OPEN ACCESS
Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 06 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

IMPROVING SPAM DETECTION FOR GERMAN USERS: A MACHINE LEARNING APPROACH TO GERMAN EMAIL CLASSIFICATION

Kashif Iqbal
Computer Science Department, Greenwich University Karachi, Pakistan.
Muhammad Khalid
Computer Science Department, Greenwich University Karachi, Pakistan.
Shamim Akhtar
Faculty of Engineering Science and Technology, IQRA University, Karachi.
Sajid Yasin
Computer Science Department, Greenwich University Karachi, Pakistan.
Noor Ahmed
Computer Science, SZABIST, Street, Karachi, 10587, Sindh, Pakistan.
Aqsa Shahid
Department of Computer Science & Software Engineering, Ziauddin University, Karachi, Pakistan.

***Corresponding author: Yusuf Yahaya Miya (jatauinitiative@gmail.com)**

Article Info

Abstract

The proliferation of unsolicited and potentially harmful emails has necessitatedthedevelopmentofrobustemailclassificationsystems.ThisstudyfocusesontheclassificationofGermanlanguageemailsusingtheCODEAALTAGdataset, which comprises a comprehensive collection of legitimate (ham) and unwanted(spam)emails.Byleveragingthisdataset,weapplyvariousmachinelearning algorithmstoaccuratelydistinguishbetweenhamandspamemails.Byleveraging this dataset, we apply various machine learning algorithms to accurately distinguishbetweenhamandspamemails.TheCODEAALTAGdatasetismeticulously curated, featuring a wide array of attributes including content-based features, header information, and technical metadata. We evaluate the performanceofseveralclassificationtechniques,includingNaiveBayes,Support Vector Machines (SVM), Random Forests, and deep learning models, in terms of accuracy, precision, recall, and F1-score. Our findings indicate that advancedfeatureselectionmethodsandensemblelearningapproachessignificantlyenhanceclassificationaccuracy.TheresultsdemonstratetheefficacyoftheCODEAALTAG dataset in training and validating high-performance email classifiers, contributing to improved email security and user experience. This studyunderscorestheimportanceofspecializeddatasetslikeCODEAALTAGinadvancingthefieldo email filtering and provides valuable insights for future research and developmentin spam detection technologies.

Keywords:
Deutsch E-Mail Klassifizierung,ham und spam Klassifikator, DeutschE-Mail classification, Textklassifikation, Spam-Erkennung, Automatische E-Mail-Sortierung.

1 Introduction

Electronic mail (email) is one of the most used forms of digital communication. It enables users to send and receive messages around the world, which makes it an essential tool for professional, commercial and personalinteractions.Automaticemailclassificationistheprocessofcategorizingemailsfortheuserautomatically[1].It isacrucialtaskforimprovingefficiencyandorganizationintoday'sdigitalworld. It allows individuals and organizations to automatically sort incoming emails into predefined categories, such as important, promotional, or spam. The volume of emails generated daily necessitates effective classification systems to manage and organize thisinformationeffectively.Table2.highlightsthenumberofemailssentandreceived daily, starting from 2017 till 2026.

Research on German email corpus is limited, one study dealt with performing sentimentalanalysisonGermancorpususingmachinelearning,theyacquireddata of private customers from a company's telecommunication sector[2]. The Code AlltagXL German Language email corpus consists of roughly 1.5M emails[3]. It is apublicly available data repository, and it provides a valuable resource for training and evaluating email classification models. This data consists of a collection of real-world emails in the German language, however, this data is unstructured and not labelled andutilizingaGermandatasetisparticularlyinsightfulforresearchtargetingusers in German-speaking regions or those working with German-language data.

Utilizing machine learning methods for the classification of emails is a beneficial approach that enhances the efficiency and accuracy of sorting large amounts of email data. Machine learning-based algorithms can learn from various datasets, and identify patterns and features which enables them to classify emails into relevant categories. The use of machine learning for email classification has proven to be much more efficient according to many research studies. One study found machine learning methods to be effective against the problem of spam email classification, they trained various machine learning models and found Naıve Bayes to be the most efficient by achieving an accuracy of 99.46% [4]. Another study proposed machine learning models, Support Vector Machine (SVM) and Artificial Neural Network (ANN) and they achieved an accuracyof98%forSVMand98.06%fortheANNmodel[5].

There is a noticeable gap in the research on the classification of emails in German. TheUnitedStatesleadstheworldindailyemailvolumewithastaggering9.7billion

S.No	Country	No.ofEmailYearly(inbillions)
1	USA	9.7
2	Deutschland	8.5
3	Ireland	8.4
4	Netherland	8.3
5	UK	8.3
6	France	8.3
7	Austria	8.2
8	Japan	8.2
9	India	8.2
10	Australia	8.1

Table: 1 Country-Specific Daily Email Statistics[6] messages. Germany follows, recording 8.5 billion daily emails, as detailed in Table 2. ThissignificantemailactivityinGermany,secondonlytotheUSglobally,provides a compelling rationale for prioritizing a German language corpus over English in our analysis.ExistingstudiesfocusonEnglishorotherspokenlanguages.

This study seeks to address this gap as the motivation behind this study is to propose a machine learning model which is capable of classifying emails in the Ger- manlanguage.ByutilizingtheCodEAlltagXLdata,thisresearchaimstoexplore the performance of different machine learning algorithms. A comparative analysis is conducted to find an effective model which can accurately classify emails into designated categories. The specific objectives of this study include:

- Gatheringthedata,labelingthedata,convertingitintoastandardformatofa dataset,preprocessingthedatasetandperformingfeatureextraction.
- Trainingandevaluatingtheperformanceofdifferentmachinelearningalgorithms (e.g.,NaiveBayes,SupportVectorMachines(SVM),RandomForest,AdaBoost and XGBoost) on the dataset.
- Identifyingthemosteffectivealgorithmforemailclassificationbasedonmetrics likeaccuracy,precision,recall,F1-score,confusionmatrixandgeometricmean.

This research contributes to the field of email classification by: Transforming the unstructureddataintoastructuredandlabeleddatasetofGermanemails.Providing a comparative analysis of various machine learning algorithms for email classification on a German dataset, highlighting the most suitable algorithm(s) for email classifica- tion tasks in a German context. This study aims to addressthelinguisticchallengesposedbytheGermanlanguage,contributingtotheadvancementofemailcl assification technology.

1.1 MotivationandContributions

Email classification is now a crucial activity for enhancing productivity and organiza- tioninbothpersonalandprofessionalcontextsduetothedailyincreaseinthevolume ofemailssentandreceived.Althoughalotofstudyhasbeendoneonclassifying

S.No	Year	No.ofEmailYearly(inbillions)
1	2017	269
2	2018	281.1
3	2019	293.6
4	2020	306.4
5	2021	319.6
6	2022	333.2
7	2023	347.3
8	2024	361.6
9	2025	376.4
10	2026	392.5
11	2027	408.2

Table: 2 No. of emails per day worldwide2017-2027 [7]

emails in English, there is a notable paucity of studies that emphasis on emails writ-ten in German. By presenting a machine learning model that can reliably distinguish between spam and ham emails, this work seeks to close this gap. In order to improve spam identification for German users and aid in the creation of more efficient email filtering systems suited to German-speaking areas, this study makes use of the CodEAlltagXL German email corpus. The significant contributions of this study are listed below. With the objective to make the unstructured CodEAlltagXL German email corpus usable for machine learning techniques, this study integrates it into a struc-turedandlabeleddataset.Thestudyisnoteworthybecauseitoffersaninvaluabletool for further research on the classification of German emails.

1. With the objective to make the unstructured CodEAlltagXL German email corpus usable for machine learning techniques, this study integrates it into a structured and labeled dataset. The study is noteworthy since it offers an invaluable tool for more research on classification of German emails.
2. InordertoclassifyGermanemails,thestudydoesanextensiveanalysisofnumerous artificial intelligence techniques, such as Naive Bayes, Support Vector Machines (SVM),RandomForest,AdaBoost,andXGBoost.Themostefficientalgorithms for this task are highlighted in this analysis along with details concerning their performance metrics, including F1-score, recall, accuracy, precision, sensitivity and specificity.
3. The study found that Random Forest had the highest geometric mean, accuracy, precision,andrecall,makingitthemosteffectivealgorithmforcategorizingGerman emails. This knowledge is important for applications that need to filter emails with high accuracy and reliability.
4. In order to enhance email classification algorithms for non-English languages, the study addresses the linguistic challenges presented by the German language, which is particularly significant for developing email filtering systems that can handle an extensive variety of linguistic and cultural settings.

1.2 Research Questions

Several possible research concerns derive from the above overview in the following ways:

1. What exactly are the most effective approaches to transform unstructured German email data into a labeled, structured dataset that machine learning algorithms can use?
2. What is the relative accuracy of several machine learning approaches (Naive Bayes, SVM, Random Forest, AdaBoost, and XGBoost) in classifying German emails asspam or ham?
3. Which machine learning approach has the greatest degree of accuracy, precision, recall,andF1scorewhenclassifyingGermantextandemails,andwhy?
4. How do the linguistic peculiarities of German language influence the effectiveness of ML algorithms in email classification, and what solutions may be applied to address these issues?

The remainder of this research article is structured as follows: Section 2 presents a review of related work in email classification using machine learning. Section 3 details the methodology, including data pre-processing,featureengineering,andthechosenmachinelearningalgorithms.Section4discussestheexperiment alresultsandanalysis. Section 5 offers a conclusion, summarizing the key findings and outlining potential future directions.

2 Literature Review

Therehasbeenmuchresearchtodevelopsuchsystemsthatcanhandleemails. One such system is developed by Van Den Poel and Coussement that segregates complaintsandnon-complaintsbyemailclassification[8].Intheirdetailedwork they have used Boosting as their main classification technique over email corpus and claimed it to be young and powerful machine learning technique. Beside this, therehas been another text classification performed by Jakub, Ahmet and Rafal [9] over three datasets Spambase Data Set (1999), Farm Advertisement (2011) and Amazon book reviews Data Set (2016) where they applied LSTM and BLSTM (bi-directional LSTM) that has performed significantly better with results LSTM performance up to 99.79%followedbyBi-directionalLSTMat99.83%onspamcollection.

Similarly,therehasbeenaUnsolicitedBulkEmail(UBE)classificationproposed by Mohammed S. et al. [10] that uses spam-ham dictionary, after pre-processing and data-mining algorithms it has been suggested that Naive Bays and SVM are the most efficient.OlaAmayrietalperformedspamfilteringwithsupportvectormachines

[11]intheirdetailedworktheyhaveclaimedthatbestresults.

In another work presented by Subramniam et al. [12] has performed spam-filtering over foreign language (Malay) using Na¨ıve Bayes. The accuracy that they achievedwas 69%. There has also been work done over Turkish language purposed by LeventOzguretal[13]wheretheyhavebrieflydiscussedworkintwosectionsfirstoneis Morphology and the second one is Learning Module and performed classification using two types of ANN; Single Layer Perceptron (SLP) and Multi-layer Perceptron (MLP) for which they have achieved 90% and 68%.

Mahmoud Jazzar et al. have proposed machine learning techniques over UCI machine learning repository for email classification. They have used 1367 spam e-mail and4361aslegitimate.TheyhavetriedJ48,SVM,ANNandNaiveBayes-classifier for this task and have achieved. With SVM being the highest in their two methodswith 93.91% accuracy in the first experiment and 94.06% accuracy in the second one [14].Similarly,inanotherworkthatcontainssupervisedmachinelearningtechniques

[15] have claimed that using FBL in naive bayes can reduce the number of attributes that are dependent thus improvement in the model. However, they have achieved93% accuracy with MLP.

K. Iqbal et al. implemented Bidirectional Encoder Representations from Trans- formers (BERT) on 6 different Enron datasets, containing ham and spam emails. They performed experiments in 3 sets, which contained different batch sizes and epochs.Accuracywasusedastheevaluationmetric.Thelowestaccuracywas84.69% andthehighestaccuracywas97.66%achievedontheir5thEnrondatasetusingbatch size of 64 and 40 epochs [5].

There has also been a method discovered to classify e-mail using SVM [16] where the authors have used Linear Kernel and Gaussian Kernel and captured results that Linear Kernel provided more test accuracy and reasoned that their dataset contained more number of features. And there has been a comparative analysis between Naive Bayes and SVM classifiers performed [17] in their experiments they have used

Multi- nomial Naive Bayes and on the other hand the Linear SVM. While on testing model they have segregated training email into different six different portions ranging from 1000 to 6000 where the test emails were 200 in each portion. On the basis of their results, they have concluded that Support Vector Machine has provided better results for classification.

S. Khan et al. proposed a new fuzzy-logic evaluation metric to evaluate the performance of email spam detection algorithms by combining accuracy, recall, and precision metrics. They measured performance of BERT and LSTM on three datasets. LSTMperformedbetterfortheEnronandPUdatasetswhile BERT performedbet- ter for the Lingspam dataset. The results showed potential for further developmentfor the proposed evaluation metric [18].

3 Methodology

The classification of emails using supervised machine learning techniques is the focus of this study. In order to do this activity, there are five main steps involved. Data collection,datacleaning,featureextraction,training,andmodelevaluationarethe

Study	Model	Year	MainFindings
Coussement.etal.[8]	AdaBoost	2008	Effectiveinclassifyingcomplaints andnon-complaintsemails.
Nowak.eal.[9]	LSTM,BLSTM	2017	LSTMachieved99.79%accuracyand BLSTMachieved99.83%accuracy.
MohammedS.etal.[10]	Na¨ıveBayes,SVM	2013	Na¨ıve Bayes andSVM were fficient forUBEclassification.
Amayri.etal.[11]	SVM	2010	SVMachievedhighprecisionandrecallrates.
Subramniam.etal.[12]	Na¨ıveBayes	2010	Achievedaccuracyof96%forspam classificationinMalaylanguage.
Ozgur.etal.[13]	SLP,MLP	2004	Accuracy90%withSLPand68% withMLPinTurkishlanguage.
Jazzar.etal.[14]	J48,SVM,ANN,Na¨ıveBay es	2021	SVMachieved93.91%and94.60%accuracy
Renuka.etal.[15]	J48,Na¨ıveBayes,MLP	2011	Na¨ıve Bayesachieved91% and MLPachieved93%accuracy.
K.Iqbal.etal.[5]	BERT+TF2.0	2022	Achievedthehighestaccuracyof97.66%.
Singh.etal.[16]	SVM(LinearKernel,Gaussi anKernel)	2018	LinearKernelhadbetteraccuracy asdatasethadmanyfeatures.
Thae.Ma.eal.[17]	Na¨ıveBayes,SVM	2020	SVMprovidedbetterresultsforclassification acrossdifferentportionsofemails.
S.Khanetal.[18]	LSTM,BERT	2022	LSTMperformedbetterwhenutilizingfor twodatasetsandBERTperformedbetter foronlyonedataset.

Table: 3 Summary of existing studies.

processes involved in performing a comparative analysis. Naive Bayes, Support Vector Machine (SVM), Random Forest, AdaBoost, and XGBoost are the five models used.The Random Forest classifier yielded the best accuracy of 99.72% in a study by M.Rathi[19].Thestepsofthetechniqueforthisinvestigationaredescribedinfullbelow

3.1 Data Collection:

Thedatasetof"CodEAlltagXLGERMAN"wasutilized[20].Thisdatasetcontains raw Deutsch language emails separated into 9 different folders containing each folder contains subfolders with emails in text files (.txt). For this study, different folders were chosentogetarandomsetofdata,compiledandthenmanuallylabelledusingGoogle Translate'sDeutschlanguagetranslationfeature.Atotalof3070fileswerelabeled. To ensure the data is in a suitable format for machine learning algorithms, after the labeling of the text files, they were compiled in a Comma-Separated Values (CSV) file format using scripting techniques in Python.

3.1.1 Data Cleaning and Preprocessing

As preprocessing is an essential part of machine learning, a study by S. Alam proposes that the accuracy of the Naive Bayes and SVM algorithms was improved after applying data preprocessing steps [21]. In this study, after converting the data into CSV file format, it was imported into Data Frame format using Pandas library of Python. Then missing values were checked to ensure data integrity. Data must be pre-processed to removeallspecialcharactersandpunctuationmarks[22].Preprocessingwasdone toensuretheremovalofHyperTextMarkupLanguage(HTML)tags,punctuations,

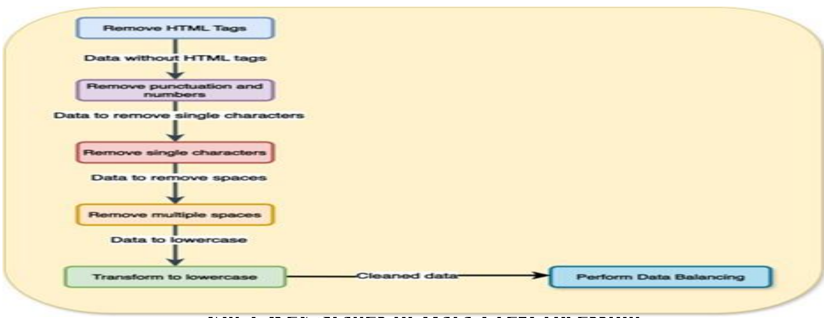

Fig.1 Key Stages in Data Pre-Processing

and numbers, single character removal, and multiple space removal were applied to clean the text further. To address the imbalance nature of the dataset, oversampling was done using the Synthetic Minority Oversampling Technique (SMOTE), which is available in the imbalanced-learn library, Imbalanced-learn is an open source library which is MIT-licensed and relies on scikit-learn. It provides various tools for dealing withimbalanceddatasets[23].SMOTEservesasapioneeringoversamplingmethod in the research field for classification of imbalanced data sets [24]. Fig. 1 shows the complete steps of the data pre-processing phase. Table 1. shows the distribution of classes before data balancing and Table 2. shows the distribution of classes after data balancing.

Class	No.ofE-mails	Class	No.ofE-mails
ham (0)	2997	ham (0)	2997
spam(1)	73	spam(1)	2997

Table:4 Distribution of classes before and after data balancing

4 Feature Extraction

Feature extraction is the step of transforming raw textual data into numerical format that machine learning algorithms can understand. Several techniques were used to vectorize the email text data, it is essential to remove noise and extract meaningful features.

4.1 Stop Words Removal

Stopwordsarethecommonwordsthatmostofthetimesdonotcontributesignificant meaning to the text such as "and", "the", and "is" in English. Removing stop words reduces the dimensionality of the text data while preserving important information or context of the text. Removal of stop words brings advantages in less usage of storage space and amount of time spent computing [25]. spaCy library was used to load the German language model's default stop words and to use it in text vectorization by the useofTermFrequency-InverseDocumentFrequency(TF-IDF).

4.1.1 spaCy

spaCy [26] is an open-source and free Python library for advanced Natural Language Processing (NLP) tasks. It is designed to use for production use cases and for building real-world applications. It offers easy to use tools for working with large-scale text data.spaCycontainsanextensivesuiteoffeaturessuchastokenization,part-of-speech tagging,dependencyparsing,namedentityrecognition,andlemmatization.spaCy has support for multiple languages. It can easily be integrated with machine learning frameworks and allows for custom models which are tailored for NLP related tasks.

4.2 Text Vectorization

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical technique to evaluatetheimportanceofawordinadocumentorcollectionofdocuments.Itper- forms the combination of two metrics that are term frequency (TF), it is a count of howoftenawordappearsinadocument,andtheinversedocumentfrequency(IDF) which performs the function of scaling down words that appear more frequently across documents. The IF-IDF score increases as the number of times a word is present in asingledocumentbutgetsoffsetbythefrequencyofthewordintheentirecorpus. Sklearn'slibraryprovidesTfidfVectorizerintheirfeatureextractionpackage,fortext vectorization, this study utilized the TfidfVectorizer and German stop words were passed in the parameters of the vectorizer.

TheTF-IDFvalueofawordtinadocumentdinsideacorpusDcanbefound using the formula:

$$TFIDF(t,d,D) = TF(t,d) * IDF(t,D) \tag{1}$$

WhereTFistheTermFrequency,thatismeasuredby:

$$TF(t,d) = \frac{ft,d}{Nd} \tag{2}$$

Here,$f_{t,d}$isthefrequencyoftermtindocumentd.AndNdisthetotalnumber of terms in a document d.

IDFisthemeasureofimportanceofatermintheentirecorpus.Itismeasuredby:

$$IDF(t, D) = log \frac{N}{nt} \quad (3)$$

Here,NisthetotalnumberofdocumentsinthecorpusD.Andntisthenumber of documents in which a term t appears.

4.3 Transforming the Data

After the setup of vectorizer, data was split using sklearn's "train test split" function intotrainingandtestingsetswith80%fortrainingsetand20%fortestingset.Training and testing features that are X train and X test were transformed using TF-IDF vectorizer. Hence, TF-IDF vectorization converted the preprocessed text data into numericaldatathatcapturestheimportanceofeachwordrelativetotheentirecorpus. This step is essential as it help machine learning algorithms to understand the dataandeffectivelytrainingthemodelstoclassifytheemailsbasedontheirtextualcontent.

5 Model Training

NaiveBayesclassifier is a probabilistic machine learning based on the BayesTheorem,whichworksbyassumingthepresenceofaspecificfeatureinaclassthat is unrelated to the presence of any other feature. In other words, it does not learn which of the features are the most important to differentiate between classes. Naive Bayes is a useful algorithm for text classification.

Support Vector Machine (SVM) is a quite powerful and one of the versatile supervised machine learning algorithms used for both classification and regression tasks. It looks to find the optimal hyperplane which best separates the data into different classes.

Random Forest is an ensemble machine learning method that builds multiple decision trees during training and outputs the class that is the mode or most repeated of the classes (classification). For regression it does mean prediction of the individual trees.RandomForestcombinestheideaof"bagging"(BootstrapAggregating)with

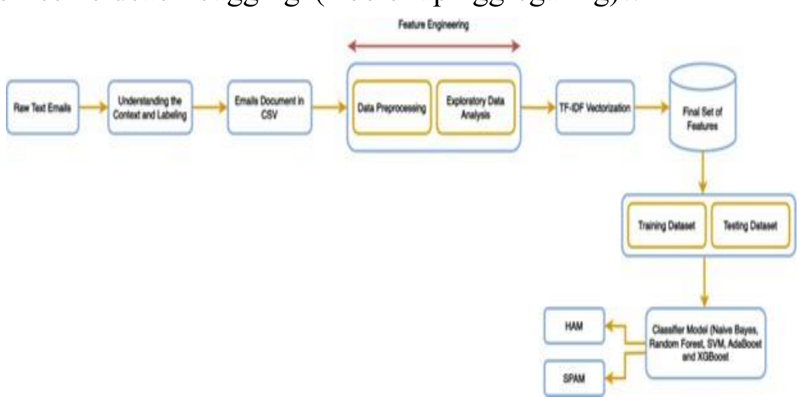

Fig. 2 Architecture of the Proposed Classification Model

randomfeatureselectionwhichenhancesthemodel'saccuracyandrobustness.

AdaptiveBoosting(AdaBoost)isanensemblemachinelearningalgorithmthatcombinesmultiplenumberofweakclassifierstoformastrongclassifier.AdaBoost takes two steps for training and merging weak classifiers: first, it determines which traininginstanceseachclassifiershouldbetrainedon,andsecond,itdetermines the weight of each classifier in the vote [27]. It adjusts the weights of misclassified instances so that subsequent classifiers focus is more on the hard-to-classify samples.

ExtremeGradientBoosting(XGBoost)isanadvancedimplementationof gradient boost algorithm, it is designed for speed and performance. It combines decisiontreesandgradientdescenttoenhancetheperformanceofthemodel.

Forthisstudy,fiveclassificationmachinelearningmodelsweretrainedwhichare the following, Naive Bayes, SVM, Random Forest, AdaBoost and XGBoost. Each modelwastrainedonthe80%trainingdata.XGBoost'smodelwashyperparameter tuned to observe any enhancements in the performance of the model. Fig. 2 shows the flow of every model. The first step begins with collecting raw text emails, which in thenextstep,areanalyzedtounderstandthecontextoftheemailandmanuallylabeled aseitherspamorham.Afterthisstep,alllabeledemailsaretransformedintoaCSV documentthenthenextstepinvolvesdatapreprocessingandexploratorydataanalysis to clean and further understand the data, which is then followed by using TF-IDF vectorization to transform text data into numerical form which is suitable for machine learning algorithms. These features can form the final dataset, this dataset is then splitintotrainingandtestingdatasets.Inthefinalstep,classifiermodelsincludingNaive Bayes, Random Forest, SVM, AdaBoost, and XGBoost, are trained on the training datasetandevaluatedonthetestingdatasettoclassifytheemailsintohamorspam.

6 Model Evaluation

As this study deals with the classification technique of machine learning, each model wasevaluatedbyclassificationperformancemetricsthatareprecision,recall,f1-score, support, accuracy, sensitivity, specificity and geometric mean. Additionally, eachmodel was evaluated by the use of confusion matrices.

Accuracyof the model measures the overall correctness of the model. Accuracycan be mathematically given as:

$$Accuracy = \frac{T.P + T.N}{T.P + T.N + F.P + F.N} \tag{4}$$

Where T.P is True Positive, T.N is True Negative, F.P is False Positive and F.N is False Negative.

Precisionis the measure of accuracy of the correct predictions made by the model. Precisionofthemodelcanbemathematicallygivenas:

$$Precision = \frac{T.P}{T.P+F.p} \tag{5}$$

HereT.PisTruePositiveandF.PisFalse Positive.

Recall measures the model's capability to identify all relevant instances. It is theratio of correctly predicted observations to all the observations in the actual class.

$$Recall = \frac{T.P}{T.P+F.N} \quad (6)$$

Intheaboveequation,TPisTruePositiveandFPisFalsePositive.

F1-scorecombinesrecallandprecisionboth[18].Itisametricthatevaluates the predictive skills of a model by examining its class-wise performance rather than examining an overall performance like done by accuracy. F1-score is mathematically given by:

$$F-1Score = 2 * \frac{precision * recall}{precision+recall} \quad (7)$$

Sensitivity measures the model's ability to correctly identify the positive pre- dictions. It can also be termed as a true positive rate. Sensitivity is mathematically calculated by:

$$\text{Sensitivity} = \frac{T.P}{T.P+F.N} \quad (8)$$

Sensitivity measures the model's ability to correctly identify the positive pre- dictions. It can also be termed as a true positive rate. Sensitivity is mathematically calculated by:

$$\text{Specificity} = \frac{T.N}{T.N+F.P} \quad (9)$$

Geometricmeanmeasurestheoverallperformanceofamodel.Itcombines both, sensitivity and specificity into a single metric to provide a balanced evaluation. Geometric mean is given by:

$$\text{GeometricMean} = \sqrt{sensitivity * specificity} \quad (10)$$

By using these metrics, it enhances the understanding of a model's performance, allowing to select the most optimal model for the email classification task.

7 ResultsandDiscussion

Eachofthe5classifiers,NaiveBayes,SVM,RandomForest,AdaBoostandXGBoost weretrainedon80%ofthedataandtotestthemodel'sperformanceonunseendata, it was tested using 20% of data. The performance of each model was evaluated using several classification metrics including accuracy, precision, recall, F1-score, confusion matricesandgeometricmean.Forcalculatingthegeometricmean,thesensitivityandspecificityofeachmodelwe realsocalculated.Theseevaluationmetricsprovidea comprehensive understanding of how well each of the models has performed by classifying emails into their respective categories—the results of the models provided below.Fig.3showsthewordcloud,representingthefrequentlyoccurringwordsin thedataset.Thisvisualrepresentationdisplaysthecommonkeytermsandthemes.

Table 5. summarizes the metrics, accuracy, precision, recall and F1-score of each model evaluated using the actual values (testing labels) and predicted values. Naive Bayes achieved the lowest accuracy of 0.91, precision of 0.88, 0.93 recall and 0.90 F1-score.ItcanbeobservedthatNaiveBayeswasthelowestineverymetricincomparison to other models. SVM achieved an accuracy of 0.98, a precision of 0.96 and a perfect recall of 1, having an F1-score of 0.98. Random Forest achieved an accuracy of 0.99, precision of 0.99, recall of 0.99 and F1-score of 0.99. For the boosting algorithms, AdaBoostachievedanaccuracyof0.98,precisionof0.98,recallof0.99andF1-scoreof

0.99. In Fig.4, you will find an accuracy report for each model, providing insights into theireffectiveness.XGBoostwastestedwithabasemodelandthenitshyperparameter was tuned by finding the best parameters using GridSearchCV while fitting the model in 3 folds. The parameters were then used to tune the model. Both the models, the base and tuned model achieved same the accuracy, precision, and F1-score i.e., 0.99 respectivelybutthetunedmodelachieved0.99recall.However,itshouldbenotedthat both models performed slightly differently in metrics, namely, the confusion matrix and geometric mean. The results of these metrics are also provided in this study.

Fig.3VisualizingtheDistributionofWordFrequenciesintheDataset(Wordcloud)

Model	Accuracy	precision	Recall	F1-Score
NaiveBayes	0.9058	0.8771	0.9312	0.9033
SVM	0.9791	0.9578	1	0.9784
RandomForest	0.9933	0.9982	0.9877	0.9929
AdaBoost	0.9867	0.9825	0.9894	0.9859
XGBoost	0.9875	0.99894	0.9841	0.9867
XGBoost(Tuned)	0.9883	0.9894	0.9859	0.9876

Table5Comparative Model Performance: Accuracy, Precision, Recall, and F1-score

Fig.5presentsasummaryofprecisionforallmodels.Additionalperformance metrics, including recall and F1-score, are shown in Figures 3 and 4, respectively. These figurescollectivelyprovideacomprehensiveoverviewofeachmodel'sperformance. Fromtheresults,itcanbeinterpretedthatRandomForestperformedthebestthan othermodels,makingitthemostefficientmodelofthisstudy.

Table6.showsthesummarizedresultsofeachmodel'sconfusionmatrix.These include the true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP). NaiveBayeshadahighnumberofFP(74)andFN(39),with558truenegatives and528truepositives.TheFPandFNcountsofNa¨ıveBayesshowthatithada reasonable performance but

it may struggle with differentiating between classes, indicating that it can have a higher rate of misclassification as compared to other models.SVMhadthelowestpossibleFNcounti.e.,0andalowcountofFPi.e., 25. It achieved 607 TN and 567 TP. Random Forest had the lowest count of FP i.e., 1and7FNalongside631TNand560TP.ThelownumberofFPandFNreflects

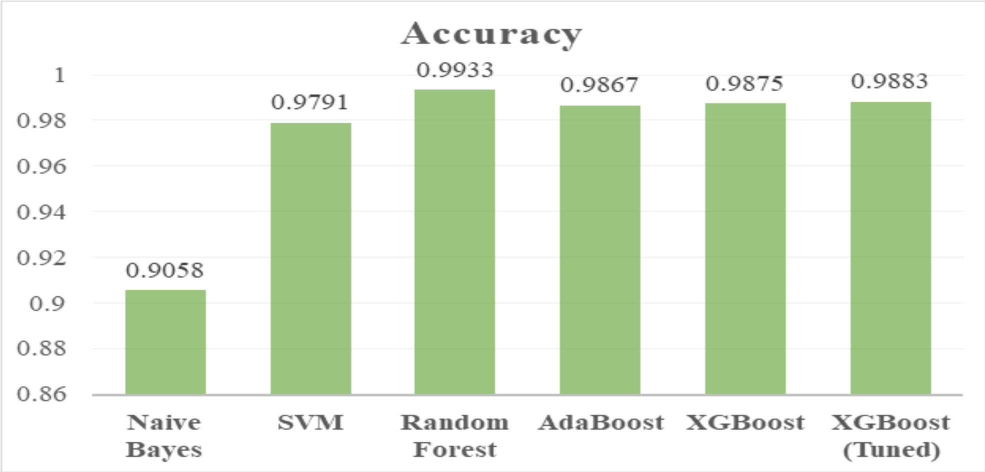

Fig.4ComprehensiveReportofOverallModelAccuracy

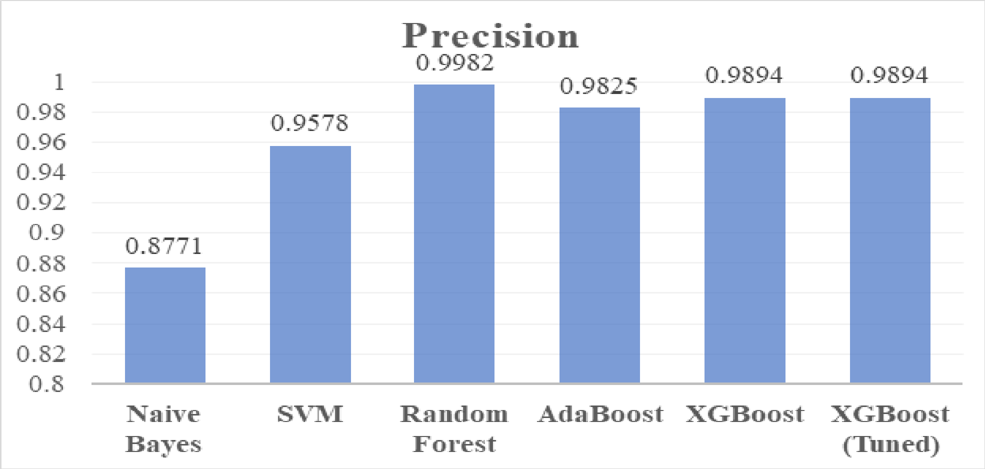

Fig.5OverallAccuracyPerformanceofthePredictiveModel

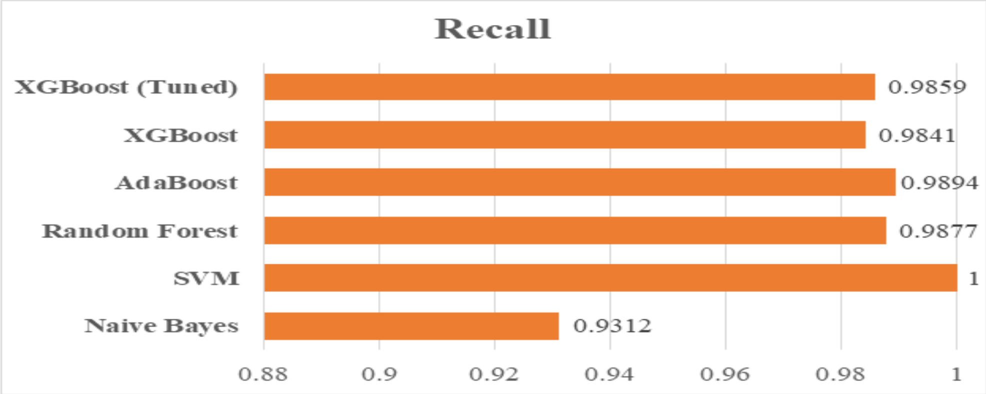

Fig.6RecallPerformancebyMode:AComparativeSummary

Model	TN	FP	FN	TP
NaiveBayes	558	74	39	528
SVM	607	25	0	567
RandomForest	631	1	7	560
AdaBoost	622	10	6	561
XGBoost	626	6	9	558
XGBoost(Tuned)	626	6	8	559

Table 6Detailed Confusion Matrix Results for Each Trained Classification Model

Random Forest's ability to accurately classify both ham and spam, which contributesto achieving high precision and recall.

For the boosting models, AdaBoost had 10 FP and 6 FN, 622 TN and 561 TP.The low count of FP and FN indicates that also AdaBoost minimizes misclassification.XGBoostperformedbetterthanAdaBoost,thetunedmodelachievedresultsof 6 FP, and 8 FN alongside 626 TN and 559 TP.

All models demonstrated strong performance, SVM, Random Forest and XGBoost(both tuned and untuned) particularly excelled with their low FP and FN counts, with Random Forest being the most optimal model

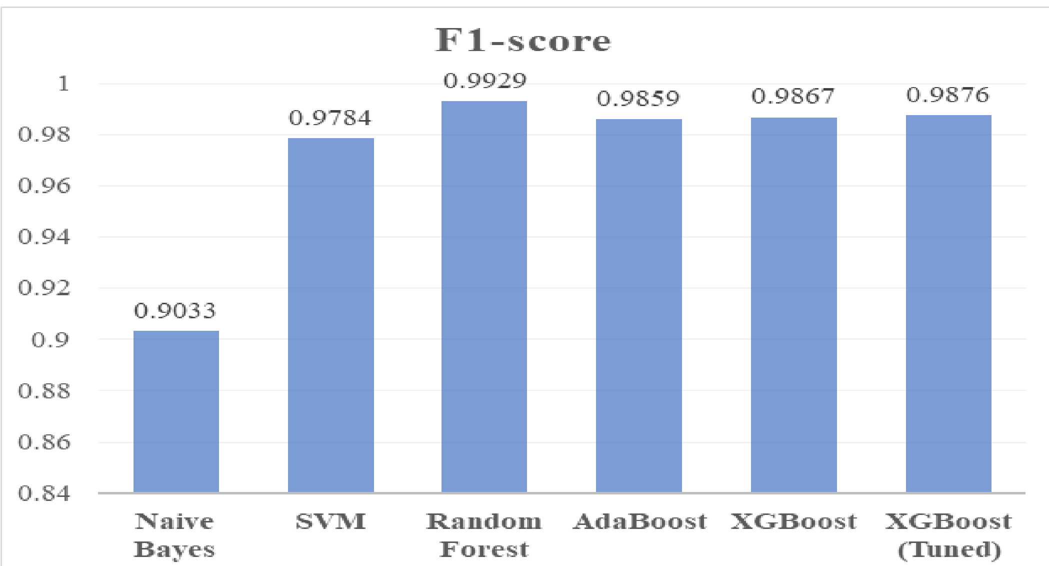

Fig.7F-1ScoreComparisonAcrossDifferentModels

A comprehensive appearance at the performance of each model we analyzed in this study—Naive Bayes, SVM, Random Forest, XGBoost, AdaBoost, and the opti-mizedXGBoost—isshowninFig.8.Aconfusionmatrix,whichisbasicallyatable that contrasts the model's predictions with the actual ground truth, can be seen for each of these different approaches.

Therefore, we can directly evaluate each model's strengths and weaknesses across several categories by looking at the confusion matrix for each model in Fig. 8. One model may, for instance, be very good at accurately recognizing a certain class (high TP), but it also has a tendency to incorrectly categorize other cases as belonging tothat class (high FP). A more conservative model might provide fewer false positivesbut possibly more false negatives.

Since it enables us to observe the tangible effects of our hyperparameter optimiza- tion efforts on the model's classification behavior, the inclusion of the tuned XGBoost model'sconfusionmatrixisveryinstructive.Todeterminewhichkindsoferrors were decreased or increased through the tuning process, we may directly compare its performance to that of the base XGBoost model.

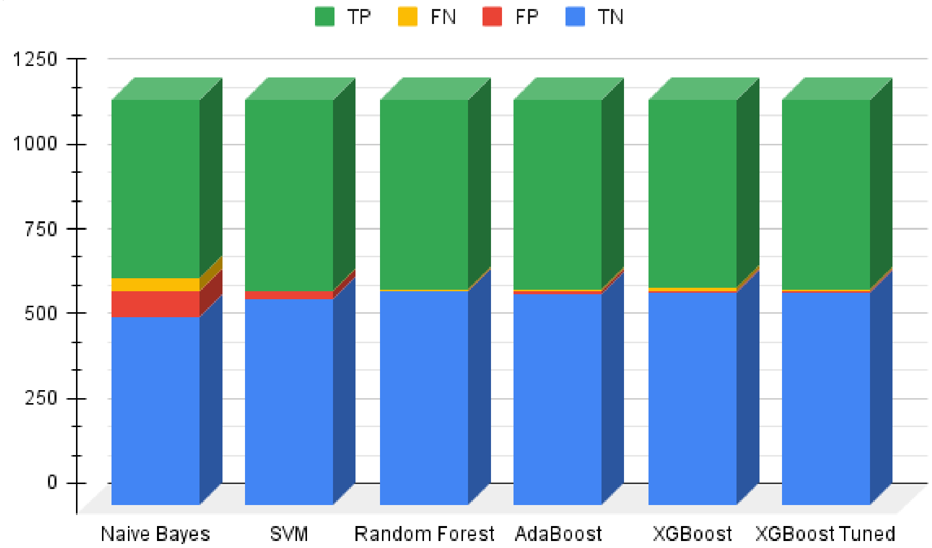

Fig. 8VisualComparisonofConfusionMatricesAcrossAllEvaluatedClassificationModels

In general, Fig. 8 provides a more comprehensive view of the performance of each model than just the total accuracy scores. It enables us to identify certain trends in misclassifications and acquire a better understanding of how each algorithm makes predictions.

Table 7. shows the summarized sensitivity, specificity, and geometric mean results ofeachmachine-learningmodel.Na¨ıveBayesachievesthelowestgeometricmean,its lowerspecificitysuggests a comparatively high numberof false positives.SVMhad the perfect sensitivity of 1 and achieving a geometric mean of 0.98, it outperforms Naive Bayes. In these metrics as well, Random Forest had the highest sensitivity, specificity and geometric mean of 0.99. Both models of XGBoost performed slightly better than AdaBoost with the tuned model achieving a geometric mean of 0.99. The high geometric mean of the tuned XGBoost model makes it a highly effective and reliable model for classification tasks.

Model	Sensitivity	Specificity	GeometricMean
NaiveBayes	0.9312	0.8829	0.9067
SVM	1	0.9604	0.9800
RandomForest	0.9876	0.9984	0.9930
AdaBoost	0.9894	0.9841	0.9867
XGBoost	0.9841	0.9905	0.9873
XGBoost(Tuned)	0.9858	0.9905	0.9881

Table7Summary of Sensitivity, Specificity and Geometric Mean

In the results of these metrics, Random Forest outperforms all the other models. The results confirm that Random Forest is the most suitable model for the classifica- tion task as it can be applied in a wide range of applications requiring high reliability and precision. Examining Fig. 8, we can see that the random forest model achievedthe best overall performance among the models evaluated in this research.

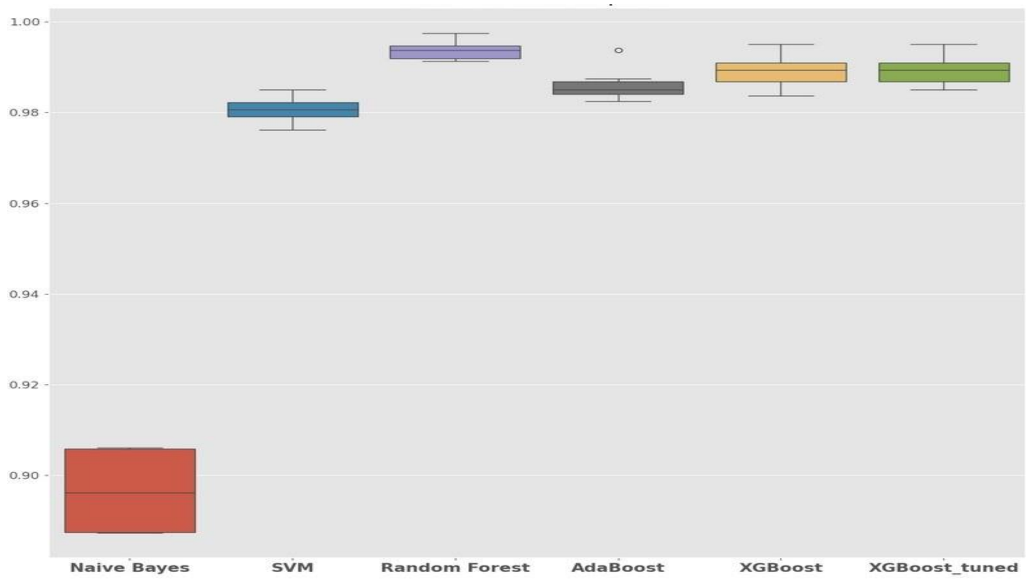

Fig.9OverallPerformanceofdifferentModels

Fig.10presentsthedistributionofsensitivity,specificity,andgeometricmean for all the models. This graphical representation shows the central tendency andvariabilityoftheseperformancemetricsacrossthemodelsevaluatedinthisstudy.

8 FutureWork

One promising avenue for future research lies in exploring deep learning techniques foremailclassification.Deeplearningmodels,withtheirabilitytohandlecomplex and unstructured data like email content, could potentially achieve even higher accuracyincategorizingemails.Thisapproachcouldinvolveutilizingarchitectures like recurrent neural networks (RNNs) or transformers, which are adept at capturing sequential information and contextual relationships within text data. By leveraging these powerful models, researchers could potentially develop more robust and adapt- able email classification systems.

Furthermore,thisstudy’sfocusonthe”CodEAlltagXLGERMAN”dataset presents an opportunity for broader exploration. Expanding the research to include datasets in various languages would provide valuable insights into the generalizabil-ity of the employed models. By evaluating their performance across diverse linguistic andculturalcontexts,researcherscouldgainadeeperunderstandingofthemodels’

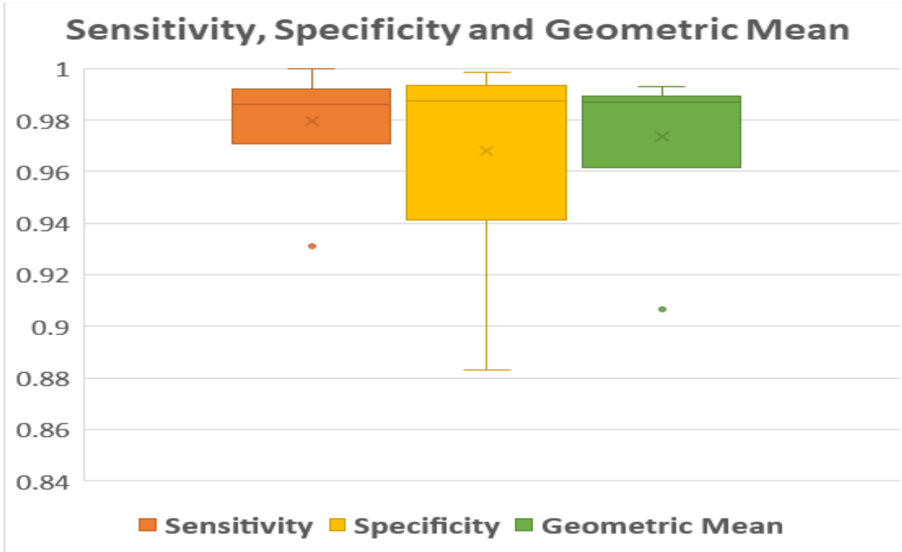

Fig. 10The effectiveness of all models is compared using sensitivity, specificity, and the geometricmean.

strengths and weaknesses. This cross-linguistic analysis would not only contribute to the development of more universally applicable email classification systems but also shedlightonhowlanguageandculturalnuancesmightinfluenceemailcommunication patterns

9 Conclusion

Emails are one of the most important forms of communication in both personal and professionalenvironments,itisamediumforinformationexchange,collaborationand coordination. Due to the growing number of users on the internet, spam emails have become commonplace in the digital world. There is extensive research done on email classification in the English language, and research on the classification of emails in foreign languages is still developing. This research aimed to classify emails in the Deutsch language using machine-learning techniques that can effectively handle the diversity and complexity of email content.

In this study, five machine learning classifiers namely, Naive Bayes, SVM, Random Forest, AdaBoost and XGBoost on a Deutsch language dataset for classifying emails into their respective categories i.e., ham or spam. The models were evaluated using the classification metrics of accuracy, precision, recall, F1-score, confusion matrices, sensitivity, specificity and geometric mean. All models showed a strong performance but with varying degrees of efficacy.

While being effective, Naive Bayes showed the lowest performance across most met- rics, indicating a high rate of misclassification. SVM excelled with high performance across all metrics while Random Forest performed the best than other models, achiev- ing the highest accuracy, precision, recall and geometric mean.

The boosting algorithms in this study which are, AdaBoost and XGBoost, also per- formed well but XGBoost both in its base and tuned forms, achieved higher accuracy than AdaBoost and balanced performance in other metrics. The hyperparameter tuning of XGBoost further enhanced its performance in confusion matrix metrics, specificity, and geometric mean, contributing to its robust performance. The overall results of this study have presented that the Random Forest algorithm is the most reliable and efficient for classifying emails. It achieved strong performance in multiple evaluation metrics, making it the ideal choice of use in applications that require high precision and reliability

References

1. Brutlag, J.D., Meek, C.: Challenges of the email domain for text classification. In: ICML, pp. 103–110 (2000)
2. Markscheffel, B., Haberzettl, M.: Sentiment analysis of german emails: A com- parison of two approaches. In: DATA, pp. 385–391 (2019)
3. Krieg-Holz, U., Schuschnig, C., Matthies, F., Redling, B., Hahn, U.: Code alltag:A german-language e-mail corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2543–2550 (2016)
4. Awad, W., ELseuofi, S.: Machine learning methods for e-mail classification. InternationalJournalofComputerApplications**16**(1),39–45(2011)
5. Iqbal, K., Khan, M.S.: Email classification analysis using machine learning techniques. Applied Computing and Informatics (2022)
6. demandsage:Numberofsentandreceivede-mailsperdayworldwide from 2017 to 2027. https://www.demandsage.com/how-many-emails-are-sent- per-day/. [Online; accessed 15-April-2025] (2025)
7. Oberlabo:Numberofsentandreceivede-mailsperdayworldwidefrom2017 to 2027. https://www.oberlo.com/statistics/how-many-emails-are-sent-per-day. [Online; accessed 15-April-2025] (2025)
8. Coussement, K., Poel, D.: Improving customer complaint management by auto- matic email classification using linguistic style features as predictors. Decision support systems **44**(4), 870–882 (2008)
9. Nowak, J., Taspinar, A., Scherer, R.: Lstm recurrent neural networks for short textandsentimentclassification.In:ArtificialIntelligenceandSoftComputing:16th InternationalConference,ICAISC2017,Zakopane,Poland,June11-15,2017, Proceedings,PartII16,pp.553–562(2017).Springer
10. Mohammed,S.,Mohammed,O.,Fiaidhi,J.,Fong,S.,Kim,T.H.:Classify- ing unsolicited bulk email (ube) using python machine learning techniques. International Journal of Hybrid Information Technology **6**(1), 43–56 (2013)
11. Amayri,O.,Bouguila,N.:Astudyofspamfilteringusingsupportvector machines. Artificial Intelligence Review **34**, 73–108 (2010)
12. Subramaniam, T., Jalab, H.A., Taqa, A.Y.: Overview of textual anti-spam filtering techniques. Int. J. Phys. Sci **5**(12), 1869–1882 (2010)
13. Özgür,L.,Güngör,T.,Gürgen,F.:Adaptiveanti-spamfilteringforagglutinative languages:aspecialcaseforturkish.PatternRecognitionLetters**25**(16),1819– 1831 (2004)
14. Jazzar,M.,Yousef,R.F.,Eleyan,D.:Evaluationofmachinelearningtech- niques for email spam classification. International Journal Of Education And Management Engineering **11**(4), 35–42 (2021)
15. Renuka, D.K., Hamsapriya, T., Chakkaravarthi, M.R., Surya, P.L.: Spam classi- fication based on supervised learning using machine learning techniques. In: 2011 International Conference on Process Automation, Control and Computing, pp. 1–7 (2011). IEEE

16. Singh,M.,Pamula,R.,*etal.*:Emailspamclassificationbysupportvector machine. In: 2018 International Conference on Computing, Power and CommunicationTechnologies(GUCON),pp.878–882(2018).IEEE

17. Ma, T.M., Yamamori, K., Thida, A.: A comparative approach to na¨ıve bayes classifierandsupportvectormachineforemailspamclassification.In:2020IEEE 9th Global Conference on Consumer Electronics (GCCE), pp. 324–326 (2020). IEEE

18. Khan, S.A., Iqbal, K., Mohammad, N., Akbar, R., Ali, S.S.A., Siddiqui, A.A.: A novel fuzzy-logic-based multi-criteria metric for performance evaluation of spam email detection algorithms. Applied Sciences **12**(14), 7043 (2022)

19. Rathi, M., Pareek, V.: Spam mail detection through data mining-a comparative performance analysis. International Journal of Modern Education and Computer Science **5**(12), 31 (2013)

20. CodeAlltag.:Numberofsentandreceivede-mailsperdayworldwidefrom2017 to 2026. https://github.com/codealltag/CodEAlltag pXL GERMAN. [Online; accessed 09-September-2023] (2016)

21. Alam, S., Yao, N.: The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Computational and Mathematical Organization Theory **25**, 319–335 (2019)

22. Iqbal,K.,AKhan,S.,Anisa,S.,Tasneem,A.,Mohammad,N.:Apreliminary study on personalized spam e-mail filtering using bidirectional encoder repre- sentations from transformers (bert) and tensorflow 2.0. International Journal of Computing and Digital Systems **11**(1), 893–903 (2022)

23. developers,T.:Numberofsentandreceivede-mailsperdayworldwidefrom2017 to 2026. https://imbalanced-learn.org/stable/. [Online; accessed 03-March-2024] (2014)

24. Pradipta,G.A.,Wardoyo,R.,Musdholifah,A.,Sanjaya,I.N.H.,Ismail,M.:Smote forhandlingimbalanceddataproblem:Areview.In:2021SixthInternational Conference on Informatics and Computing (ICIC), pp. 1–8 (2021). IEEE

25. Silva, C., Ribeiro, B.: The importance of stop word removal on recall values in text categorization.In:ProceedingsoftheInternationalJointConferenceonNeural Networks, 2003., vol. 3, pp. 1661–1666 (2003). IEEE

26. 2016-2024Explosion,h..u.y...n..O.a..-M.-.title=Spacy:ANaturalLanguage Processing

27. Borg,A.,Boldt,M.,Rosander,O.,Ahlstrand,J.:E-mailclassificationwith machine learning and word embeddings for improved customer support. Neural Computing and Applications **33**(6), 1881–1902 (2021)