Detecting Gender-based, Homophobic, Religious, and Ethnic Hate Speech in Turkish Tweets

Final Report

Kübranur Umar
Artificial Intelligence Engineering
TOBB ETU
Ankara, Türkiye
kumar@etu.edu.tr

***Abstract*—In this project, a detailed study on detecting and categorizing hate speech in Turkish tweets was proposed. The project is examined in two parts: first part is about determining whether a tweet contains hate speech, and second part is about classifying the target of the detected hate speech into categories such as gender-based, homophobic, religious, or ethnic. An existing dataset is used for the first part. For the classification, we created a dataset from multiple datasets using a keyword-based labeling. Two NLP models BerTurk and TurkishBerTweet also, traditional machine learning models including Naive Bayes and Logistic Regression were trained and evaluated on this datasets to analyse the hate speech. In addition, a pre-trained large language model ChatGPT 4o was tested on test dataset in order to compare its performance with Bert-based models. The results show that the BerTurk model outperformed both TurkishBerTweet and ChatGPT in detection of hate speech and categorization of the tweets. This research aims to provide a detailed explanation on newly created target-based and already existing datasets and compare the model's performances on these datasets.**

***Index Terms*—Hate Speech, BERTurk, TurkishBerTweet, Target-based**

I. Introduction

A text that contains insults, demeans and violent words is accepted as hate speech. In the world of technology hate speech can be encountered regularly especially on social media platforms and, this problem is becoming a significant concern due to its potential of taking lead in violence and discrimination. The detection of this issue is crucial in order to prevent its harmful effects and ensure safer online environments. Understanding and identifying a text including hate speech is challenging because of the structure and cultural differences of the language studied. Since hate speech can be constructed in different ways according to the structure and culture of the used language it is important to make a comprehensive study on the language to detect hate speech more effectively. This study aims to develop a system to detect and classify hate speech in Turkish tweets by addressing these challenges. The project is split into two parts which are detection and classification of hate speech into four categories: gender-based, homophobic, religious or ethnic. This categorization is mainly based on predetermined keywords. For the first part of the project we utilized already existing dataset created by Tanyel et al. [10] in order to identify whether a text is hate speech. In the second part of the project multiple existing datasets were utilized to create a comprehensive labeled dataset by using keyword-based labelling to categorize hate speech into distinct types: gender-based, homophobic, religious, or ethnic. Two pre-trained language models, BerTurk and TurkishBerTweet, were trained for both detection and categorization tasks and, evaluated to determine their reliability and effectiveness in these datasets. Additionaly, ChatGPT 4o being best version of large language model ChatGPT was utilized for target based classification task. In addition to language models we trained traditional machine learning models such as Naive Bayes and Logistic regression for both tasks of the project. The comparison of all model's performances showed that BerTurk model outperformed other algorithms by resulting 88% F1 score for the first task and 94% F1 score for the second, target-based, task in the training.

II. Related Work

In hate speech detection Transformer-based models, especially the BERT architecture based ones, have shown significantly better results. Because of the ability to capture contextual relationships within the text of The Bert Model, researchers have commonly been using the Bert model as a first option in their studies. A Turkish hate speech dataset specifically targeting tweets related to the Istanbul Convention and refugees was created and analysed by Beyhan et al. (2022) [1]. The study trained both the BERTurk model and traditional methods in order to develop a machine learning system. They annotated 1278 tweets about refugees and 1206 tweets on gender-based violence. With the BERTurk model they achieved an accuracy of 77% on the Istanbul Convention dataset and 71% on the Refugee dataset. In the study of Hüsnübeyi et al. [5] they aimed at identifying religious and ethnic hate speech in Turkish news articles. A dataset of 18,316 manually annotated articles was developed. The study achieved an F1-score of 90.7% with the BERT model by

using BERT and Hierarchical Attention Network (HAN) models. Karayiğit et al. [7] developed the Homophobic-Abusive Turkish Comments (HATC) dataset, including 31,290 Instagram comments categorized into homophobic, hateful, and neutral. Multilingual BERT (M-BERT) model was employed for classifying the comments and, the model achieved F1-scores of 82.64% for homophobic, 91.75% for hateful, and 96.08% for neutral comments. In the study of Tanyel et al. [10] a comprehensive dataset for detection of offensive language in Turkish tweets was created. They utilized from multiple sources to form a balanced dataset of 42,486 samples. They evaluated several models and found that a BERT-CNN-BiLSTM pipeline achieved the best results, with 96% F1 score in the test set of the created dataset without using text normalization. Hate Speech Detection in Turkish and Arabic Tweets was provided by Uludoğan et al. (2024) [11]. Two sub-tasks were presented: Sub-task A includes hate speech detection in Turkish tweets, sub-task B includes hate speech detection in Arabic tweets. For sub-task A they employed ConvBERTurk model and, the model managed to produce successful results for task A achieving an F1 score of 0.69645 on private test set thus, outperforming other Bert based models (ReBERT: 0.73786 public, 0.68886 private; VRLLab: 0.70642 public, 0.66432 private). In the study of Uludoğan et al. [12] They introduced HateTargetBERT, a Bert based model, for hate speech detection. In addition a dataset of 6681 Turkish news articles annotated for hate speech targeting ethnicity, nationality, and religious identity was created with the name of TurkishHatePrintCorpus. They evaluated the HateTargetBert model in this dataset and, it outperformed BERTurk with an F1-score of 89.60%. Dehghan et al. (2024) [3] used ChatGPT for hate speech detection in Turkish Tweets using zero- and few-shot paradigms. ChatGPT achieved a 65.81% accuracy in few-shot settings, whereas the BERT model with supervised fine-tuning achieved 82.22% accuracy. This clearly shows that the Bert model is better suited for NLP classification tasks. In the project of Najafi and Varol (2024) [9] the TurkishBERTweet model was proposed in order to develope a hate speech detection system for Turkish tweets. They fine-tuned the model with Low-Rank Adaptation (LoRA). They utilized a dataset including 9,140 tweets focused on conflicts and anti-refugee content, provided by the Hrant Dink Foundation. The proposed model, TurkishBERTweet+LoRA, achieved a weighted F1-score of 0.8137 in a 5-fold cross-validation. The study of Zampieri et al. (2023) [13] developed the Target-Based Offensive language (TBO) dataset including both post-level annotations regarding the harmfulness of offensive posts and token-level annotations that identify the target and offensive argument, comprising over 4,500 English Twitter posts. The authors employed RACL-BERT model achieving the highest performance with a Target F1 score of 0.442, Argument F1 score of 0.256, and a Harmfulness Macro F1 score of 0.693. İhtiyar et al. (2023) [6] developed a dataset including 28,000 Turkish tweet-reply pairs. This dataset consists of offensive language against unvaccinated people during the Covid-19 pandemic. They evaluated various models, such as traditional models like Logistic Regression (LR) and Support Vector Machines (SVM), and deep learning models like BERTurk. The results indicated that BERTurk achieved an F1-score of 73.57% when using contextual information. In this study we utilized the ChatGPT, BerTurk and TurkishBerTweet model to compare their performances for hate speech detection in both not target-based dataset [2] and target-based dataset created from various manually labeled datasets. In addition, we developed a system from the best model in order to detect hate speech and determine its target audience.

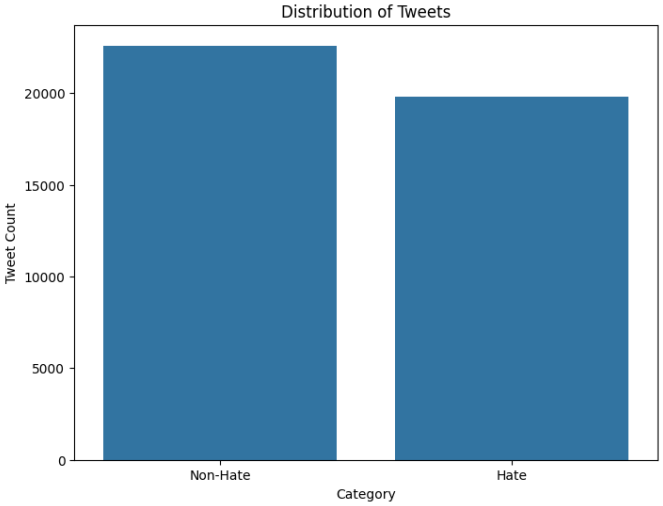

Fig. 1. Distribution of Non-Targeted Dataset

III. Data

A. Data Collection

In this study a collection of datasets from various studies were utilized to create a successful hate speech detection task for Turkish tweets. For the non-target-based part of the project the dataset from Tanyel et al. [10] which is a combination of Coltekin's OffensEval 2020 [2], Mayda et al. [8] and Kaggle dataset were used for training of models. This combined dataset is a significant resource for non-target-based hate detection by being a balanced dataset among other already existing datasets. The distribution of hate and non-hate data is avaliable at Fig. 1. This distribution shows that the data is balanced thus, suitable for accurate analysis.

In the second part of the project three already existing datasets were used for annotation into specific targets: gender-based, homophobic, religious and ethnic. The first dataset utilized from Mayda et al. [8]. This dataset is already target-based annotated for gender-based, ethnic and religious targets and, its 50 religious targeted data entries were included in the final target-based dataset. Beyhan et al. [1] dataset was utilized for gender-based and homophobic. The hate speech entries of the set were composed of two distinct collections: the Istanbul Convention Dataset and the Refugee Dataset. The Istanbul Convention parts covers both gender-based and homophobic hate thus, a manual labelling was performed in order to get

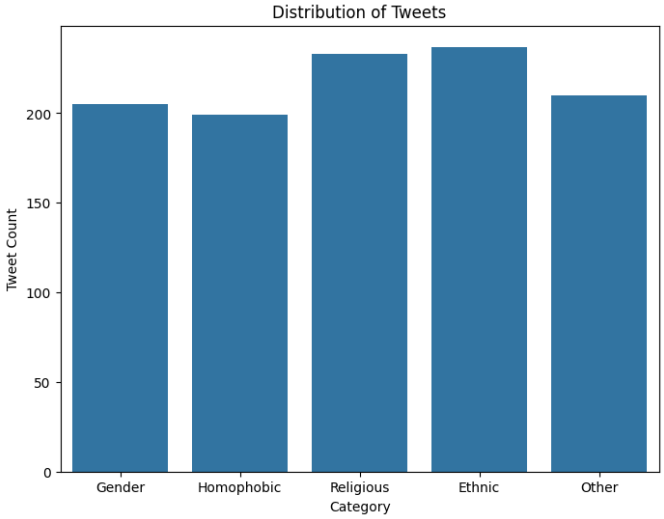

Fig. 2. Distribution of Targeted Dataset

TABLE I
THE NUMBER OF ENTRIES TAKEN

	Mayda et al.	Beyhan et al.	Tanyel et al.
Gender-Based	-	61	144
Homophobic	-	160	41
Religious	50	-	183
Ethnic	-	-	237
Other	-	-	210

separate entries for both targets. The last dataset that was used in the final set is from Tanyel et al. [10]. This dataset was manually labelled into gender-based, homophobic, religious, ethnic and other based on predetermined keywords. The target determined as "other" covers the targets except the four targets already studied. To accurately label the data for target-based hate speech detection, the predetermined keywords were crucial. The keywords for each target category are as follows:

- Ethnic: suriyeli, ermeni, ingiliz, kürt, türk, yunan, arap, rum, çinli
- Religious: alevi, müslüman, yahudi, ateist, gavur, hıristiyan, kafir, putperest
- Homophobic: top, ibne, lgbt
- Gender: orospu, amk, yollu, fahişe, kaşar, kaltak

In order to construct final target-based dataset all three manually labeled and split entries were combined and prepared for model training. There are total 1086 entries including 2 duplicates thus, these duplicates removed. The size of the final dataset resulted as 1084. Table 1 shows that how many number of entries from already existing datasets are included in the final dataset. In addition, Fig. 2. presents the distribution of targets in the final dataset, showing that the dataset is balanced and appropriate for model training.

An example texts and target-based labels is showed in Fig. 3. for better understanding of the targets.

Text	Label
Kaltak karı ödev yapıyoz burda kes bi çeneni. Her mahallede varmı amk bu çeneli karılar	Gender-Based
gay misin yuvarlak misin top musun lastik misn nesin sen	Homophobic
Umarım birgün bana da denk gelir şu gavur tohumlarından biri...	Religious
Mutlu günlerinde , kötü günlerinde arkadaşlarımın yanında olamıyorum.. Allah cezanı versin piç çinli!	Ethnic

Fig. 3. Examle Target-Based Hate Speech Texts

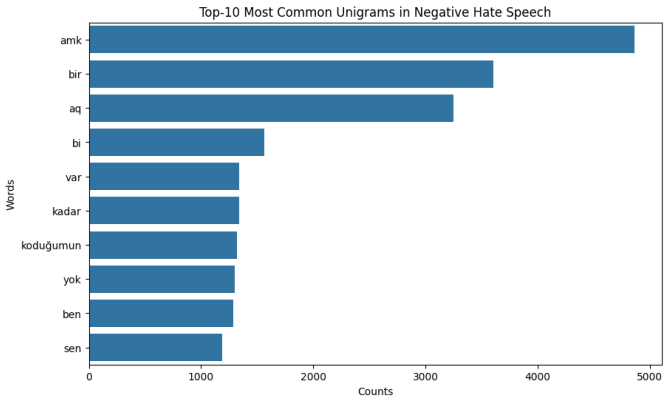

Fig. 4. Most Common Unigrams in Hate Speech Texts

B. Data Preprocessing

For the purpose of ensuring quality and consistent datasets several preprocessing steps were applied to the data. The non-target-based Tanyel et al. [10] and the final target-based dataset were normalized with these preprocessing steps including removing user handles, hashtags, emojis, and punctuation marks, and converting all text to lowercase. In Fig. 4. the common unigram words in the non-target-based dataset after normalization can be seen and, Fig. 5-8. show the most common unigrams in the target-based dataset after normalization step. This normalization step was crucial for standardizing the text and improving the performance of the NLP and traditional machine learning models.

C. Challenges and Solutions

Labeling the data based on specific targets was challenging because, deciding the accurate label for each instance was difficult, as some sentences exhibited characteristics of more than one target, or in some cases, did not clearly fit any specific target. This ambiguity required careful consideration and consistent annotation guidelines to ensure accuracy in the

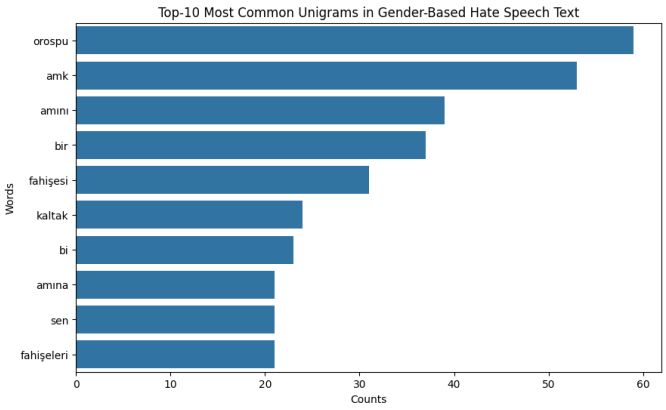

Fig. 5. Most Common Unigrams in Gender-Based Hate Speech Texts

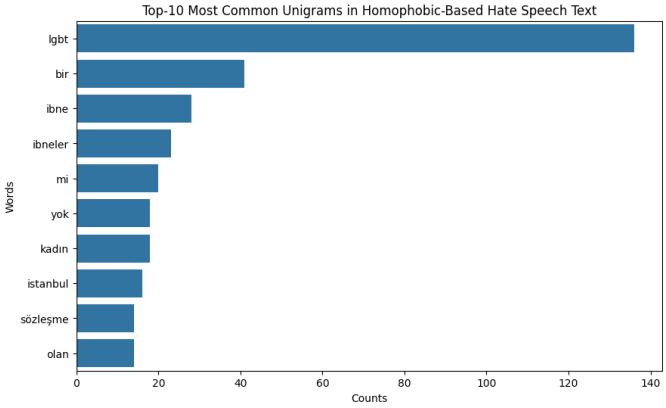

Fig. 6. Most Common Unigrams in Homophobic-Based Hate Speech Texts

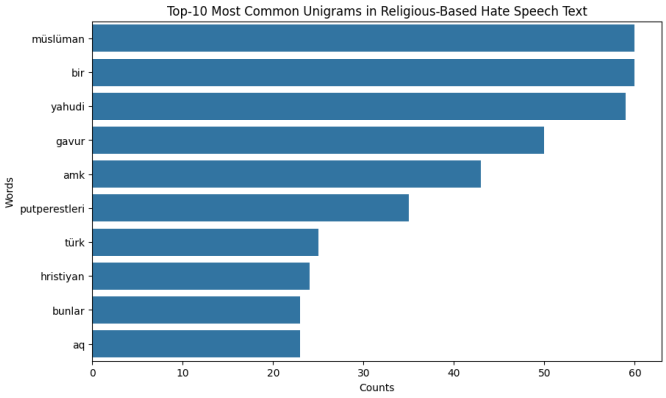

Fig. 7. Most Common Unigrams in Religious-Based Hate Speech Texts

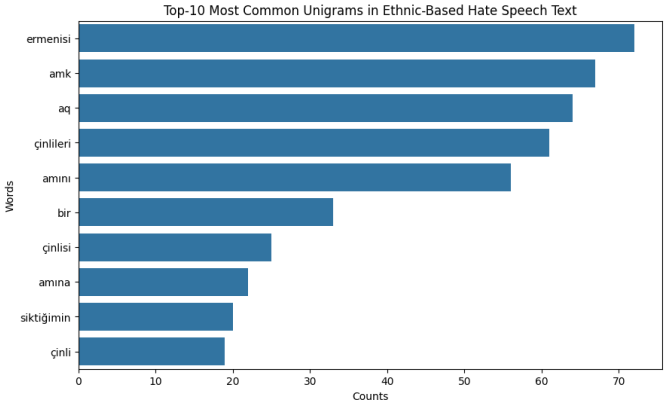

Fig. 8. Most Common Unigrams in Ethnic-Based Hate Speech Texts

labeling process. Some examples of this ambiguity are shown below:

- Bak bana orospu evladı. Ben yatay geçiş yaptım diye hala aktif olarak okumama rağmen benim geri ödeyeceğim kredim kesiliyor. Tüm para bu ibnelere gidiyor dememiz gayet normal. Sikerim belanı.
  - Target: Gender-Based, Homophobic
- en azından ben ağzımı bozdum sizin gibi karakterimi değil amına koduğumun salakları
  - Target: None

IV. Methodology

In this study we utilized BERT [4] based models being state-of-the-art language model with high performance in NLP tasks. The first model used for hate speech detection task is BERTurk which is a pre-trained model especially designed for Turkish, leveraging the BERT architecture. Thanks to the its superior ability to understand and process Turkish syntax and semantics, the BERTurk model was selected to accurately detect hate speech in Turkish tweets.

In addition to BERTurk model, TurkishBerTweet model, also leveraging the BERT architecture, proposed by Najafi and Varol (2024) [9] was utilized with its fine-tuned version with low-rank adaptatition (LoRA). TurkishBERTweet being a new language model was specifically trained on nearly 894M Turkish tweets.

We also utilized from ChatGPT 4o in order to classify hate speech text into targets by using zero-shot prompt inspired by the technique proposed in Dehghan et al. (2024) [3]. ChatGPT is an AI chatbot with natural language processing (NLP) that allows you to have human-like conversations to complete various tasks[1]. We used zero-shot prompt that is available at Fig. 9. for GPT to classify the texts.

Finally, traditional machine learning models which are Naive Bayes and Logistic Regression was employed as base models for hate speech detection. In the field of text classification these models offer a comparative baseline to evaluate

[1]https://www.zdnet.com/article/what-is-chatgpt-and-why-does-it-matter-heres-everything-you-need-to-know/

Zero-shot prompt	You are an AI language model trained to analyze and detect hate speech. Given a tweet, decide which target the hate speech text contains. Assign a target number: gender:0, homophobic:1, religious:2, ethnic:3, other:4 to each text in a format of: Target: 'Target' **Tweet:** toplumu bozar sapkinliktir lanetlenmistir gecmiste bir toplum sirf bunlar yuzunden helak olmustur lgbt ye hayir
ChatGPT Response	**Target:** 1

Fig. 9. Zero-shot prompt and ChatGPT response for an hate speech

the performances of BERT based models. The study aims to highligth the improvements and success of transformer-based models in the context of hate speech detection.

In order to provide a comprehensive analysis of hate speech detection in Turkish, this study compare both BERT-based models and traditional models and, demonstrate the effectiveness of transformer-based approaches.

V. Results

In this study, a BERTurk model on the non-target-based dataset from Tanyel et al. [10] was trained using specific training parameters. The model demonstrated impressive performance on the test dataset, achieving an evaluation accuracy of 92.29% and an F1-score of 92.29%. In order to compare BERTurk with another BERT-based model, we trained TurkishBertTweet+LORA model for accurate model selection. Similar to BERturk model, this model is trained on the same non-target-based dataset, achieving accuracy of 91%, F1 score of 91%. To provide a comparative analysis, we also trained Naive Bayes and Logistic Regression models on the same non-target-based dataset. The Naive Bayes model yielded an accuracy of 90.00%, resulting in an F1-score of 0.90 for class 0 and 0.89 for class 1. Similarly, the Logistic Regression model achieved an accuracy of 90.00%, leading to an F1-score of 0.91 for class 0 and 0.90 for class 1. For non-target-based dataset BERTurk model outperformed both TurkisBerTweet and traditional machine learning models in terms of accuracy and F1-score. The comperative results can be found in Table 2 for test dataset.

The target-based dataset was utilized to train BERTurk, TurkisBerTweet, Naive Bayes and Logistic Regression with 5 classes. These classes are gender, homophobic, religious, ethnic and other labelled as 0,1,2,3,4 respectively. In addition to other models, some samples from target based retrieved and submitted to GPT for classification. GPT was tasked with analyzing a series of tweets containing hate speech and classifying each one based on its target. GPT examined the content of each tweet and assigned the appropriate target number, resulting a 60% F1-score and 61% accuracy. BERTurk model, again, outperformed other models in the test dataset, leading to 96% F1-score in detecting targeted hate speech in Turkish tweets. The comperative results can be found in Table 3 for test dataset.

TABLE II
Model Results For Non-Target-Based Dataset in The Test Data

	Accuracy	F1	Recall	Precision
BERTurk	92.29	92.29	92.29	92.30
TurkishBerTweet	90.61	90.60	90.61	90.65
Naive Bayes	90.00	89.50	89.50	90.00
Logistic Regression	90.00	90.50	90.50	91.50

TABLE III
Model Results For Target-Based Dataset in The Test Data

	Accuracy	F1	Recall	Precision
BERTurk	96.33	96.34	96.70	96.33
TurkishBerTweet	87.15	87.15	88.06	87.15
ChatGPT 4o	61.46	59.66	61.46	78.28
Naive Bayes	69.00	67.40	68.60	68.60
Logistic Regression	72.00	73.60	72.60	79.80

The higher performance of BERTurk underscores the effectiveness of transformer-based models in hate speech detection tasks. The confusion matrices for Naive Bayes and Logistic Regression models (provided in Fig. 10. and 11.) for non-target-based dataset further illustrate the distribution of predictions across different classes, highlighting the precision and recall trade-offs in each approach. Example results of model tested by sample tweets different from test dataset are shown in Fig. 12.

VI. Conclusion

This study made significant advancements in detecting hate speech in Turkish tweets by utilizing both traditional and advanced machine learning models. The BERTurk model, trained on a non-target-based dataset from Tanyel et al. [10],

Tweet: bu lgbtli sapıklar hep bizi buluyor bunlar defolsun gitsin ibneler **Target:** Homophobic
Tweet: suriyelileri sevenin amk hepsi orospu çocuğu siktirsinler kendi ülkelerine **Target:** Ethnic
Tweet: bu karılar böyle yollu hep işleri güçleri saçmalık aq **Target:** Gender
Tweet: bu gavur ateistlerin kökünü kurutmak lazım **Target:** Religious
Tweet: orospu çocuğu gavurlar olmasa dünyamız daha iyi bi yer olurdu **Target:** Religious
Tweet: ülkemde mülteci istemiyorum **Target:** Ethnic
Tweet: bu ne saçma sapan bir iş ya ne yaptığınız belli değil asalaklar **Target:** Other
Tweet: yunan salakları boş konuşmayın **Target:** Ethnic

Fig. 10. Bert Test Results

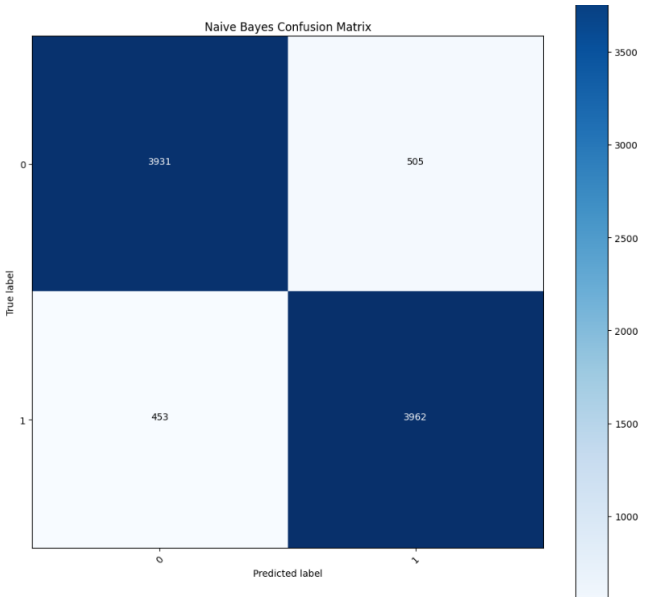

Fig. 11. Confusion Matrix for Naive Bayes

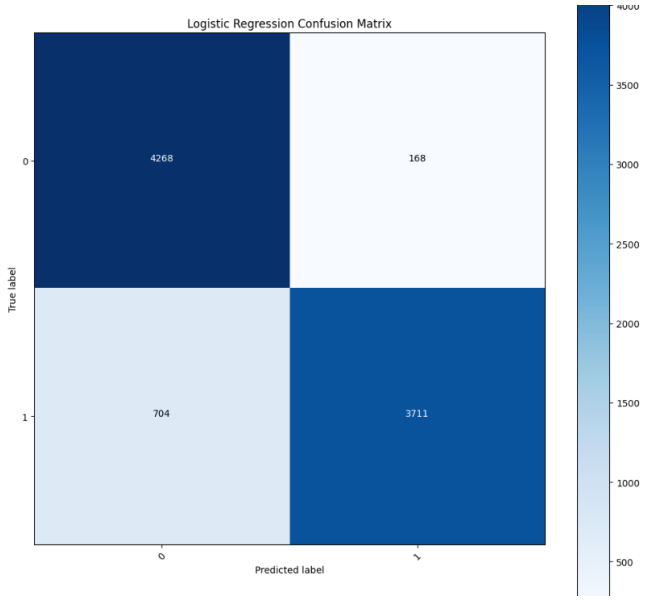

Fig. 12. Confusion Matrix for Logistic Regresion

achieved an accuracy of 92.29% and an F1-score of 92.29%, outperforming ChatGPT, TurkishBerTweet, Naive Bayes and Logistic Regression models. Crucial preprocessing steps included normalizing the dataset to remove user handles, hashtags, emojis, and punctuation, and converting text to lowercase. Additionally, target-based data from Beyhan et al. [1], Mayda et al. [8] and Tanyel et al. [10] were manually labeled for categories like gender, homophobia, ethnicity, and religion, totaling 1084 entries. In the future works, more targets will be taking consideration for hate speech detection and, more

annotating for targets will be planning to completed aiming to obtain larger dataset. The presentation video for the project can be accessible in https://youtu.be/yIh9ZSKJtKY

REFERENCES

[1] Fatih Beyhan, Buse Çarık, Inanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. A turkish hate speech dataset and detection system. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4177–4185, 2022.

[2] Çağrı Çöltekin. A corpus of turkish offensive language on social media. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6174–6184, 2020.

[3] Somaiyeh Dehghan and Berrin Yanikoglu. Evaluating chatgpt's ability to detect hate speech in turkish tweets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 54–59, 2024.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Zehra Melce Hüsünbeyi, Didar Akar, and Arzucan Özgür. Identifying hate speech using neural networks and discourse analysis techniques. In *Proceedings of the first workshop on language technology and resources for a fair, inclusive, and safe society within the 13th language resources and evaluation conference*, pages 32–41, 2022.

[6] Musa Ihtiyar, Ömer Özdemir, Mustafa Erengül, and Arzucan Özgür. A dataset for investigating the impact of context for offensive language detection in tweets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1543–1549, 2023.

[7] Habibe Karayiğit, Ali Akdagli, and Çiğdem İnan Aci. Homophobic and hate speech detection using multilingual-bert model on turkish social media. *Information Technology and Control*, 51(2):356–375, 2022.

[8] İslam Mayda, Yunus Emre Demir, Tuğba Dalyan, and Banu Diri. Hate speech dataset from turkish tweets. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2021.

[9] Ali Najafi and Onur Varol. Vrllab at hsd-2lang 2024: Turkish hate speech detection online with turkishbertweet. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 185–189, 2024.

[10] Toygar Tanyel, Besher Alkurdi, and Serkan Ayvaz. Linguistic-based data augmentation approach for offensive language detection. In *2022 7th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6. IEEE, 2022.

[11] Gökçe Uludoğan, Somaiyeh Dehghan, Inanç Arın, Elif Erol, Berrin Yanikoglu, and Arzucan Özgür. Overview of the hate speech detection in turkish and arabic tweets (hsd-2lang) shared task at case 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 229–233, 2024.

[12] Gökçe Uludoğan, Atıf Emre Yüksel, Ümit Tunçer, Burak Işık, Yasemin Korkmaz, Didar Akar, and Arzucan Özgür. Detecting hate speech in turkish print media: A corpus and a hybrid approach with target-oriented linguistic knowledge. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 205–214, 2024.

[13] Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. Target-based offensive language identification. 2023.