ENGINEERING IN ADVANCED
RESEARCH SCIENCE AND
TECHNOLOGY

ISSN 2350-0174
Vol.03, Issue.04
July-2023
Pages: 893-995

AN EFFICIENT METHOD FOR TEXT CLASSIFICATION USING NAIVE BAYES

[1]BALLA MOUNICA, [2]Ms.K.V.DURGA DEVI
[1]M.Tech, Department of Artificial Intelligence, Kakinada Institute of Engineering & Technology for Women, Korangi
[2]Guide,Assistant Professor, Department of Artificial Intelligence, Kakinada Institute of Engineering & Technology for Women, Korangi

ABSTRACT: Spam emails are known as unrequested commercialized emails or deceptive emails sent to a specific person or a company. Spams can be detected through natural language processing and machine learning methodologies. Machine learning methods are commonly used in spam filtering. These methods are used to render spam classifying emails to either ham (valid messages) or spam (unwanted messages) with the use of Machine Learning classifiers. The proposed work showcases differentiating features of the contentof documents. There has been a lot of work that has been performed in the area of spam filtering which is limited to some domains. Research on spam email detection either focuses on natural language processing methodologies on single machine learning algorithms or one natural language processing technique on multiple machine learning algorithms. In this Project, a modeling pipeline with Naive Bayes is developed to review the machine learning methodologies.

INTRODUCTION: Technology has become a vital part of life in today’s time. With each passing day, the use of the internet increases exponentially, and with it, the use of email for the purpose of exchanging information and communicating has also increased, it has become second nature to most people. While e-mails are necessaryfor everyone, they also come with unnecessary, undesirable bulk mails, which are also called Spam Mails. Anyone with access to the internet can receive spam on their devices. Email system is one of the most effectiveand commonly used sources of communication. The reason of the popularity of email system lies in its cost effective and faster communication nature. Unfortunately, email system is getting threatened by spam emails.Spam emails are the uninvited emails sent by some unwanted users also known as spammers with the motiveof making money. The emails users spend most of their valuable time in sorting these spam mails. Multiple copies of same message are sent many times which not only affect an organization financially but also irritatesthe receiving user. Spam emails are not only intruding the user’s emails but they are also producing large amount of unwanted data and thus affecting the network’s capacity and usage. In this paper, a Spam Mail Detection (SMD) system is proposed which will classify email data into spam and ham emails. The process of spam filtering focuses on three main levels: the email address, subject and content of the message. All mails have a common structure i.e. subject of the email and the body of the email. A typical spam mail can be classified by filtering its content. The process of spam mail detection is based on the assumption that the content of the spam mail is different than the legitimate or ham mail. For example words related to the advertisement of any product, endorsement of services, dating related content etc. The process of spam emaildetection can be broadly categorized into two approaches: knowledge engineering and machine learning approach. Knowledge engineering is a network based approach in which IP (internet protocol) address, network address along with some set of defined rules are considered for the email classification. The approachhas shown promising results but it is very time consuming. The maintenance and task of updating rules is notconvenient for all users. On the other hand, machine learning approach does not involve any set of rules andis efficient than knowledge engineering approach. The classification algorithm classifies the email based on the content and other attributes. For most of the classification problems the process of feature extraction and selection is very important. Features play a vital role in the process of classification. In this paper, a correlationbased feature selection (CFS) method is used for feature extraction. The CFS approach extracts the best features from the pool of features for efficient classification results. In order to remove the drawbacks of current model a novel hybrid bagged technique is introduced in the proposed spam mail detection (SMD) system. The proposed spam mail detection system is inspired from the effectiveness of machine learning approach. In spam mail detection system, initially email data is collected. The email data collected is raw and unstructured in nature. In order to reduce the computations and to obtain accurate results,email data needs to be pre-processed.

LITERATURE SURVEY

Email system is one of the most common and popular communication systems. Organizations from all over the world are making their efforts in order to identify the spam mails. The work of authors to identify the ham and spam emails is discussed here. Table 1 illustrates the comparative work of authors by stating the classification techniques, dataset, feature extraction approaches and drawbacks. In order to classify the email as spam, a filtering technique is required for its classification. Mohammad and Selamat have proposed a spam email filtering system using two different features selection methods to classify the emails.They haveconsidered English and Malay email dataset and after the pre-processing of the dataset features are selected using TF-IDF and rough set theory method. Then machine learning technique is applied for the classification purpose showing some reasonably good results. Another machine learning algorithm based work for the classification of email data was presented by Harisinghaney et al.. The algorithmic implementation includesKNN, Naïve Bayes and DBSCAN algorithms and showing effective results when the algorithms are appliedon pre-processed data. Further, Youn and Mcleod proposed an ontology based email filtering method. The considered dataset is classified using J48 decision tree based algorithm. A RDF language based ontology is created by Jena in order to test the results obtain after the classification. Authors have also adapted the optimization techniques. Faris et al. have used feed forward Decision Tree based method to detect the spam emails and to optimize the results as well. The Decision Tree is trained with the help of Krill Herd algorithm.The pre-processed dataset is equally divided into two halves for the training and testing purpose. The optimized classification results obtained from Decision Tree are compared with other optimization algorithms like Genetic algorithm and Back propagation. The experimental results shown by Kill Herd algorithm are more accurate than the other two algorithms. Another optimization based system is proposed by Al-Shboul et al. for the detection of spam mails. The authors have considered a hybrid approach for the email filtration process. In the first phase, Particle Swarm Optimization based algorithm is considered in order to select the best and optimized features. In the second phase, Random forest algorithm is trained with the selected features form the previous phase in order to classify the email dataset into ham and spam emails.

TEXT CLASSIFICATION: Text classification is also known as text-tagging and text categorization, it is a process in which text which can be unstructured or structured is classified into organized groups and according to the requirements. Various machine learning models use Natural Language Processing to analyze text and perform other operation and then assign them group based on their content. Figure depicts the text classification flowchart and the steps involved

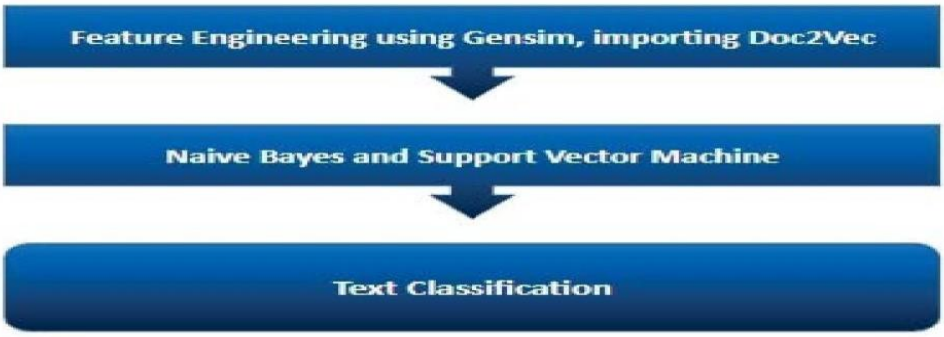

Feature Engineering using Gensim, importing Doc2Vec
Naive Bayes and Support Vector Machine
Text Classification

Fig 1 Text classification

NATURAL LANGUAGE PROCESSING: Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP combines computational linguistics rule based modeling of human language with statistical, machine learning, and deep learning models.

Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chat bots, and other consumer conveniences. But NLP also plays a growing role inenterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-

EXISTING SYSTEM

Due to the increase in the number of email users, the amount of spam emails have also risen in number in the past years. It has now become even more challenging to handle a wide range of emails for data mining and machine learning. Therefore, many researchers have executed comparative studies to see various classification algorithms performances and their results in classifying emails accurately with the help of a number of performance metrics.

Hence, it is important to find an algorithm that gives the best possible outcome for any particular metric for correct classification of emails and spam or ham. The present systems of spam detection are reliant on three major methods:- Linguistic Based Methods: Unlike humans, who can grasp linguistic constructs along with their exposition,machines cannot and hence it is necessary to teach machines some languages to help them understand these constructs. This is the technique that is used in places like search engines in order to ascertain the next terms for suggestions to the user while they are typing their search. Sentences are divided into two Unigrams (wordstaken are one by one) and two Bigrams (words that are taken two at a time). Since this technique requires thatevery expression be remembered, this method is not feasible and also time-intensive. Behavior -Based Methods: This technique is Metadata-based. This approach requires that users generatea set of rules, and the users must have a thorough understanding of these rules. Since the attributes of spam change over time so the rules also need to be reformed from time to time. As a result, it still requires a humanto scrutinise the details and is majorly user-dependent.

PROPOSED SYSTEM:

The dataset is taken from Spam Assassin , non spam messages belong to easy ham and they should be easilydifferentiated from spam. Instead of using sophisticated and hybrid models, this study relies on relatively simple classification algorithms to solve this problem like Logistic Regression, Naive Bayes, and Support Vector Machine.

The concept of Decision Trees is also used to select the best activation function for spam detection. The dataset is in the form of TEXT files which are converted into plaintext during text pre processing. This paper has used two feature sets to find the most optimal feature set and respective models. In order to perform efficient operations, Compressed Sparse Row (CSR) is used to feed data to models.

Hence,the data is converted into a compressed sparse row matrix format for modeling. A perfect (or best) model should be the one that reduces under fitting or overfitting. There are three practices for identification. They are datasets splitting, cross-validation, and bootstrap. In proposed work to prevent underfitting and overfitting, the modeling results will be evaluated first through a 10-fold cross-validation score, and then evaluated by evaluation metrics of classification.

SYSTEM DESIGN:

DATA FLOW:

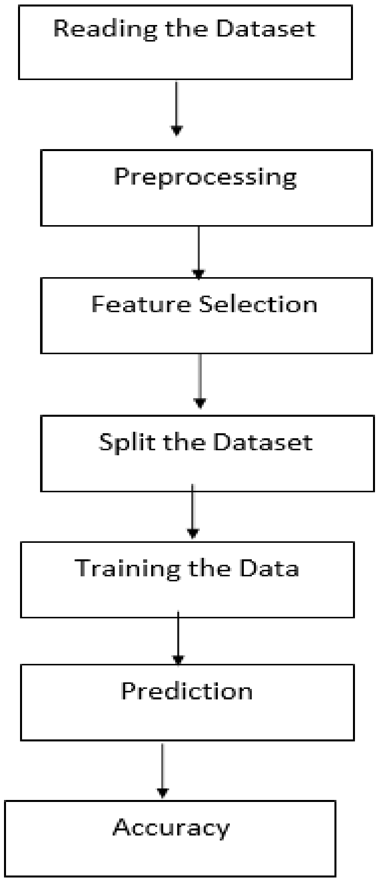
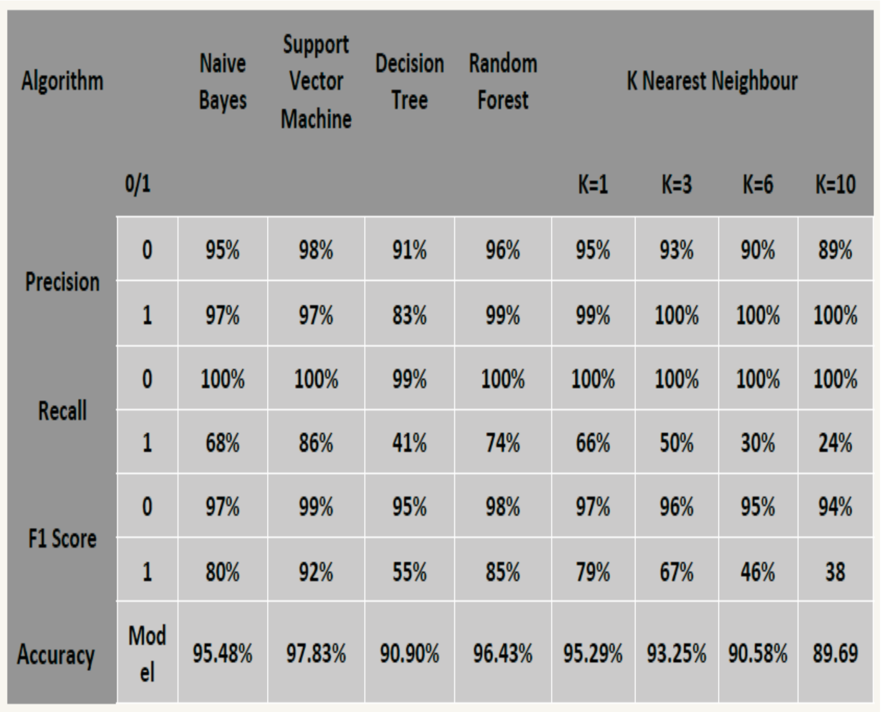
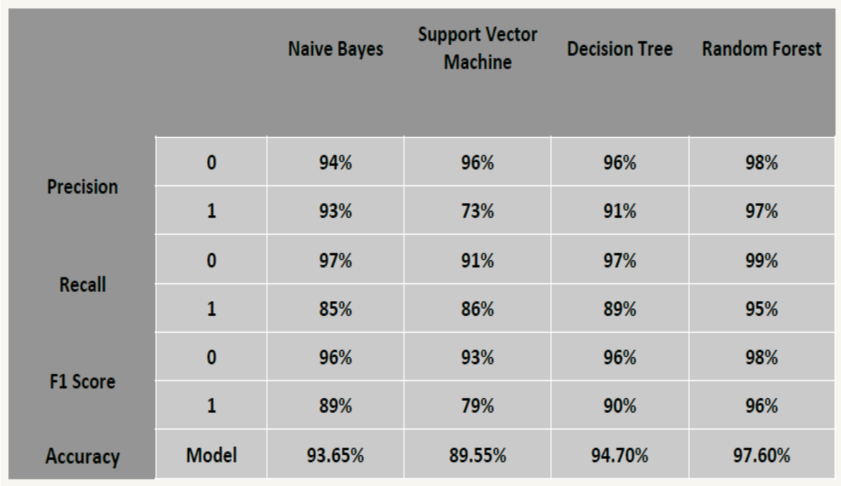


Fig 2: Data flow



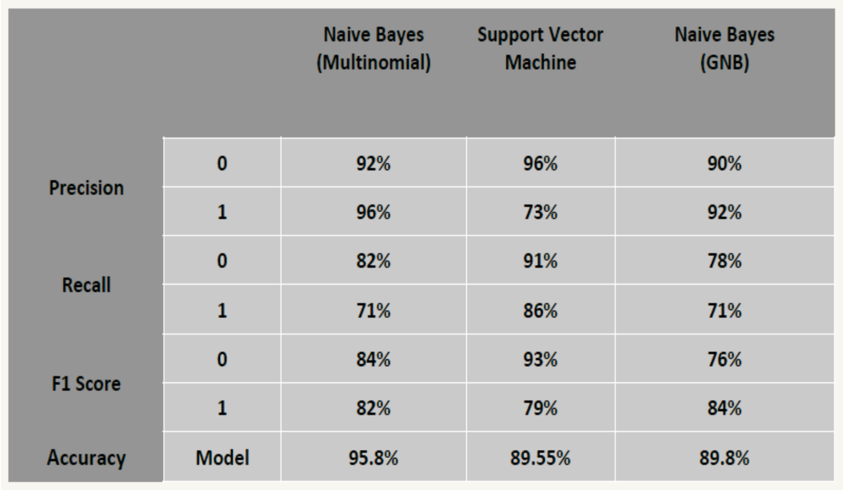
Algorithm		Naive Bayes	Support Vector Machine	Decision Tree	Random Forest	K Nearest Neighbour			
	0/1					K=1	K=3	K=6	K=10
Precision	0	95%	98%	91%	96%	95%	93%	90%	89%
	1	97%	97%	83%	99%	99%	100%	100%	100%
Recall	0	100%	100%	99%	100%	100%	100%	100%	100%
	1	68%	86%	41%	74%	66%	50%	30%	24%
F1 Score	0	97%	99%	95%	98%	97%	96%	95%	94%
	1	80%	92%	55%	85%	79%	67%	46%	38
Accuracy	Model	95.48%	97.83%	90.90%	96.43%	95.29%	93.25%	90.58%	89.69

Comparative Tabular Analysis 1



		Naive Bayes	Support Vector Machine	Decision Tree	Random Forest
Precision	0	94%	96%	96%	98%
	1	93%	73%	91%	97%
Recall	0	97%	91%	97%	99%
	1	85%	86%	89%	95%
F1 Score	0	96%	93%	96%	98%
	1	89%	79%	90%	96%
Accuracy	Model	93.65%	89.55%	94.70%	97.60%

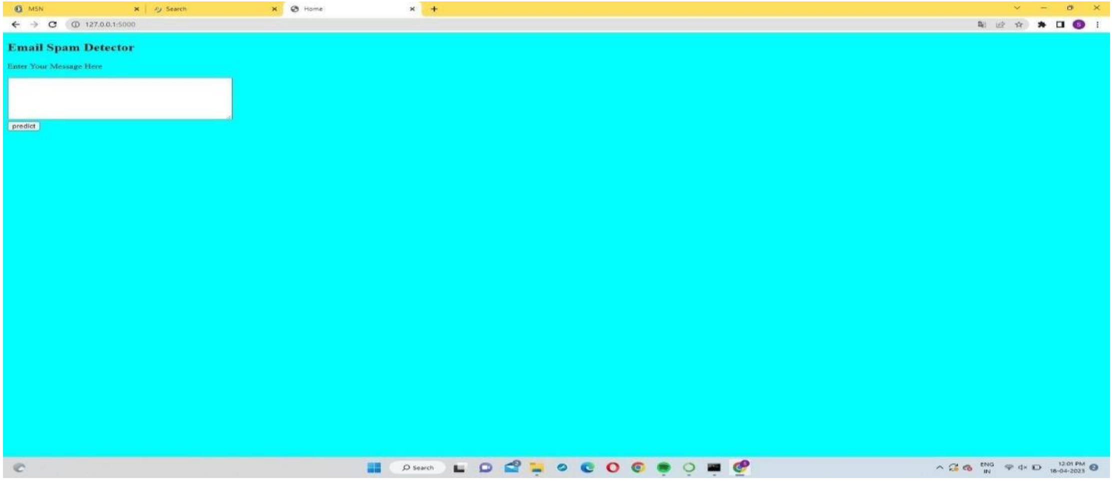
Comparative Tabular Analysis 2



		Naive Bayes (Multinomial)	Support Vector Machine	Naive Bayes (GNB)
Precision	0	92%	96%	90%
	1	96%	73%	92%
Recall	0	82%	91%	78%
	1	71%	86%	71%
F1 Score	0	84%	93%	76%
	1	82%	79%	84%
Accuracy	Model	95.8%	89.55%	89.8%

Comparative Tabular Analysis 3

RESULTS AND DISCUSSIONS

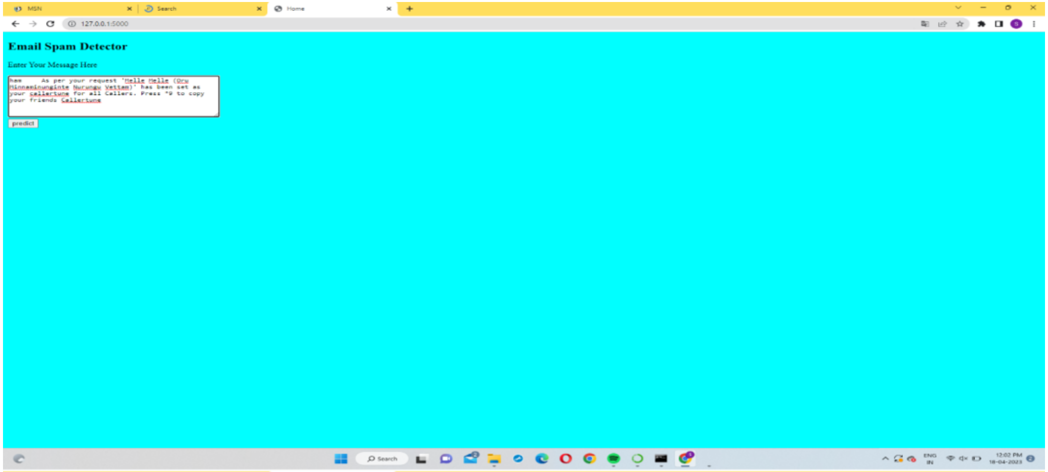


Email Spam Detector

Enter Your Message Here

Fig3: Browser view

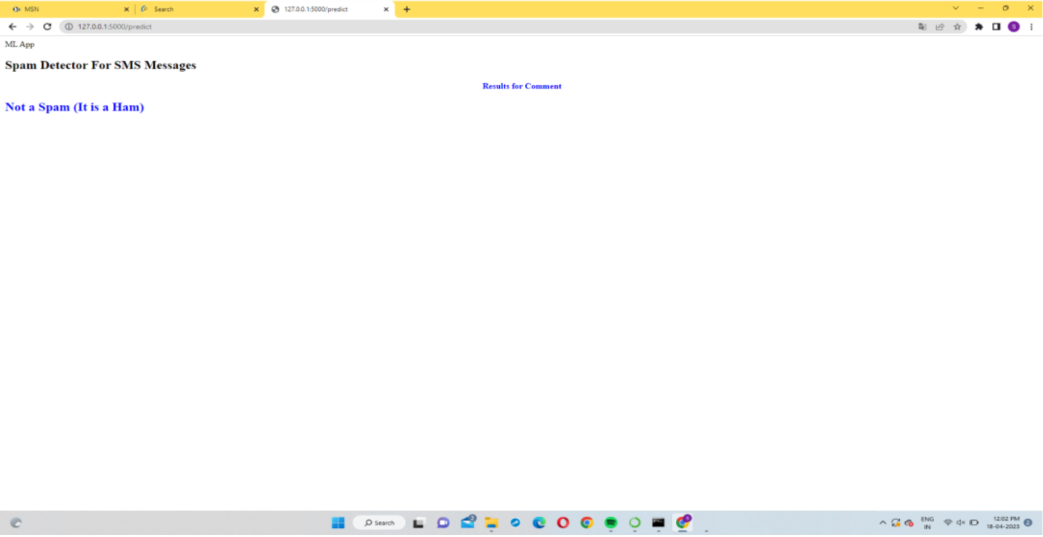
The message we got in email will be entered into the Spam detector and it will verify whether the received message is spam or ham. It will show the display as not a Spam as shown in above output result.



Email Spam Detector

Enter Your Message Here

ham As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune



ML App

Spam Detector For SMS Messages

Results for Comment

Not a Spam (It is a Ham)

SAMPLE DATASETS

The csv file contains 5172 rows, each row for each email. There are 3002 columns. The first column indicatesEmail name. The name has been set with numbers and not recipients' name to protect privacy. The last columnhas the labels for prediction: 1 for spam, 0 for not spam. The remaining 3000 columns are the 3000 most common words in all the emails, after excluding the non-alphabetical characters/words. For each row, the count of each word (column) in that email (row) is stored in the respective cells. Thus, information regarding all 5172 emails are stored in a compact data frame rather than as separate text files..

Category	Message
Ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there gotamore wat...
Ham	Ok lar... Joking wif u oni...
Spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receiveentry question(std txt rate)T&C's apply 08452810075over18's
Ham	U dun say so early hor... U c already then say...
Ham	Nah I don't think he goes to usf, he lives around here though
Spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up forit still? Tb ok! XxX std chgs to send, Â£1.50 to rcv

Ham	Even my brother is not like to speak with me. They treat me like aids patent.
Ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as yourcallertune for all Callers. Press *9 to copy your friends Callertune
Spam	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward!To claim call 09061701461. Claim code KL341. Valid 12 hours only.
Spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles withcamera for Free! Call The Mobile Update Co FREE on 08002986030
Ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've criedenough today.
Spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word:CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take yourhelp for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or clickhere>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatÂ’s the way u feel. ThatÂ’s the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 egENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ãº1.20 POBOXox36504W45WQ 16+
ham	Is that seriously how you spell his name?
ham	Iâ€˜m going to try for 2 months ha ha only joking
ham	So Ã¼ pay first lar... Then when is da stock comin...
ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
ham	Fffffffff. Alright no way I can meet up with you sooner?
ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. Heknows I'm sick when I turn down pizza. Lol
ham	Lol your always so convincing.
ham	Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom'sleft over dinner ? Do you feel my Love ?
ham	I'm back &amp; we're packing the car now, I'll let you know if there's room
ham	Ahhh. Work. I vaguely remember that! What does it feel like? Lol
ham	Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why xdoesn't want to live with us
ham	Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and hegot caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You?
ham	K tell me anything about you.
ham	For fear of fainting with the of all that housework you just did? Quick have a cuppa

spam	Thanks for your subscription to Ringtone UK your mobile will be charged Â£5/month Pleaseconfirm by replying YES or NO. If you reply NO you will not be charged
ham	Yup... Ok i go home look at the timings then i msg Ã¼ again... Xuhui going to learn on 2nd may toobut her lesson is at 8am
ham	Oops, I'll let you know when my roommate's done
ham	I see the letter B on my car
ham	Anything lor... U decide...
ham	Hello! How's you and how did saturday go? I was just texting to see if you'd decided to doanything tomo. Not that i'm trying to invite myself or anything!
ham	Pls go ahead with watts. I just wanted to be sure. Do have a great weekend. Abiola
ham	Did I forget to tell you ? I want you , I need you, I crave you ... But most of all ... I love you mysweet Arabian steed ... Mmmmmm ... Yummy
spam	07732584351 - Rodger Burns - MSG = We tried to call you re your reply to our sms for a free nokiamobile + free camcorder. Please call now 08000930705 for delivery tomorrow
ham	WHO ARE YOU SEEING?
ham	Great! I hope you like your man well endowed. I am &lt;#&gt; inches...
ham	No calls..messages..missed calls
ham	Didn't you get hep b immunisation in nigeria.
ham	Fair enough, anything going on?
ham	Yeah hopefully, if tyler can't do it I could maybe ask around a bit
ham	U don't know how stubborn I am. I didn't even want to go to the hospital. I kept telling Mark I'mnot a weak sucker. Hospitals are for weak suckers.
ham	What you thinked about me. First time you saw me in class.
ham	A gram usually runs like &lt;#&gt; , a half eighth is smarter though and gets you almost a wholesecond gram for &lt;#&gt;
ham	K fyi x has a ride early tomorrow morning but he's crashing at our place tonight
ham	Wow. I never realized that you were so embarassed by your accomodations. I thought you liked it,since i was doing the best i could and you always seemed so happy about "the cave". I'm sorry I didn't and don't have more to give. I'm sorry i offered. I'm sorry your room was so embarassing.
spam	SMS. ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice Hockey. Correct orIncorrect? End? Reply END SPTV
ham	Do you know what Mallika Sherawat did yesterday? Find out now @ &lt;URL&gt;
spam	Congrats! 1 year special cinema pass for 2 is yours. call 09061209465 now! C Suprman V, Matrix3,StarWars3, etc all 4 FREE! bx420-ip4-5we. 150pm. Dont miss out!
ham	Sorry, I'll call later in meeting.
ham	Tell where you reached
ham	Yes..gauti and sehwag out of odi series.
ham	Your gonna have to pick up a $1 burger for yourself on your way home. I can't even move. Pain iskilling me.
ham	Ha ha ha good joke. Girls are situation seekers.
Ham	Its a part of checking IQ
Ham	Sorry my roommates took forever, it ok if I come by now?

ham	Ok lar i double check wif da hair dresser already he said wun cut v short. He said will cut until ilook nice.
spam	As a valued customer, I am pleased to advise you that following recent review of your Mob No.you are awarded with a Â£1500 Bonus Prize, call 09066364589
ham	Today is "song dedicated day.." Which song will u dedicate for me? Send this to all ur valuablefrnds but first rply me...
spam	Urgent UR awarded a complimentary trip to EuroDisinc Trav, Aco&Entry41 Or Â£1000. To claimtxt DIS to 87121 18+6*Â£1.50(moreFrmMob. ShrAcomOrSglSuplt)10, LS1 3AJ
spam	Did you hear about the new "Divorce Barbie"? It comes with all of Ken's stuff!
ham	I plane to give on this month end.
ham	Wah lucky man... Then can save money... Hee...
ham	Finished class where are you.
ham	HI BABE IM AT HOME NOW WANNA DO SOMETHING? XX
ham	K..k:)where are you?how did you performed?
ham	U can call me now...
ham	I am waiting machan. Call me once you free.
ham	Thats cool. i am a gentleman and will treat you with dignity and respect.
ham	I like you peoples very much:) but am very shy pa.
ham	Does not operate after &lt;#&gt; or what
ham	Its not the same here. Still looking for a job. How much do Ta's earn there.
ham	Sorry, I'll call later
ham	K. Did you call me just now ah?
ham	Ok i am on the way to home hi hi
ham	You will be in the place of that man
ham	Yup next stop.
ham	I call you later, don't have network. If urgnt, sms me.
ham	For real when u getting on yo? I only need 2 more tickets and one more jacket and I'm done. Ialready used all my multis.
ham	Yes I started to send requests to make it but pain came back so I'm back in bed. Double coins atthe factory too. I gotta cash in all my nitros.
ham	I'm really not up to it still tonight babe
ham	Ela kano.,il download, come wen ur free..
ham	Yeah do! Donâ€˜t stand to close tho- youâ€˜ll catch something!
ham	Sorry to be a pain. Is it ok if we meet another night? I spent late afternoon in casualty and that means i haven't done any of y stuff42moro and that includes all my time sheets and that. Sorry.
ham	Smile in Pleasure Smile in Pain Smile when trouble pours like Rain Smile when sum1 Hurts U Smilebecoz SOMEONE still Loves to see u Smiling!!
spam	Please call our customer service representative on 0800 169 6031 between 10am-9pm as youhave WON a guaranteed Â£1000 cash or Â£5000 prize!
ham	Havent planning to buy later. I check already lido only got 530 show in e afternoon. U finish workalready?
spam	Your free ringtone is waiting to be collected. Simply text the password "MIX" to 85069 to verify.Get Usher and Britney. FML, PO Box 5249, MK17 92H. 450Ppw 16
ham	Watching telugu movie..wat abt u?

ham	i see. When we finish we have loads of loans to pay
ham	Hi. Wk been ok - on hols now! Yes on for a bit of a run. Forgot that i have hairdressers appointment at four so need to get home n shower beforehand. Does that cause prob for u?"

CONCLUSION: A comprehensive and efficient spam classification system has been created which follows a two step methodology to completely ensure that the mail received is spam or not. Initially, text classification takes place which is followed by URL analysis and filtering in order to determine ifany link present in the mail is malicious or not. For text classification, five machine learning algorithms were studied and analyzed, out of which Naive Bayes and Support Vector Machine having the highest accuracy were included in the final model. Various data-sets have been referredto for a list of spam trigger words and a list of blacklisted URLs. This model was hosted as an API which was then called by the JavaScript code in the google apps script in order to classify mails in real time in Gmail.

FUTURE ENHANCEMENT: Further research in this topic can be done across various sub-domains. Initially, the focus can be on improving accuracy by using some more computationally expensive but accurate machine learning classifiers like XGBoost. Further more, different word embedding algorithms other than Gensim word2Vec can be explored. Research in the field of deep learning could include transformer based deep learning models which was introduced in 2017. It enables training on humongous data sets, and also includes pre-trained systems which are used for text summarizationand translation. Lastly, real time learning of email classifiers is something which the current data-sets do not focus on. It is important because real time factors play a huge role in determining the classification accuracy.

REFERENCES

[1] AKINYELU, A. A., & ADEWUMI, A. O. (2014). “Classification of phishing email using random forest machine learning technique”. Journal of Applied Mathematics.

[2] Vinodhini. M, Prithvi. D, Balaji. S “Spam Detection Framework using ML Algorithm” inIJRTE ISSN: 2277- 3878, Vol.8 Issue.6, March 2020.

[3] YUsKSEL, A. S., CANKAYA, S. F., & UsNCUs, It. S. (2017). “Design of a Machine Learning Based Predictive Analytics System for Spam Problem.” Acta Physica Polonica, A.,132(3).[26] GOODMAN, J. (2004, July). “IP Addresses in Email Clients.” In CEAS.

[4] Deepika Mallampati, Nagaratna P. Hegde “A Machine Learning Based Email Spam Classification Framework Model” in IJITEE, ISSN: 2278-3075, Vol.9 Issue.4, February 2020.

[5] Javatpoint, “Machine Learning Tutorial” 2017 https://www.javatpoint.com/machine-learning

[6] SpamAssassin, “Spam and Ham Dataset", Kaggle, 018.https://www.kaggle.com/veleon/ham-and-spam-dataset

[7] Apache, “open-source Apache SpamAssassin Dataset”, 2019 https://spamassassin.apache.org/old/publiccorpus/

[8] SpamAssassin, “Spam Classification Kernel”, 2018 https://www.kaggle.com/veleon/spam-classification

[9] SpamAssassin, “REVISION HISTORY OF THIS CORPUS”, 2016 https://spamassassin.apache.org/old/publiccorpus/read me.html

[10] Jason Brownlee, “Naive Bayes for Machine Learning” The Machine Learning Mastery,

April 11, 2015. https://machinelearningmastery.com/naive-bayes-formachine- learning/

[11] Wikipedia, “History of email spam,” Internet Free Encyclopedia, 2001. https://en.wikipedia.org/wiki/History_of_email_spam

[12] E.G.Dada, J.S.Bassi, H.Chiroma, S.M.Abdulhamid, A.O.Adetunmbi, O.E.Ajibuwa, “ Machine learning for email spam filtering: re- view, approaches and open research problems”, Heliyon (2019) DOI:doi.org/10.1016/j.heliyon.2019.e01802

[13] 455 Spam Trigger Words to Avoid in 2019, accessed 3 November 2020, https://prospect.io/blog/455-email-spam-trigger-words-avoid-2018/

[14] PhishTank, accessed 3 November 2020, https://www.phishtank.com/

[15] Word2vec skip gram and cbow, accessed 3 November 2020, https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and- cbow-93512ee24314

[16] Asif Karim, Sami Azam, Bharanidharan Shanmugam, Krishnan Kan- noorpatti, and Mamoun Alazab - “A Comprehensive Survey for Intelli- gent Spam Email Detection”, College of Engineering, IT and Environ- ment, Charles Darwin University, Casuarina, NT 0810, Australia.

[17] Two Simple Adaptations of Word2Vec for Syntax Problems - Scientific Figure on ResearchGate, accessed 3 November 2020, https://www.researchgate.net/figure/Illustration-of-the-Skip-gram-and-Continuous-Bag-of-Word-CBOW-modelsfig1281812760

[18] Enron Spam data set accessed on 3 November 2020, http://nlp.cs.aueb.gr/softwareanddatasets/Enron-Spam/index.html

[19] Kaggle data set accessed on 3 November 2020, https://www.kaggle.com/uciml/sms-spam-collection-dataset

[20] Y. Lin and J. Wang, ”Research on text classification based on SVM- KNN,” 2014 IEEE 5th International Conference on Software Engineer- ing and Service Science, Beijing, 2014, pp. 842-844, doi: 10.1109/IC-SESS.2014.6933697.