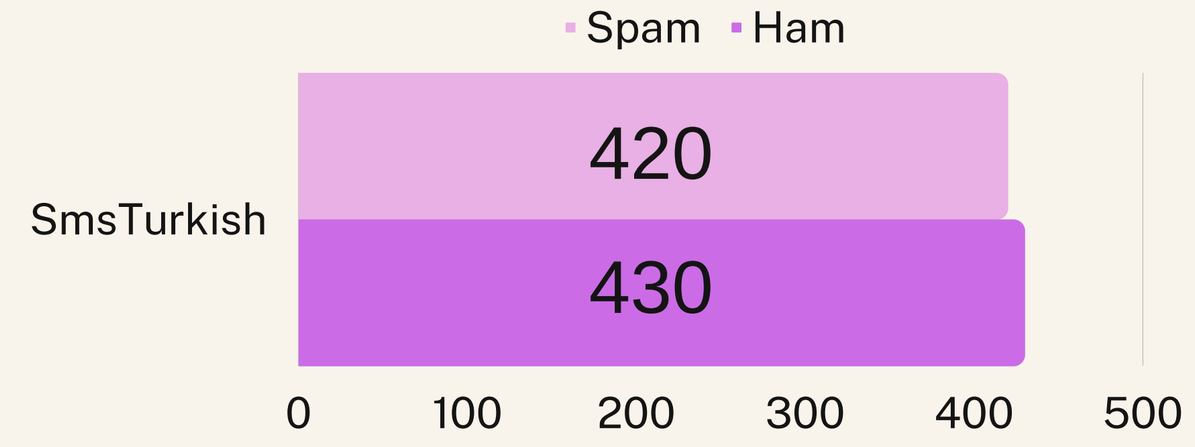
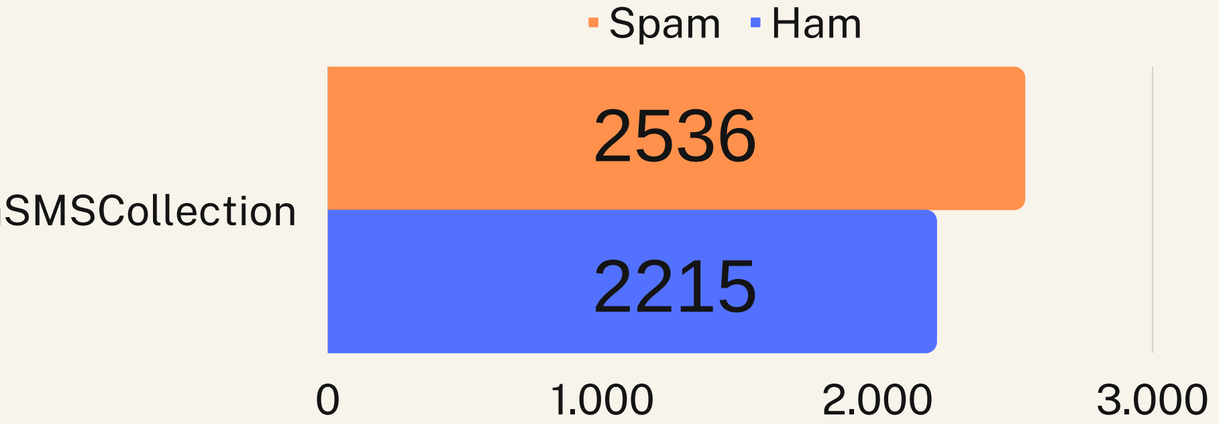
SPAM DETECTION FOR TURKISH SHORT MESSAGE SERVICE

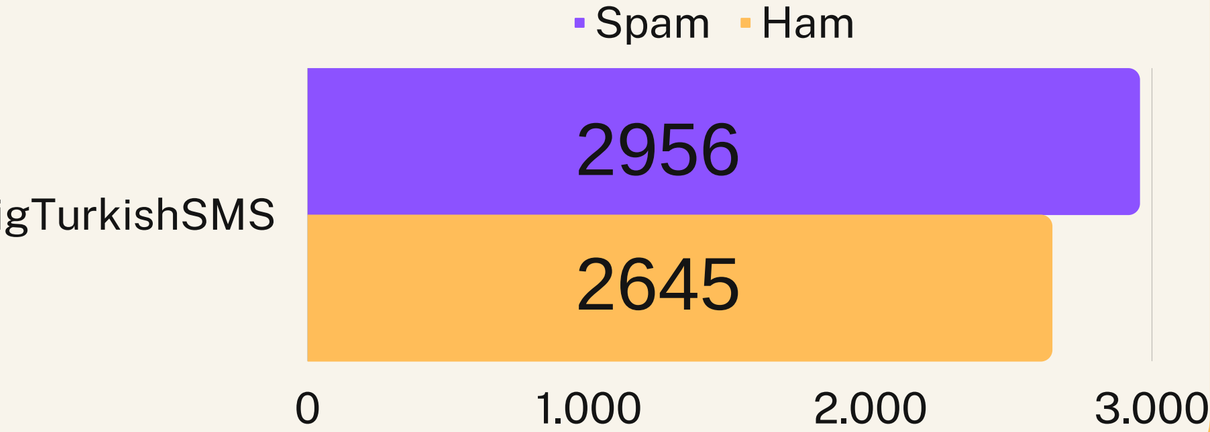
Zeynep Çakmak 1700003250
Mehmet Salih Çifçi 1800004594

INTRODUCTION

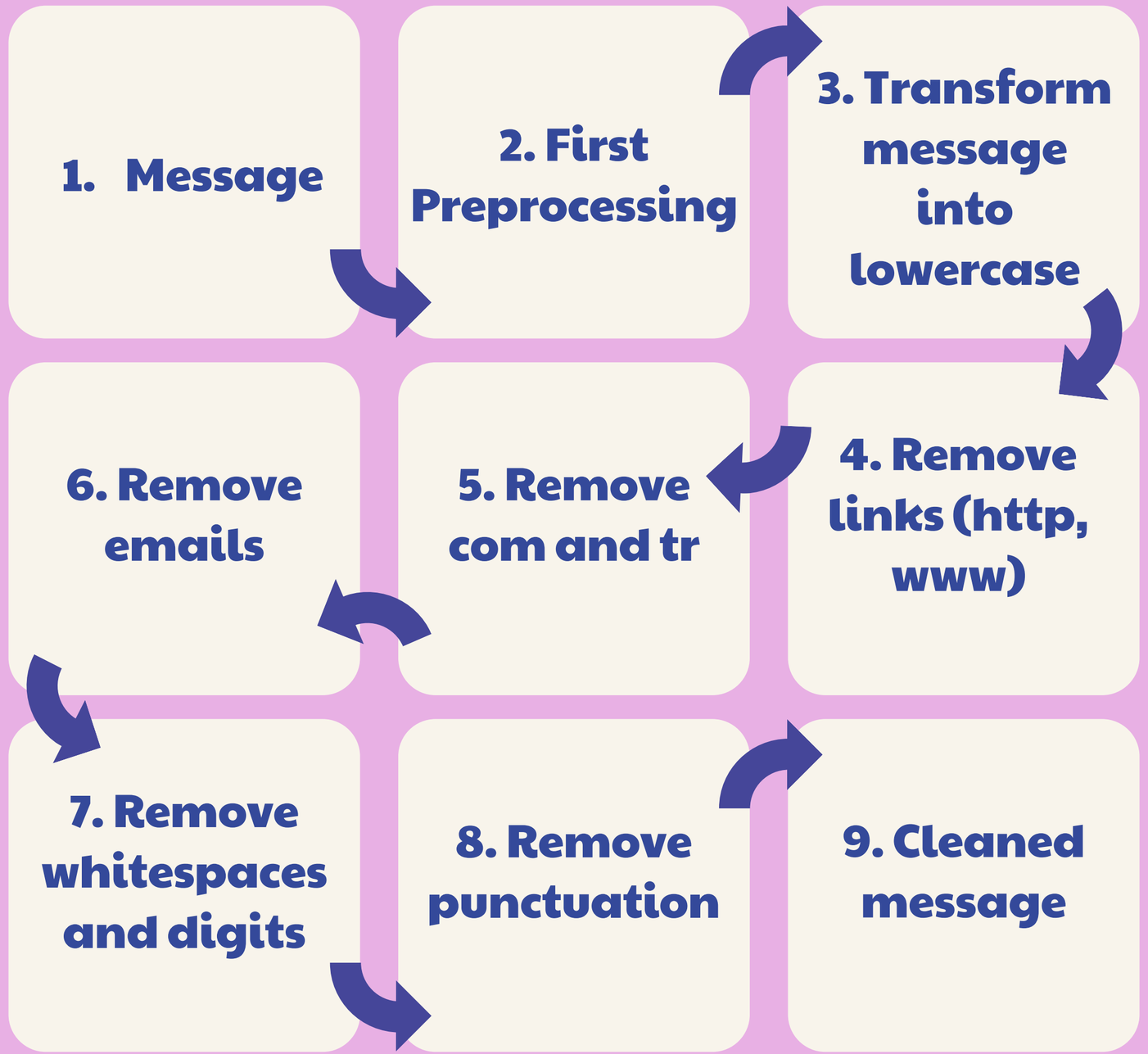
In this study, the classification of messages was carried out using machine and deep learning algorithms on a dataset consisting of Turkish SMS.







Data Preprocessing : First Preprocessing



Second Preprocessing

CROSS VALIDATION

Cross validation is statistical resampling method used to measure the productivity of the ML model on data it does not see, as objectively and accurately as possible.

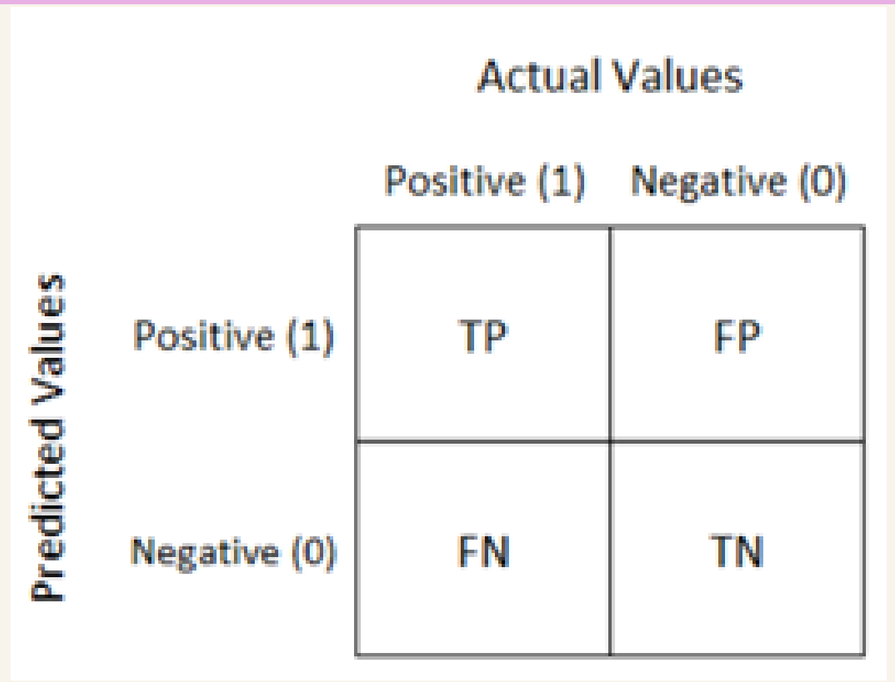
Procedure takes a one variable called k. This shows the count of groups that need to divide a significant data instance. For this reason, this method is often cited as K fold cross validation.

Cross validation scores for all algorithms (Smsturkish)

	cv = 3	cv=5	cv=10
KNN	0.5888	0.6426	0.6981
RandomForest	0.9672	0.9603	0.9726
Logistic Regression	0.9656	0.9716	0.9722
Multinomial Naive Bayes	0.9548	0.9545	0.9676
Support Vector Machine	0.9661	0.9710	0.9770
Decision Tree Classifier	0.9156	0.9147	0.9172
Ada Boost Classifier	0.9561	0.9589	0.9545
XGBoost Classifier	0.9406	0.9445	0.9483
Stochastic Gradient Descent	0.9396	0.9441	0.9649
ANN	0.9543	0.9762	0.9941
CNN	0.9078	0.9385	0.9588
LSTM	0.4983	0.5189	0.5353

Cross validation scores for all algorithms (SmsEnglish)

	cv = 3	cv=5	cv=10
KNN	0.8024	0.8277	0.8315
RandomForest	0.9395	0.9459	0.9479
Logistic Regression	0.9424	0.9480	0.9592
Multinomial Naive Bayes	0.8909	0.8841	0.8825
Support Vector Machine	0.9347	0.9518	0.9548
Decision Tree Classifier	0.7949	0.7893	0.8070
Ada Boost Classifier	0.8846	0.9003	0.8944
XGBoost Classifier	0.9324	0.9339	0.9448
Stochastic Gradient Descent	0.9389	0.9457	0.9475
ANN	0.9448	0.9683	0.9776
CNN	0.9378	0.9591	0.9767
LSTM	0.4918	0.5029	0.5348



TP: If prediction is positive and it’s true.
TN: If prediction is negative and it’s true.
FP: If prediction is positive and it’s false.
FN: If prediction is negative and it’s false.

CONFUSION MATRIX

Confusion matrix is a productivity indicator for ML classification difficulty, and the output can be in further than one class. It is a table of four distinct combining estimated and real values.

Accuracy: From all the classes , how many of them we have predicted correctly.

Precision: From all the classes we have predicted as positive, how many are actually positive.

Recall: From all the positive classes, how many we predicted correctly.

F1 Score: F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recal = \frac{TP}{TP + TN}$$

$$F1\ Score = \frac{Precision * Recall}{Precision + Recall} * 2$$

EXPERIMENTAL RESULTS

Smsturkish

FIRST PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.7353	0.7385	0.7353	0.7338
RandomForest	0.9412	0.9422	0.9412	0.9412
Logistic Regression	0.9412	0.9422	0.9412	0.9412
Multinomial Naive Bayes	0.9412	0.9472	0.9412	0.9409
Support Vector Machine	0.9471	0.9477	0.9471	0.9471
Decision Tree Classifier	0.8588	0.8762	0.8588	0.8576
Ada Boost Classifier	0.9471	0.9487	0.9471	0.9471
XGBoost Classifier	0.9235	0.9266	0.9235	0.9235
Stochastic Gradient Descent	0.9647	0.9649	0.9647	0.9647

SECOND PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.8941	0.8999	0.8941	0.8938
RandomForest	0.9470	0.9475	0.9470	0.9470
Logistic Regression	0.9294	0.9304	0.9294	0.9294
Multinomial Naive Bayes	0.9470	0.9475	0.9470	0.9470
Support Vector Machine	0.9411	0.9413	0.9411	0.9411
Decision Tree Classifier	0.8705	0.8813	0.8705	0.8699
Ada Boost Classifier	0.9352	0.9353	0.9352	0.9353
XGBoost Classifier	0.9352	0.9358	0.9352	0.9353
Stochastic Gradient Descent	0.9235	0.9236	0.9235	0.9235

TurkishSMSCollection

FIRST PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.6635	0.7474	0.6635	0.6101
RandomForest	0.9284	0.9322	0.9284	0.9278
Logistic Regression	0.9747	0.9757	0.9747	0.9748
Multinomial Naive Bayes	0.9747	0.9756	0.9747	0.9746
Support Vector Machine	0.9831	0.9837	0.9831	0.9832
Decision Tree Classifier	0.9158	0.9196	0.9158	0.9161
Ada Boost Classifier	0.9663	0.9664	0.9663	0.9663
XGBoost Classifier	0.9674	0.9678	0.9674	0.9674
Stochastic Gradient Descent	0.9863	0.9867	0.9863	0.9863

SECOND PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.9652	0.9667	0.9652	0.9653
RandomForest	0.9821	0.9821	0.9821	0.9821
Logistic Regression	0.9737	0.9740	0.9737	0.9737
Multinomial Naive Bayes	0.6624	0.7640	0.6624	0.6039
Support Vector Machine	0.9821	0.9821	0.9821	0.9821
Decision Tree Classifier	0.9463	0.9495	0.9463	0.9465
Ada Boost Classifier	0.9842	0.9842	0.9842	0.9842
XGBoost Classifier	0.9779	0.9779	0.9779	0.9779
Stochastic Gradient Descent	0.9810	0.9810	0.9810	0.9810

BigTurkishSms

FIRST PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.6975	0.7482	0.6975	0.6726
RandomForest	0.9393	0.9422	0.9393	0.9390
Logistic Regression	0.9803	0.9807	0.9803	0.9803
Multinomial Naive Bayes	0.9741	0.9751	0.9741	0.9740
Support Vector Machine	0.9857	0.9859	0.9857	0.9857
Decision Tree Classifier	0.9072	0.9129	0.9072	0.9073
Ada Boost Classifier	0.9634	0.9643	0.9634	0.9634
XGBoost Classifier	0.9607	0.9611	0.9607	0.9607
Stochastic Gradient Descent	0.9848	0.9849	0.9848	0.9848

SECOND PREPROCESSING	Accuracy	Precision Score	Recall Score	F1 Score
KNN	0.9536	0.9575	0.9536	0.9536
RandomForest	0.9759	0.9761	0.9759	0.9759
Logistic Regression	0.9661	0.9673	0.9661	0.9661
Multinomial Naive Bayes	0.6440	0.7630	0.6440	0.5830
Support Vector Machine	0.9696	0.9697	0.9696	0.9696
Decision Tree Classifier	0.9295	0.9344	0.9295	0.9295
Ada Boost Classifier	0.9759	0.9763	0.9759	0.9759
XGBoost Classifier	0.9705	0.9709	0.9705	0.9705
Stochastic Gradient Descent	0.9705	0.9706	0.9705	0.9705

KNN testing accuracy

Looking at the results, we reached lower values in the KNN algorithm compared to the others. Therefore, we examined the effect on the accuracy value by changing the number of neighbors in the KNN.

	Accuracy
n_neighbors = 3	0.7352
n_neighbors = 4	0.6882
n_neighbors = 5	0.7235
n_neighbors = 6	0.6352
n_neighbors = 7	0.6588
n_neighbors = 8	0.5882
n_neighbors = 9	0.6176
n_neighbors = 10	0.5411

FINAL COMPARISON

	1.preprocess	2.preprocess	1.preprocess	2.preprocess	1.preprocess	2.preprocess
(cv=10)	Smsturkish	Smsturkish	Turkish SMS Collection	Turkish SMS Collection	bigDataset	bigDataset
KNN	0.6981	0.9436	0.6459	0.9790	0.6942	0.9767
RandomForest	0.9726	0.9537	0.8977	0.9885	0.8972	0.9807
Logistic Regression	0.9722	0.9606	0.9892	0.9878	0.9888	0.9807
Multinomial Naive Bayes	0.9676	0.9472	0.9874	0.6501	0.9786	0.6506
Support Vector Machine	0.9770	0.9660	0.9924	0.9863	0.9903	0.9744
Decision Tree Classifier	0.9172	0.9127	0.9573	0.9779	0.9477	0.9620
Ada Boost Classifier	0.9545	0.9464	0.9787	0.9848	0.9785	0.9788
XGBoost Classifier	0.9483	0.9521	0.9802	0.9868	0.9760	0.9801
Stochastic Gradient Descent	0.9649	0.9505	0.9915	0.9870	0.9901	0.9749
ANN	0.9941	0.9412	0.9632	0.9853	0.9563	0.9741
CNN	0.9588	0.9412	0.9579	0.9832	0.9572	0.9715
LSTM	0.9000	0.9294	0.9769	0.9779	0.5987	0.9634

THANK YOU FOR LISTENING TO US