containing othering language, the ratio of sentences containing imperative language, and the ratio of sentences containing reported speech forms. These three dimensional numerical feature vectors are concatenated to the one-hot encoded vectors of reported speech expressions to form *ling_set2*.

Document embedding with *ling_set1*

It has been shown that the embedding representations of documents with similar semantics of context belong to a related part of space (Le and Mikolov, 2014). Considering that previous studies obtained effective results (Nobata et al., 2016; Alorainy et al., 2019) in the detection of hate speech by using document embeddings, which provide the semantics of texts to be captured, we have created document embedding for our problem. *Ling_set1* and documents not including patterns have been processed along with The Distributed Memory Model of Paragraph Vectors (PV-DM) (Le and Mikolov, 2014) to obtain low dimensional vectors of the documents with vector size = 300, window size = 5, and number of training epochs = 30.

3.3. Proposed Deep Learning Models

Hate speech reflected in the national and local press, unlike social media texts, is implicit and representative. While the explicit hate speech language often contains sexist or racial slur words, they are usually not applied in implicit media language. Abusive language is disguised by vague terms, ridicule, profanity, and other means, rather than using explicit language. As Van Dijk pointed out, discourse that controls semantic markers, such as media, can only be considered along with its context (Van Dijk, 2011). HAN and BERT based models have been implemented to address the contexts and changing meanings of words and sentences in different texts.

3.3.1. HAN for Hate-speech Detection

HAN (Yang et al., 2016) uses knowledge of the hierarchical structure of texts. The architecture of the model consists of word encoder, word attention, sentence encoder and sentence attention layers. Words of delivered sentence have been embedded and relevant context of each sentence which is called annotations of words have been extracted through Bidirectional GRU (Bahdanau et al., 2014). To emphasize connotation words for representing sentence meaning, word annotation layer gets output of encoder layer and produces a sentence vector with indicative words. Likewise the word level calculations, document vector has been obtained by feeding the sentence vector to the network.

We implemented the HAN model based on (Yang et al., 2016) using domain specific word embeddings,

which were trained on the training and validation splits of the proposed dataset through fastText (Bojanowski et al., 2017). The hyperparameters were tuned on the validation set as 100 hidden units in the GRU layers, 200 hidden units in the attention layers, and RMSprop optimizer with learning rate 10^{-3}.

HAN with Novel Linguistic Features

As well as obtaining the semantic content of documents with HAN, hate discourse patterns in news articles have been also taken into account to improve our model. For this purpose, *ling_set1* and *ling_set2* have been concatenated to HAN both separately and jointly, and their performance in identifying hate speech was analyzed.

Initially, the pre-trained *ling_set1* were combined HAN model. We processed paragraph vectors through two fully connected layers with 200 and 100 hidden units, respectively, and the Rectified Linear Unit (ReLU) activation function is applied. Before concatenation with document representations, the dropout regularization with a rate of 0.3 was implemented to the attention layer of HAN. The concatenated vectors were fed into a fully connected layer with 200 hidden units through ReLU activation function. Lastly, predictions were generated using the softmax activation function.

In the second case, *ling_set2* were combined HAN model. These external features are concatenated with the output of the attention layer of HAN. We fed the concatenated vectors through a fully connected layer with 200 hidden units and the ReLU activation function. Then, the dropout regularization with a rate of 0.1 was performed to the hidden layer. Finally, the softmax activation function was utilized to create predictions.

In the third case, the pre-trained *ling_set1* as well as *ling_set2* were concatenated. Our proposed architecture was presented in Figure1. Essentially, the previous two models were merged. *ling_set1* and the output of the attention layer with dropout regularization were concatenated and fed to a fully connected layer with 200 hidden units and the ReLU activation function. Then, these document vectors were concatenated to *ling_set2* and processed through a fully connected layer with 200 hidden units and the ReLU activation function. We implemented dropout regularization with a rate of 0.2 to the hidden layer. Lastly, the predictions are created along with the softmax activation function.

3.3.2. BERT for Hate-speech Detection

Transformers based BERT offers a powerful solution for context heavy texts with its structure that bidirectionally examines the incoming text and combines the masked language and next sentence prediction models. The pre-trained Turkish language model, BERTurk with 12 transformers blocks was trained on several Turkish corpora such as the OSCAR corpus[4], a recent Wikipedia dump, and various OPUS corpora[5].

[4]https://oscar-corpus.com/

[5]https://opus.nlpl.eu/

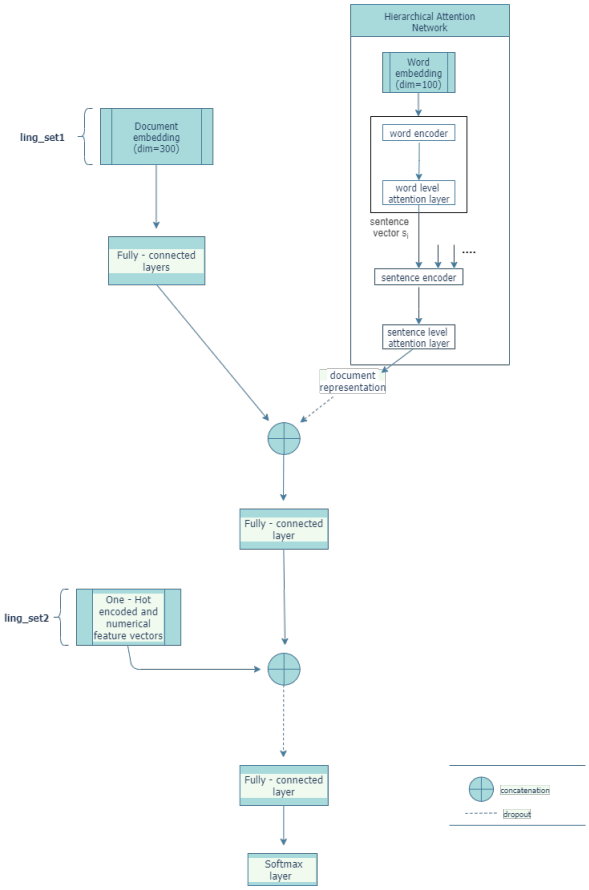

Figure 1: The proposed model incorporates HAN with linguistic features, *ling_set1* and *ling_set2*.

We fine-tuned this uncased BERT model for the detection of hate speech task using the compiled media data. The recommended hyperparameters by (Devlin et al., 2019) were evaluated. Batch-size and common learning rate were chosen as 16 and 5e-5, respectively. Document embeddings were constituted through obtaining the vectors corresponding to the [CLS] token from the final Transformer layer of this fine-tuned BERT model.

BERT embeddings with Novel Linguistic Features

The 768 dimensional feature vector taken from the final transformer layer of BERT model were concatenated with *ling_set1* and *ling_set2*, as similarly in Figure 1. Instead of the output vectors of the attention layer of HAN, vector sequences provided by BERT have been merged with *ling_set1* and passed to a fully connected layer with 200 hidden units and the ReLU activation function. Then, concatenation of *ling_set2* and feature vectors from the previous layers followed same fully connected layer and Dropout regularization steps as in the proposed model. Finally, softmax activation function is performed for obtaining predictions.

4. Experiments and Results

To the best of our knowledge, the only prior work on hate speech detection in Turkish news articles has been conducted by (Coban and Filatova, 2019). They used a data set, where the not-hate speech articles have been sampled from CNN[6] and BBC[7] news under the assumption that they do not include hate-inciting content. We compiled a larger data set, where both hate speech and not-hate speech classes have been manually annotated and evaluated the methods proposed by (Coban and Filatova, 2019), namely SVM with linear kernel, Naive Bayes and Multilayer Perceptron with TF-IDF weighted character and word n-grams, in order to serve as baseline for our proposed linguistically enhanced neural models. In addition, we implemented a Logistic regression classifier and performed a grid search through the validation data set split to get the optimized parameters of the classifiers.

According to our results in Table 1, the overall scores with word n-gram are higher than the ones achieved with char n-grams. Logistic regression obtain the highest scores in all metrics, while the second-highest scores are reached through SVM with linear kernel.

Additionally, the performance of HAN has been compared with CNN and CNN-GRU. It is stated that in the literature, CNN and RNN have been used separately (Le and Mikolov, 2014) and together (Zhang et al., 2018) on social media data and significant performances have been obtained. We have evaluated a CNN model that is based on the model of (Kim, 2014) with 3 parallel convolution layers and kernel sizes of 3, 4, 5 of words with filter size 100 of each for feature extraction. As a state-of-art based model, we have also replicated the CNN-GRU architecture in (Zhang et al., 2018). To maintain consistency, these models have been evaluated on the test set with 3662 documents, 20% of the overall dataset. The word embedding vectors trained via fastText were applied in all deep artificial neural network models. Also, the average evaluation scores with three different fixed seeds and three experimental runs in each fixed seed have been computed for the sake of reliability of the results.

We have observed that HAN outperforms the evaluation scores of traditional ML-based approaches and CNN-based approaches showed in Table 1 and Table 2 in all metrics. Addition of the GRU recurrent layer to the CNN improved the accuracy and macro average f-score with 0.2%. While CNN is good at feature extraction in comparison to the traditional machine learning models, GRU brings the capability of learning sequence dependencies. It can be stated that the attention based HAN model is more compatible with the features of our dataset for the task of hate speech detection. We have compared the performance of HAN with the proposed feature sets. The results in Table 2 show that both *ling_set1* and *ling_set2* enhance the performance of the HAN model and the best results are achieved when the two feature sets are used together.

Our experiments have been extended to BERT which is among the recent state-of-the-art models for hate speech detection and categorization (Wiedemann et al., 2020). Although the BERT constrained with 512 characters long, BERT base model performed better than both HAN base and HAN with linguistic feature sets model. We examined the effect of linguistic features on the BERT model, which significantly affected the performance of the HAN model, as explained in Randomization test section. According to the Table 2, although the linguistic features slightly increased the performance of the BERT model, it is concluded that there is no statistical difference between the two models. (Rogers et al., 2020) stated that BERT embeddings hold especially semantic and syntactical knowledge through multi-head attention layers. Obtained result showed the possibility that the BERT model implicitly capturing the linguistic features which are beneficial for hate speech detection task.

5. Analysis

5.1. Randomization test

We have performed a randomization test (Yeh, 2000), which is widely used in NLP, to examine if there is a significant difference between proposed models. The null hypothesis that *'there is no significant difference between the models'* is rejected when p is less than 0.025 with significance level of alpha = 0.05. With the 9 outcomes from each model, 81 different p-values and their harmonic mean is calculated. With comparison of HAN base and HAN with the linguistic features models, the p-value scores were calculated for hate speech class is 0.007 and the p-value for not hate speech class is 0.008. According to our test statistics the null hypothesis is rejected for both classes. It proves that there is a significant difference between HAN base and HAN with the linguistic features models, suggesting that the novel linguistic features bring further improvement to the HAN model. Another comparison was made between BERT model and BERT with the linguistic features model and the obtained p-value for the hate speech class is 0.188 and for not hate speech class is 0.147. These test statistics state that the null hypothesis is not rejected for both classes and there is no significant difference between these two models.

5.2. Error analysis

Dependency parsing and POS tagging errors affected the feature extraction process and the overall performance. These errors are caused by the parser as well as by OCR.

In addition, we observed that many errors were caused by the incorrect classification of news articles that contain discriminatory language, but not hate speech, as

[6]https://www.cnnturk.com/

[7]https://www.bbc.com/turkce

			accuracy	precision	recall	fscore	fscore macro avg
word (1,2)-gram + tf-idf	SVM	hate_speech	0.857	0.849	0.862	0.856	0.857
		not_hate_speech		0.865	0.852	0.858	
	Logistic Regression	hate_speech	0.864	0.856	0.869	0.862	0.864
		not_hate_speech		0.872	0.858	0.865	
	MultinomialNB	hate_speech	0.810	0.790	0.835	0.812	0.81
		not_hate_speech		0.831	0.785	0.807	
	Multilayer Perceptron	hate_speech	0.834	0.839	0.821	0.830	0.834
		not_hate_speech		0.830	0.847	0.839	
char 2-gram + tf-idf	SVM	hate_speech	0.781	0.771	0.789	0.780	0.781
		not_hate_speech		0.792	0.774	0.783	
	Logistic Regression	hate_speech	0.777	0.768	0.784	0.776	0.777
		not_hate_speech		0.787	0.771	0.779	
	MultinomialNB	hate_speech	0.721	0.741	0.666	0.701	0.720
		not_hate_speech		0.705	0.775	0.739	
	Multilayer Perceptron	hate_speech	0.789	0.764	0.826	0.794	0.789
		not_hate_speech		0.817	0.753	0.784	

Table 1: Evaluation scores for the traditional machine learning based methods

		accuracy	precision	recall	fscore	fscore macro avg
CNN word	hate_speech	0.872	0.862	0.887	0.872	0.872
	not_hate_speech		0.890	0.859	0.872	
CNN + GRU	hate_speech	0.874	0.893	0.849	0.869	0.874
	not_hate_speech		0.864	0.899	0.879	
HAN base	hate_speech	0.889	0.880	0.899	0.888	0.889
	not_hate_speech		0.902	0.879	0.890	
HAN with *ling_set1*	hate_speech	0.895	0.867	0.927	0.898	0.896
	not_hate_speech		0.928	0.861	0.893	
HAN with *ling_set2*	hate_speech	0.893	0.860	0.935	0.896	0.893
	not_hate_speech		0.932	0.853	0.891	
HAN with *ling_set1* + *ling_set2*	hate_speech	0.897	0.883	0.911	0.897	0.897
	not_hate_speech		0.911	0.883	0.897	
BERT	hate_speech	0.904	0.905	0.914	0.906	0.904
	not_hate_speech		0.909	0.894	0.902	
BERT with *ling_set1* + *ling_set2*	hate_speech	0.906	0.901	0.909	0.907	0.906
	not_hate_speech		0.910	0.903	0.904	

Table 2: Evaluation scores of neural network based approaches

belonging to the hate speech class by the classification models, revealing the challenge of distinguishing hate speech from discriminatory language.

6. Conclusion

In this study, a dataset for detection of hate speech in Turkish has been compiled by retrieving 18316 national and local print media news articles. The manual annotations were obtained from the Hrant Dink Foundation, who have been working on manually detecting hate speech in the Turkish media since 2009 and have been releasing annual reports to raise awareness. By utilizing these manually annotated data, a hybrid approach based on deep learning and linguistic features has been developed for Turkish hate speech detection.

Considering the qualitative analysis of hate discourse in the Turkish language, several linguistic features have been designed. The HAN and BERT models were en-

hanced with these novel features and the performance of the new models was analyzed. Our results indicated that the HAN model is able to address the changing interest weights of words based on the context by taking account of the natural segmentation of documents. Better results compared to CNN and CNN-GRU based models have been obtained for hate speech detection using the HAN base model. Combining HAN with pretrained othering and imperative language based features as well as with information about reported speech forms further enhanced the performance. BERT based models have also been fine-tuned for the task of hate speech detection, which achieved the highest performances. The BERT model with the linguistic features closely follows the BERT base model in terms of F-score, and randomization test has shown that there is no significant difference between these two models. It concluded that, BERT model may be implicitly capturing the linguistic features which are beneficial for hate speech detection task. With the developed methods, we aim to minimize dependence on human labor for the identification of hate speech, which is crucial for the elimination of discrimination.

As future work, we are planning to investigate other linguistic properties and what features are most relevant for hate speech detection in the Turkish Language as well as exploring the inductive bias provided by linguistic features with various sizes of the data.

7. Acknowledgements

This research was partially supported by the Swedish Consulate-General, İstanbul Turkey (Project number: UM2021/10687/ISTA). We are also grateful to Hrant Dink Foundation for their collaboration and for sharing their resources which provided us with invaluable data.

8. Bibliographical References

Alorainy, W., Burnap, P., Liu, H., and Williams, M. L. (2019). "the enemy among us" detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, 13(3):1–26.

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

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

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

Burnap, P. and Williams, M. L. (2016a). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Burnap, P. and Williams, M. L. (2016b). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Coban, E. B. and Filatova, E. (2019). Incendiary news detection. *Association for the Advancement of Artificial Intelligence*.

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Lrec*, volume 6, pages 449–454.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Hüsünbeyi, Z. M. (2020). Detecting hate speech in turkish texts. Master's thesis, Bogaziçi University.

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

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Kress, G. and Hodge, R. (1997). Language as ideology. *The Modern Language Journal*, 64:512.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Liu, P., Li, W., and Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Mishraa, A. K., Saumyab, S., and Kumara, A. (2020). Iiit_dwd@ hasoc 2020: Identifying offensive content in indo-european languages.

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

Oktar, L. and Değer, A. C. (1999). Gazete söyleminde kiplik ve İşlevleri. *Dilbilim Araştırmaları Dergisi*, pages 45–53.

Oktar, L. (2001). The ideological organization of representational processes in the presentation of us and them. *Discourse & Society*, 12(3):313–346.

Ozdemir, A. and Yeniterzi, R. (2020). Su-nlp at semeval-2020 task 12: Offensive language identification in turkish tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2171–2176.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

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

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

Van Dijk, T. A., (2011). *Discourse, knowledge, power and politics*, pages 27–64.

Wang, S., Liu, J., Ouyang, X., and Sun, Y. (2020). Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455.

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

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

Wiedemann, G., Yimam, S. M., and Biemann, C. (2020). UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online), December. International Committee for Computational Linguistics.

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

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

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

Çınar, M. (2013). Habercilik ve nefret söylemi. In *Medya ve Nefret Söylemi: Kavramlar, Mecralar, Tartışmalar. Ed. Mahmut ÇINAR.*, İstanbul. Hrant Dink Vakfı Yayınları.

9. Language Resource References

Hrant Dink Foundation. (2021). *The Hate Speech Digital Archive*. Media Watch on Hate Speech project.

Appendix: Reported Speech Forms List

dedi	's/he said'
söyledi	's/he told'
açıkladı	's/he explained/ announced'
açıklar	's/he/it explains/ announces'
açıkladılar	's/he/it explained/ announced'
belirtir	's/he states that'
belirtti	's/he stated that'
belirttiler	'they stated that'
diye konuştu	's/he stated that'
diye konuştular	'they stated that'
kaydetti	's/he noted'
kaydettiler	'they noted'
dile getirdi	's/he mentioned'
dile getirir	's/he mentions'
dile getirdiler	'they mentioned'
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/ accused'
suçlar	's/he/it blames/ accuses'
suçladılar	'they blamed/ accused'
tepkilere yol açtı	'it caused reactions'
tepkilere yol açtılar	'they caused reactions'
şikayet etti	's/he reported/ complained'
şikayet eder	's/he reports/ complains'
şikayet ettiler	'they reported/ complained'
karşılık verdi	's/he responded'
karşılık verdiler	'they responded'

Table 3: The list with 30 tokens in Turkish and their English translations to detect news articles including reported speech forms