A New Hybrid LSTM-RNN Deep Learning Based Racism, Xenomy, and Genderism Detection Model in Online Social Network

Sule Kaya
Department of Software Engineering, Firat University, Elazig-23000
Email: sule.kaya@firat.edu.tr
ORCID: 0000-0001-5527-8913
Bilal Alatas
Department of Software Engineering, Firat University, Elazig-23000
Email: balatas@firat.edu.tr
ORCID: 0000-0002-3513-0329

ABSTRACT

Hate speech, which is a problem that affects everyone in the world, is taking on new dimensions and becoming more violent every day. The majority of people's interest in social media has grown in recent years, particularly in the United States. Twitter placed 5th in social media usage figures in 2022, with an average of 340 million users globally, and human control of social media has become unfeasible as a result of this expansion. As a result, certain platforms leveraging deep learning approaches have been created for machine translation, word tagging, and language understanding. Different strategies are used to develop models that divide texts into categories in this way. The goal of this research is to create an effective a new hybrid prediction model that can recognize racist, xenophobic, and sexist comments published in English on Twitter, a popular social media platform, and provide efficient and accurate findings. 7.48 percent of the data were classified as racist, genderist, and xenophobic in the used dataset. A new hybrid LSTM Neural Network and Recurrent Neural Network based model was developed in this study and compared with the most popular supervised intelligent classification models such as Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, and K-Nearest Neighbors. The results of these several models were thoroughly examined, and the LSTM Neural Network model was found to have the best performance, with an accuracy rate of 95.20 percent, a recall value of 48.94 percent, a precision of 60.95 percent, and an F1 Score of 51.32 percent. The percentage of test data was then modified, and the comparison was made by attempting to get various findings. With a larger dataset, these deep learning models are believed to produce substantially better outcomes.

Keywords - **Artificial Intelligence, Deep Learning, Genderism, Racism, Xenophobia**

Date of Submission: Jul 28, 2022 Date of Acceptance: Aug 27, 2022

I. INTRODUCTION

Internet users can utilize online social networking sites to stay in touch, share information about their everyday activities and interests, and upload and access documents, images, and videos. You can publish professional postings, have a list of colleagues to communicate with, and post and read posts from your circle of friends and others on social networks like Facebook, Twitter, Ask.fm, Instagram, LinkedIn, and Google+. Social networks, like search engines, are among the most visited websites [1]. However, social media platforms are ideal for the dissemination of damaging information. Social attacks such as cyberbullying, genderism [2], and pushing people to self-harm are some of the most effective outcomes of the broadcast of bad information. Many of these attacks are carried out by a single individual, but they can also be carried out by groups. For example, a celebrity's fan base or football team followers can make such hate remarks in groups. Trolls like these are primarily addressed at specific victims, but they can also be directed towards large groups of people who are discriminated against for reasons like hatred, ethnicity, or gender. Large groups of individuals may participate in such hate campaigns, and such feelings may lead to physical violence or violent crimes.

The use of social media has expanded dramatically in recent years, and people are spending significantly more time on these sites. Particularly popular social media programs such as Instagram, Twitter, and Facebook are able to bring people from all countries and backgrounds together around common interests. The vast majority of social media users openly communicate their sentiments and opinions on these platforms, and even unwittingly divulge too much about their personal lives, compromising user privacy. Furthermore, many users publish their sentiments and thoughts without applying any filters, owing to the confidence provided by anonymity, which is one of the characteristics provided by social networks, and do not consider how the other person may be impacted by this circumstance. Those who perceive persons of a different race as second-class citizens, in particular, frequently disseminate these opinions on social media and justify the situation as freedom of thought. People with religious, racial, or political fanaticism, in particular, can be exceedingly bigoted toward people who are physically or mentally different from them, and they can engage in sexist behavior by going beyond hate speech's bounds.

With the rise in social media usage, manually controlling the plethora of comments, messages, and information has become nearly impossible. Because such hate crimes are not reported to the police, the shortage of official information brings attention to the present concerns on social media. In this situation, social networks have a lot of content but are not very reliable. Despite this issue, due to the rich content of social media, the processing of users' data is extremely important. This helpful information is processed thanks to data mining, which uncovers hidden tendencies in datasets and allows for efficient investigations with the required decisions [3].

The aim of this study is to identify hate speech on Twitter, which includes genderism, racism, and xenophobia, and to classify it using some classification methods and forecast the tags in the test dataset. The main contributions of this study are listed below:

- The problem of detection of genderism, racism, and xenophobia was handled as a classification problem. A new hybrid Long-Short Term Memory-Recurrent Neural Network (LSTM-RNN) deep learning model for hate speech detection, which is one of the most popular social media analysis problems were adapted for the first time.
- Higher performance outcomes for hate speech detection were produced by the proposed a new hybrid LSTM-RNN deep learning approach compared to machine learning models.
- More accurate, effective, and dependable results were obtained with LSTM-RNN model, which can be simply modified to handle many additional social network and media problems.
- To identify solutions for the detection of genderism, xenophobia, and racism; a deep learning approach and five separate shallow machine learning were used instead of a single method.

In this study, while information about related studies is given in the 2nd section, the dataset is introduced in the 3rd section, detailed explanations about preprocessing are made, how the modeling is done and information about k-fold cross validation is given. In the 4th section, the experimental results are detailed through figures and tables, and in the 5th section, the results obtained from this study are given briefly.

II. Related Work

As shown in Fig 1, our focus in this study is on the issues of genderism, racism and xenophobia, which are included in the scope of 'hate speech' under the heading of 'extremism'. As part of the Online Taxonomy of Hate, the topics that are important to this study are therefore highlighted in green.

There has been some research on other related terminologies that serve relevant notions to the hate speech phenomena (e.g., cyberbullying, radicalization identification, abusive language). The analysis of these many terminologies will undoubtedly aid in gaining insights into the current situation from various viewpoints, as well as in discovering and recognizing the interrelationships between them.

Hate speech and racism are broad terms that refer to any harmful language. Hate speech falls under the category of abusive words. Profanity is also included in this vocabulary (use of inappropriate words). However, abusive language is often referred to as offensive language in studies.

Extremism

In order to identify extremism, [4] combine religious and radical characteristics with abusive and aggressive language. The authors make use of five different datasets, including the ISIS Kaggle Dataset [5] as the Radical Corpus, the ISIS-related Dataset [6] as the Neutral Corpus, the ISIS Kaggle Religious Text [7] Dataset with text from Rumiyah, the Dabiq dataset as the Religious Corpus, and a new dataset with both non-extremist and extremist tweets. The CtrlSec group, a nonprofit organization that tracks ISIS activity on Twitter, recognized the terrorist posts in the new dataset. On these datasets, the authors conduct exploratory analysis for terms related to radicalism and religion. Using TF-IDF, the authors extract features related to radicalism and religion from the Radical Corpus and the Religious Corpus, respectively. The authors divide literature into conservative and liberal. The algorithms Naive Bayes, Random Forest, and SVM are utilized for classification.

In the context of the Afghan combat zone, Sharif and his colleagues [8] discovered extremism inside Twitter networks. Extremist and non-extremist tweets were separated out by the authors. Pro-Afghanistan and pro-Taliban tweets are subcategorized as extremist tweets. The Kunduz Madrassa attack was a focus of the writers' data collection in Afghanistan [9]. The gathered tweets have been manually labeled as pro-Afghan, pro-Taliban, irrelevant, and neutral. The feature extraction employs unigrams, bigrams, and TF-IDF. Using PCA, the dimensionality is reduced.

Hate Speech

Chen and his colleagues [10] used the YouTube comments as a dataset to find offensive language. To anticipate future user behavior, they used a combination of syntactic and lexical data as well as the user's typing style.

Furthermore, Wiegand and colleagues believed that they might filter abusive phrases from negative polar expressions [11]. They used a rudimentary vocabulary to categorize a small subset of negative-pole statements and then crowdsourced the offensive words.

Xiang and his colleagues [12] presented a similar approach for detecting objectionable content on Twitter [6]. Its features are based mostly on linguistic regularities of flipped phrases and statistical subject modeling on a large dataset.

Chen et al. [13] also employed FastText as a neural network classifier to detect abusive texts on a variety of social media platforms. They discovered that FastText outperformed the Support Vector Machines (SVM) as a classifier.

Alrehili [14] did a short survey on Automatic Hate Speech Identification in social media and found that the widely used TF-IDF, Bag of Words (BoW), n-gram, Part of Speech (POS), sentiment analysis, rule-based approach, and template for automatic hate speech detection were all frequently used. emphasized eight strategies, including the most common Natural Language Processing (NLP) approaches used in the identification of hate speech are addressed and reviewed in this paper to see which ones contribute considerably to hate speech detection. They grouped the features into three categories for this purpose: linguistic pre-processing, token frequencies, and content analysis. N-gram, BoW, TF-IDF, Profanity windows, and dictionaries are examples of token frequencies, whereas linguistic preprocessing includes POS, template-based approach, rule-based method, and type dependencies. Finally, a sentiment analysis is specified as part of the content analysis. In light of all of this data, it's clear that n-gram is the most commonly utilized feature in token frequencies, and type dependencies are generally employed in linguistic preprocessing features [14].

Several attempts to classify internet abuse started by carefully examining particular types of harm. With an emphasis on cyberbullying, published a system for annotation that takes into account the presence, severity, author role (victim, harasser, or bystander), and many fine-grained categories, such as threats and insults [15].

Seven types of abuse were examined and modeled by [16] on sexist social media posts, the bulk of which go beyond physical violence against women. In a manner similar to these studies, we dissect class descriptions into fine-grained categories whenever possible in an effort to clarify possibly vague requirements. In lieu of cyberbullying, we refer to genderism, xenophobia, and racism as harmful language and hate speech.

While constructing a corpus of hate speech from Twitter data [17] investigated how the provision of associated definitions affects the reliability of the annotations. They contrasted annotations in which Twitter's definition of hate speech was offered with annotations in which no definition was provided. While annotators who were told the definition were more likely to prohibit the tweet, the authors discovered that even when Twitter's definition was presented, inter-annotator agreement, as evaluated by Krippendorf's alpha, was at most 0.3, depending on the question addressed (For annotations to be regarded reliable, Krippendorff (2004) specifies a minimum score of 0.80, with 0.667 being the lowest possible limit.). Ross et al. found that more specific coding techniques are required to differentiate hate speech from other types of information.

The community of natural language processing has mostly concentrated on detecting hate speech and cyberbullying [18]. As a result, a number of research datasets have been created [19], but none of them have used the same terminology or marked only partial phenomena (e.g. annotating sexist and racist speech, but not hate speech directed to all groups that require protection).

When creating hate speech datasets, Davidson et al. also note the challenge of reaching high rates of interannotator agreement. They found that only 1.3 percent of tweets were consistently identified as having hate speech, while 5 percent of tweets were classified as such by the majority of annotators [20]. The 2018 Kaggle Toxic Comment Classification Challenge creators state that although the dataset was created with ten annotators per label, agreement was poor (Krippendorff alpha of 0.45) [21].

Racism/Xenophobia

In order to detect racism using deep learning and text mining techniques on Twitter for Arabic, Slotaibi and Hasanat employed a Convolutional Neural Network (CNN) and a nature inspired optimization method for classification. The data was split 70-15-15 percent for training, test, and validation sets, accordingly. As a result of this research, it has been discovered that deep learning architectures perform better than intelligent suopervised learning models in detecting racism in Arabic, and that adapting a model that takes into account the Arabic language complexity compensates for the lack of Arab cyberracism detection [22].

Lee and his colleagues [23] attempted to detect racist language in Tweets using sentiment analysis. Gated Recurrent Unit (GRU), CNN, and Group Constrained-Neural Network (GCR-NN) recursive neural networks RNN have been integrated to produce a stacked ensemble deep learning architecture, thanks to deep learning's improved performance. To extract relevant and conspicuous characteristics from raw text, the GRU outperforms the GCR-NN model, and CNN has extracted key features for RNN to generate correct predictions. The proposed GCR-NN model was successful in detecting 97 percent of racist tweets [23].

Genderism

Park & Fung [24] used the dataset provided by Waseem and Hovy [33] to compare the performances of one-step and two-step classifiers in a deep learning scenario to detect genderism and hate speech. They concluded that combining two classifiers (for example, Convolutional Neural Network and Logistic Regression) could improve performance.

While Istaiteh and his colleagues [25] examined the five most commonly used datasets in their study of racist and sexist hate speech, an overview of the most commonly used feature representation methodologies is offered, including n-gram, word-gram, word embedding, Term Frequency-Inverse Document Frequency (TF-IDF), and deep learning. Deep learning-based approaches for feature extraction produce the best results, while word embedding feature representations such as GloVe and Word2vec produce competitive results in hate speech. Classic feature representations such as character n-grams and word n-grams produce outdated and unsuccessful results.

Researchers in [26] combined genderism-related lexicons with 1-3 word n-grams and 1-7 char n-grams.

Using one of the genderism datasets, the SVM classifier was used to categorize the tweets with those attributes, earning an F1 score of 89 percent.

Only sexist hate speech was the focus of data sets [27], [28], and [29]. Three datasets were used by the authors of paper [26], and their work is described in [27], [28], and [29]. The best results were obtained on the datasets and using char n-grams as features and SVM as the classification method, which had accuracy rates of 78.77 percent and 75.44 percent, respectively. With an accuracy of 89.32 percent on dataset [29], the BoW and sequences of words feature with SVM as the classification method offered the greatest performance. Additionally, writers in [30] combined sentence embeddings, BoWV, and TF-IDF, however their outcome was noticeably worse to the model offered by [26] on the same dataset.

Last but not least, [31] worked on their dataset, using bag and word sequences as features and NB as the classification technique, reaching 76 percent accuracy.

When all of these studies are analyzed, it is observed that no LSTM-RNN deep learning studies are used for the problem of racism, xenophobia and, genderism detection in online social network and media. By categorizing three classes of racism, xenophobia, and genderism in this study, the lack in this respect was filled, and rather satisfactory results were reached despite the instability of the data set.

III. Methodology

Genderism, racism, and xenophobia detection was dealt with as a classification issue. For the first time, a new hybrid Long-Short Term Memory-Recurrent Neural Network (LSTM-RNN) deep learning model was adopted for hate speech detection, one of the most common social media analysis problems. These adapted models are compared with the most popular classification models such as Random Forest (RF), SVM, Logistic Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (K-NN).

III.I. Dataset

This research aims to identify racism, xenophobia, and genderism in tweets. To put it another way, tweets are labeled as racist or sexist in order to identify information that contains hate speech.

The dataset named Hate Speech and Offensive Language Dataset published by Andrii Samoshyn on Kaggle was used for the data set [34]. Our aim is to predict the labels in the test dataset using a training set in which the label "1" indicates that the tweet includes racism/genderism and the label "0" indicates that the tweet does not include racism/genderism. The training set for this study has 31,951 labeled data, while the test set contains 17,197 data.

III.II. Preprocessing

There are numerous data preparation methods. One of them is data cleaning, which is an application for removing noise and correcting errors in data [32]. The first stage in data preparation is to deal with missing or erroneous data in the dataset. It was first assessed whether there were any missing values, which there were none. Fig 2 depicts the machine learning process diagram. Some advantages of data preparation include the capacity to execute pertinent data analyses, understanding the nature of the data, and extracting valuable information from the dataset [32]. These issues can preclude any type of analysis from being performed on the data.

Tokenization was performed to sentences and words in the second step of data preprocessing, and upper- and lower-case letters were separated. In addition, by analyzing 2000 words in the texts, the tokenizer class was updated according to twitter values, and special characters and numerical values were removed from the data.

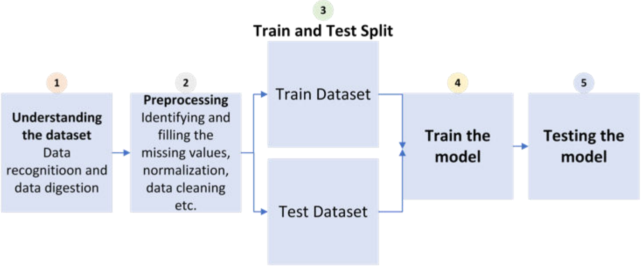
3
Train and Test Split
1
Understanding the dataset
Data recognitioon and data digestion
2
Preprocessing
Identifying and filling the missing values, normalization, data cleaning etc.
Train Dataset
Test Dataset
4
Train the model
5
Testing the model

Fig 2. Machine Learning Process Diagram

As shown in Fig 3, LSTM-based deep RNN architecture begins with dataset pre-processing, data cleaning, and tokenization processes. The LSTM-RNN model is then used to produce prediction and performance measures after feature selection. The architecture is then completed after a performance analysis study.

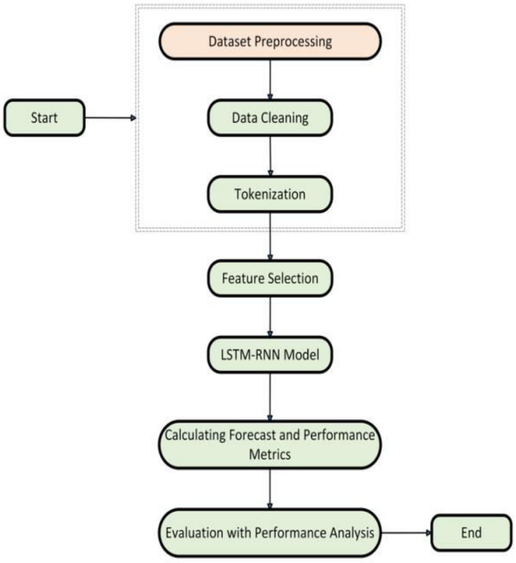

Fig 3. Flowchart of LSTM Based Deep RNN Architecture

Fig 4 depicts a general view of working in the framework. There are two types of text data mining approaches. The first is data preprocessing, which includes feature extraction using Natural Language Processing Techniques (TF-IDF), which are commonly employed by standard machine learning systems to encode words for fundamental procedures (SVM, LR, etc.). The second one uses Word Embedding methods. The proposed model is developed by deep learning classifiers after data preprocessing and feature extraction using Word Embedding.

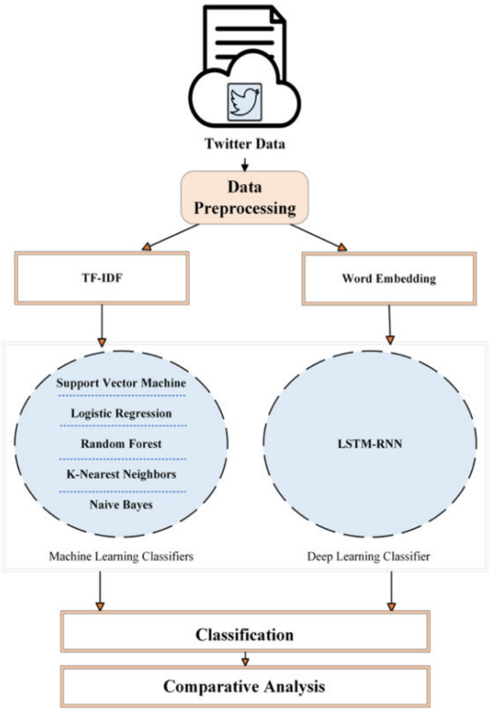

Fig 4. Racism, Xenophobia, and Genderism Detection Framework

IV. Experimental Results

Dense layer was utilized as a hidden layer, sigmoid activation function was employed as a neuron, and LSTM architectures were used as n-gram/word embedding/Content Encoder. The Embedding Layer was chosen with vector length 128, 128 neurons in each hidden layer, batch size 32, and epoch number 13. For all task categorization, the ADAM optimizer was used with Keras' default parameters. Training and test scores indicated in the tables were calculated according to the model's accuracy metric.

Table 1. Logistic Regression Results

LR	Recall	Precision	F1 Score
0	0.93	1.00	0.96
1	0.00	0.00	0.00
Macro avg	0.46	0.50	0.48
Weighted avg	0.86	0.93	0.89
F1 Score	0.0		
Train Score	0.9301390068386001		
Test Score	0.9283554072374596		
Accuracy	**0.93**		

Table 2. Random Forest Results

RF	Recall	Precision	F1 Score
0	1.00	0.94	0.97
1	0.24	0.99	0.39
Macro avg	0.62	0.97	0.68
Weighted avg	0.95	0.95	0.93
F1 Score	0.38777908343125733		
Train Score	0.9996424261386493		
Test Score	0.9448326207112316		
Accuracy	**0.95**		

Table 3. K-Nearest Neighbors Results

K-NN	Recall	Precision	F1 Score
0	0.99	0.94	0.96
1	0.13	0.59	0.22
Macro avg	0.56	0.77	0.59
Weighted avg	0.93	0.91	0.91
F1 Score	0.21930870083432655		
Train Score	0.9379609350556475		
Test Score	0.9316925643967046		
Accuracy	**0.93**		

Table 4. Support Vector Machine Results

SVM	Recall	Precision	F1 Score
0	1.00	0.93	0.96
1	0.02	1.00	0.05
Macro avg	0.51	0.97	0.50
Weighted avg	0.93	0.94	0.90
F1 Score	0.045714285714285714		
Train Score	0.9317927859473473		
Test Score	0.9303368443007612		
Accuracy	**0.93**		

Table 5. Naïve Bayes Results

NB	Recall	Precision	F1 Score
0	0.01	0.96	0.03
1	0.99	0.07	0.05
Macro avg	0.50	0.52	0.08
Weighted avg	0.08	0.90	0.03
F1 Score	0.1338590438639724		
Train Score	0.08219729137800026		
Test Score	0.08363750130357701		
Accuracy	**0.08**		

The accompanying tables show the findings of the classification models employed in this paper. 30% of the dataset was used for testing, while 70% was used for training. With 128 layers, batch size 35 was chosen and lstm_out value 64 was chosen. LR Results for 3000-word size as shown in Table 1, the accuracy rate was 93%, while the F1 Score was 0%. When precision, recall, and F1 Score measures are investigated, the RF model has the greatest values among other machine learning models, as shown in Table 2. At the same time, this model produced training and test score rates of 99.96 percent and 94.48 percent, respectively. The accuracy rate in the K-NN Results was 93 percent, and the F1 Score was 21.93 percent; these numbers are shown in Table 3. The accuracy rate of SVM Results was lower than RF, and it was found to be 93 percent, similar to other models. Table 4 shows SVM findings, which have a lower recall rate than all models except NB, and Table 5 shows NB results, which have lowered all metrics including accuracy rate than other models.

When looking at the LSTM Neural Network Results, Table 6 shows that while the F1 Score is lower than the machine learning classification models, the accuracy, precision, and recall rates are significantly higher, and the LSTM Neural Network is the model that offers the best results. Despite the little amount of data identified as hate speech, the algorithm is nevertheless able to accurately predict that the text does not include hate speech. In this way, the F1 Score has become a metric for comparing models, and it takes precision and recall into calculation.
While the accuracy graph of the models given in Fig 6 is given, the error graph is shown in Fig 7 and the test and training information is given according to 13 epochs in this graph. In Fig 8, training and test data are visualized according to the epoch number by giving the graph of the F1 Score of the model. Fig 8 depicts the accuracy, error, and F1 Score bar graphs for all classification models utilized.

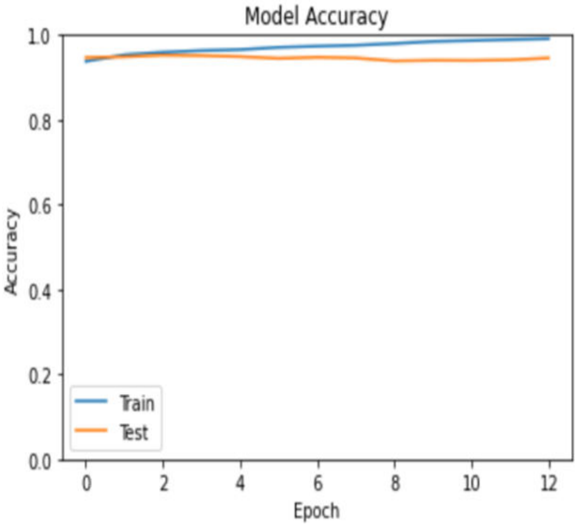

Fig 6. Model Accuracy Graph

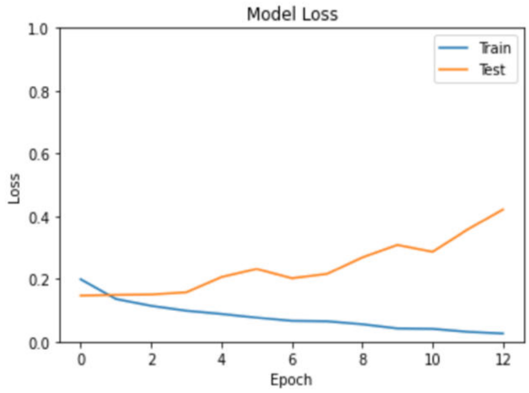

Fig 7. Model Loss Graph

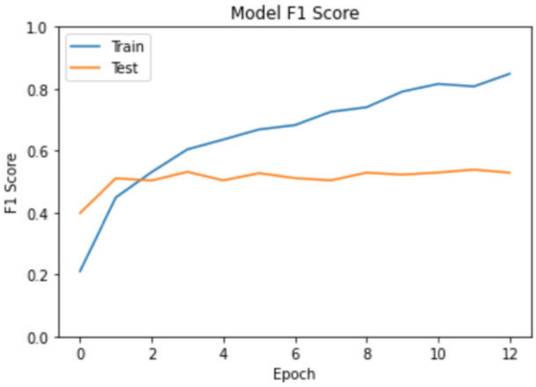

Fig 8. Model F1 Score Graph

Table 6. LSTM Neural Network Results

LSTM Neural Network	
F1 Score	0.5132381916046143
Accuracy	0.9520333409309387
Precision	0.6095554828643799
Recall	0.4894445240497589

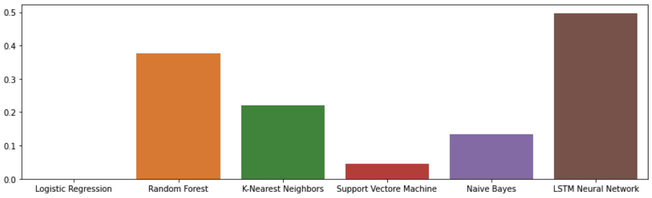

Fig 9. Column Chart of Models

Furthermore, it is intended in this study to compare the outcomes by modifying specific settings in order to see if different results and more accurate results can be produced. Testing took up 20% of the dataset, while training took up the other 80%. The batch size is 32, and the lstm_out value is 16, with 64 layers. LR When the training score in the study with 30 percent test data was compared to the training score in the study with 20 percent test data, the training score in the study with 30 percent test data offered a higher outcome, while it gave a greater percentage in the test score. Furthermore, based on the data in Table 7, it was discovered that all measures produced the same findings in both test rates. The training score in the study where the test rate was given as 20% in the RF Results was better than the 30% test rate, while the recall and F1 values, excluding precision, in the data labeled as racist in the metrics were higher than the study where the test rate was given as 30%, as shown in Table 8. All metrics in the data categorized as racist and xenophobic were higher in the study employing 20% test data, according to the K-NN Results, as shown in Table 9. The metrics in the results of the two test ratios were the same in the SVM results in Table 10, and the F1 Score metric, the results of the NB model, which delivered higher results in the 20% test data, are reported in Table 11. The accuracy rate was greater in the model employing 30% test data, although the recall rate was substantially lower, as shown in Table 12.

Table 7. Logistic Regression Model Results Using 20% Test Data

LR	Recall	Precision	F1 Score
0	1.00	0.93	0.96
1	0.00	0.00	0.00
Macro avg	0.50	0.46	0.48
Weighted avg	0.93	0.86	0.89
F1 Score	0.0		
Train Score	0.9305408893582072		
Test Score	0.9260128265290161		
Accuracy	**0.93**		

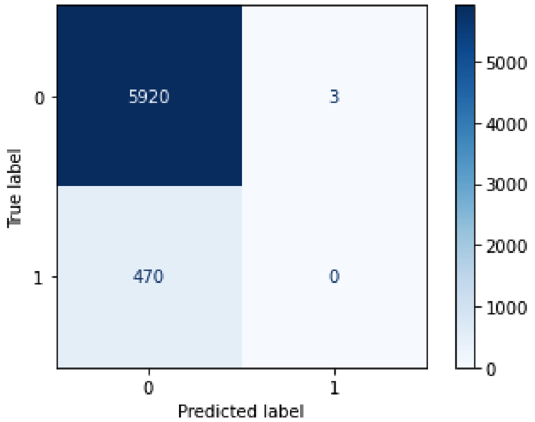

Fig 10. Logistic Regression Model Confusion Matrix Using 20% Test Data

Based on the accuracy results of all models, the column chart of the models using 20% test data in the study is given in Fig 9. In the figures, the confusion matrix of the models employed in this study is also shown. When the estimated and actual negative results are studied, the rate is quite high, but the positive value is 0. When the confusion matrix of the LR model is evaluated in Fig 10, it is apparent that the rate is quite high when the estimated and actual negative results are examined. The number of inaccurate predictions between the classes in the RF model's confusion matrix is substantially higher than the number of correct predictions, as seen in Fig 11. Furthermore, while looking at the K-NN model's confusion matrix (Fig 12), it's worth noting that the accurately predicted numbers of the classes labeled with 1 are relatively high, but the ratio of those labeled with 0 is low. Fig 13 and Fig 14 demonstrate the confusion matrix of SVM and NB, respectively.

Table 8. Random Forest Model Results Using 20% Test Data

RF	Recall	Precision	F1 Score
0	1.00	0.94	0.97
1	0.26	0.98	0.41
Macro avg	0.63	0.96	0.69
Weighted avg	0.95	0.95	0.93
F1 Score	0.4134453781512605		
Train Score	0.9995306816848527		
Test Score	0.9454090411387455		
Accuracy	**0.95**		

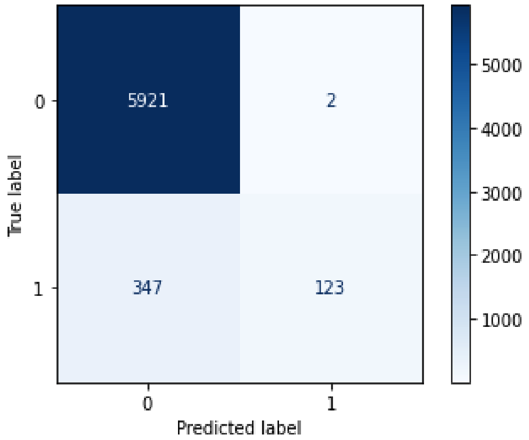

Fig 11. Random Forest Model Confusion Matrix Using 20% Test Data

Table 9. K-Nearest Neighbors Model Results Using 20% Test Data

K-NN	Recall	Precision	F1 Score
0	0.99	0.94	0.96
1	0.14	0.61	0.23
Macro avg	0.57	0.77	0.59
Weighted avg	0.93	0.91	0.91
F1 Score	0.22530329289428078		
Train Score	0.9393406077672181		
Test Score	0.9300797747536368		
Accuracy	**0.93**		

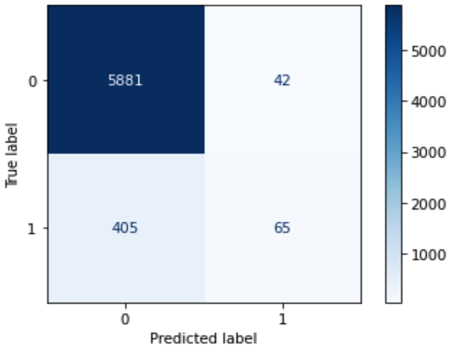

Fig 12. K-Nearest Neighbors Model Confusion Matrix Using 20% Test Data

Table 10. Support Vector Machine Model Results Using 20% Test Data

SVM	Recall	Precision	F1 Score
0	1.00	0.93	0.96
1	0.03	1.00	0.05
Macro avg	0.51	0.96	0.51
Weighted avg	0.93	0.93	0.90
F1 Score	0.05383022774327121		
Train Score	0.9320661738824357		
Test Score	0.9285155638980135		
Accuracy	**0.93**		

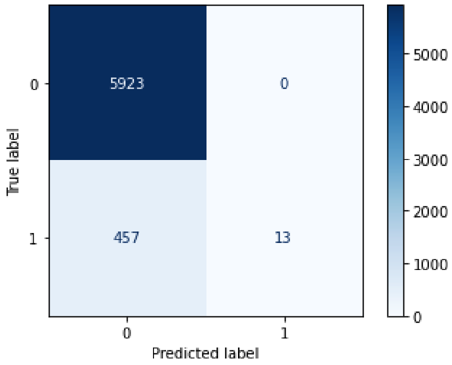

Fig 13. Support Vector Machine Model Confusion Matrix Using 20% Test Data

Table 11. Naïve Bayes Model Results Using 20% Test Data

NB	Recall	Precision	F1 Score
0	0.01	0.97	0.96
1	1.00	0.07	0.05
Macro avg	0.50	0.52	0.51
Weighted avg	0.08	0.91	0.90
F1 Score	0.137870083959346		
Train Score	0.07884547694473777		
Test Score	0.08446738620366025		
Accuracy	**0.93**		

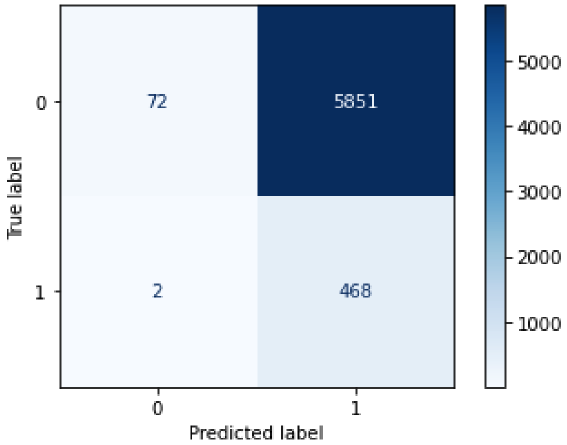

Fig 14. Naïve Bayes Model Confusion Matrix Using 20% Test Data

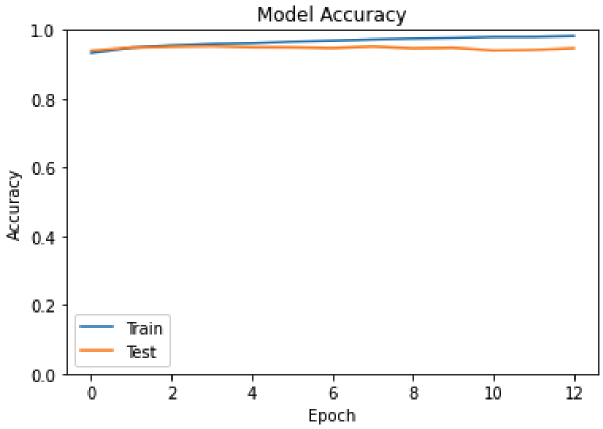

Fig 15. Accuracy Graph of Models Using 20% Test Data

In the sample study using 20% test data, Accuracy, Error and F1 Score graphs are shown in Fig 15, Fig 16 and Fig 17 respectively. Based on the accuracy results of all models, the column chart of the models using 20% test data in the study is given in Fig 18.

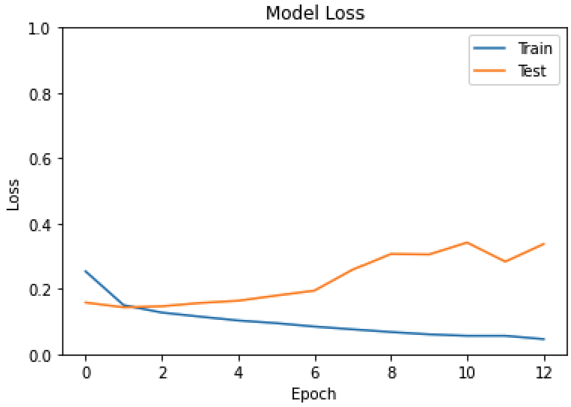

Fig 16. Loss Graph of Models Using 20% Test Data

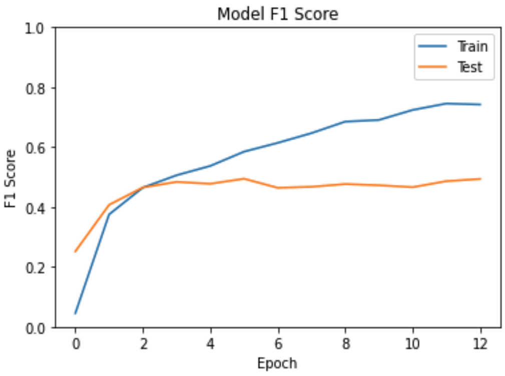

Fig 17. F1 Score Graph of Models Using 20% Test Data

Table 12. LSTM Neural Network Results Using 20% Test Data

LSTM Neural Network	
F1 Score	0.5132381916046143
Accuracy	0.9520333409309387
Precision	0.6095554828643799
Recall	0.4894445240497589

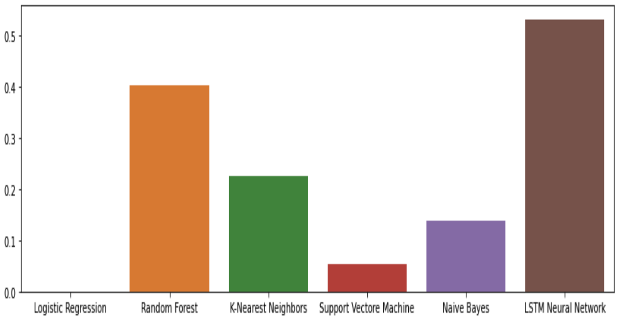

Fig 18. Bar Chart of Models Using 20% Test Data

The classification models used in the study were tested using k-fold cross validation with a k number of 5. Table 13 shows the k-fold cross validation results of the LR, RF, K-NN, SVM, and NB models based on this selection.

The LSTM Neural Network model's accuracy rate was 94.97 percent, the recall value was 52.65 percent, the precision rate was 60.90 percent, and the F1 Score was 53.06 percent as a result of these improvements, as shown in Table 14.

The LSTM Neural Network model with an accuracy rate of 95.20 percent, a recall value of 48.94 percent, a precision of 60.95 percent, and an F1 Score of 51.32 percent was deemed to have the greatest performance after examining these classification models and the outcomes of the neural network. In addition, the selected test data rate of 30% was adjusted to 20%, and the batch size, lstm_out, and number of layers were changed to see whether different and more accurate findings could be found.

When the accuracy rates were compared, the sample with 30% test data produced a superior result, but the F1 Score value, recall, and precision rates were greater in the sample with 20% test data.

Table 13. 5-Fold Cross Validation Results

Model	5-Fold Cross Validation
Logistic Regression	93.032%
Random Forest	93.12%
K-Nearest Neighbors	93.16%
Support Vector Machine	93.10%
Naive Bayes	94.10%

Table 14. Results of Models

			Test Percentage	
			20%	30%
Models	LR	F1 Score	0.0	0.0
		Accuracy	0.93	0.93
	RF	F1 Score	0.41	0.38
		Accuracy	0.95	0.95
	K-NN	F1 Score	0.22	0.21
		Accuracy	0.93	0.93
	SVM	F1 Score	0.05	0.04
		Accuracy	0.93	0.93
	NB	F1 Score	0.13	0.13
		Accuracy	0.08	0.08
	LSTM-RNN	F1 Score	0.53	0.51
		Accuracy	0.949	0.952

V. Conclusion

This research compensates for the lack of identification of racism, xenophobia, and genderism in English tweets. On a dataset encompassing racism, hate speech, and genderism, comparisons were done between several models such as SVM, LR, RF, K-NN, and NB. A model based on LSTM Neural Network and RNN was constructed, and the results were compared using bespoke metrics.

Given the low rate of hate speech in the content of the dataset utilized in this investigation, it is expected that more accurate and reliable findings would be obtained in future studies using alternative deep learning models and a more balanced dataset.

As a result of the use of 20% and 30% test data in the study, the LSTM Neural Network gave a much higher accuracy rate than other models. In addition, it is aimed to find new results by changing the batch size, number of layers and lstm_out values in order to measure whether different and more accurate findings can be found.

When the classification models are examined, it is seen that the RF model gives much higher results. The lack of balanced data in the dataset was seen as the reason why the models did not have significantly greater accuracy rates. For this reason, new studies can be performed to increase the accuracy rates with a balanced data set.

References

[1] B. B. Gupta, S. R. Sahoo, *Online social networks security: principles, algorithm, applications, and perspectives* (CRC Press, 2021) .

[2] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta ve V. ... & Pirrelli, The PAISA Corpus of Italianweb Texts, *9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden, 2014.

[3] S. De, et al. An introduction to data mining in social networks, *Advanced Data Mining Tools and Methods for Social Computing*. Academic Press, 2022. 1-25.

[4] R. Z. Ul, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz, M. Hanif & M. A. Saeed, M. A. Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning, 2021.

[5] FifthTribe. (2015). How ISIS uses Twitter. Accessed: May, 10, 2022. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[6] ActiveGalaxy, Kaggle. (2016). ISIS Related Dataset. Accessed: May, 10, 2022. [Online]. Available: https://www.kaggle.com/datasets/activegalaxy/isis-related-tweets

[7] FifthTribe, Kaggle. (2017). ISIS Religious Text. Accessed: May, 10, 2022. [Online]. Available: https://www.kaggle.com/datasets/fifthtribe/isis-religious-texts

[8] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain & G. S. Choi, An empirical approach for extreme behavior identification through tweets using machine learning. *Applied Sciences*, *9*(18), 2019, 3723.

[9] T. Ruttig, Kunduz Madrassa Attack Al Jazeera, 2018, Accessed: May. 10, 2022. [Online], Available: https://www.aljazeera.com/opinions/2018/4/5/kunduz-madrassa-attack-losing-the-moral-high-ground.

[10] Y. Chen, Y. Zhou, S. Zhu and H. Xu, Detecting offensive language in social media to protect adolescent online safety, *Proceedings,* 2012, 71-80.

[11] M. Wiegand, J. Ruppenhofer, A. Schmidt and C. Greenberg, Inducing a Lexicon of Abusive Words – a Feature-Based Approach, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, 2018.

[12] G. Xiang, B. Fan, L. Wang, J. Hong and C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, *Proc. 21st ACM Int. Conf. Inf. Knowl. Manag.- CIKM'12*, 2012.

[13] H. Chen, S. McKeever, & S. J. Delany, Abusive Text Detection Using Neural Networks, In *AICS*, (December 2017) 258-260.

[14] A. Alrehili, Automatic hate speech detection on social media: A brief survey, In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, (2019, November), 1-6.

[15] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, ... & V. Hoste, Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing,* (2015, September), 672-680.

[16] M. Anzovino, E. Fersini & P. Rosso, Automatic identification and classification of misogynistic language on twitter, In *International Conference on Applications of Natural Language to Information Systems*, Springer, Cham, (2018, June), 57-64.

[17] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky & M. Wojatzki, Measuring the reliability of hate speech annotations: The case of the european refugee crisis, *arXiv preprint arXiv:1701.08118,* 2017.

[18] A. Schmidt, & M. Wiegand, A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, Association for Computational Linguistics, (2019, January), 1-10.

[19] B. Vidgen & L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, *Plos one*, *15*(12), e0243300, 2020.

[20] Z. Waseem, T. Davidson, D. Warmsley & I. Weber, Understanding abuse: A typology of abusive language detection subtasks, *arXiv preprint arXiv:1705.09899,* 2017.

[21] E. Wulczyn, N. Thain & L. Dixon, Ex machina: Personal attacks seen at scale, CoRR abs/1610.08914, 2016.

[22] A. Alotaibi& M. H. A. Hasanat, Racism detection in Twitter using deep learning and text mining techniques for the Arabic language, In *2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTEC),* IEEE, (2020, November), 161-164.

[23] E. Lee, F. Rustam, P. B. Washington, F. El Barakaz, W. Aljedaani & I. Ashraf, Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. *IEEE Access*, *10*, 2022, 9717-9728.

[24] J. H. Park ve P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter", *AICS Conference*, 2017.

[25] O. Istaiteh, R. Al-Omoush & S. Tedmori, October). Racist and sexist hate speech detection: Literature review, In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* IEEE, 2020, 95-99.

[26] S. Frenda, B. Ghanem, M. Montes-y-Gómez & P. Rosso, Online hate speech against women: Automatic identification of misogyny and sexism on twitter, *Journal of Intelligent & Fuzzy Systems*, *36*(5), 2019, 4743-4752.

[27] E. Fersini, D. Nozza & P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), *EVALITA Evaluation of NLP and Speech Tools for Italian*, *12*, 2018, 59.

[28] E. Fersini, P. Rosso M. & Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, *Ibereval@ sepln*, *2150*, 2018, 214-228.

[29] J. Andreas, E. Choi & A. Lazaridou, Proceedings of the naacl student research workshop. In *Proceedings of the NAACL Student Research Workshop,* (2016, June).

[30] P. Saha, B. Mathew, P. Goyal & A. Mukherjee, Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700, 2018*.

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

[32] L. Hickman, S. Thapa, L. Tay, M. Cao & P. Srinivasan, Text preprocessing for text mining in organizational research: Review and recommendations, *Organizational Research Methods*, *25*(1), 2022, 114-146.

[33] Z. Waseem and D. Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, *Proc. NAACL Student Res. Work, 2016,* 88-93.

[34] "Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset.

BIOGRAPHIES AND PHOTOGRAPHS

Şule Kaya

Şule Kaya is currently working as a research assistant at the Department of Software Engineering at Firat University. She is receiving a Master's degree in Software Engineering. She received her bachelor's degree in software engineering from Firat University, Turkey, in 2020.

Bilal Alatas

Prof. Dr. Bilal Alatas received his B.S. and M.S. degrees in Computer Engineering from Firat University in 2001 and 2003, respectively. He received Ph.D. degree from Firat University in 2007. Currently, he is head of the Software Engineering Department at Firat University in Elazig, Turkey and works as a Professor of Software Engineering at this department. He served as the chair of department of computer engineering at Munzur University during 2010-2014. He is the founder head of the Computer Engineering Department of Munzur University and Software Engineering Department of Firat University. His research interests include artificial intelligence, data mining, social network analysis, metaheuristic optimization, and machine learning. Dr. Alatas has published over 150 papers in many well-known international journals, proceedings of the refereed conferences, and books since 2001. Over 400 citations of his works have been reported in the Google Scholar.