RESEARCH ARTICLE



Application of Nave Bayes Algorithm for SMS Spam Classification Using Orange

Nosiel1*; Sigit Andriyanto2; Muhammad Said Hasibuan3

Published online: 10 January 2022

Abstract

Mobile phones have become a necessity for everyone. SMS is a communication service that is used to send and receive short messages in the form of text on mobile phones. Among all the advantages of SMS, there is a very annoying activity called spam (unsolicited commercial advertisements). Spam is the continuous use of electronic devices to send messages. called spammers. Spam messages are sent by advertisers with the lowest operating costs. Therefore, there are a lot of spammers and the number of messages requested is huge. Therefore, many aspects are harmed and disturbed. When SMS enters the user's mobile device, this study aims to classify spam and ham SMS. SMS classification adopts naive Bayes method. By looking at the contents of the SMS, the application of the naive Bayes method in data mining can distinguish unwanted SMS from non-spam. Results The classification accuracy rate is 0.999%. Based on the research that I have done, the Naive Bayes method can classify 1000 SMS spam data contained in the SMS spam data set file correctly.

Keyword: Classification; SMS; Spam; Ham; Naive Bayes

Introduction

In today's technological era, mobile phones are very common and are owned by almost all lower, middle and upper classes. Even in practice, mobile phones have become a secondary necessity for various circles of society. Sms (Short Message Service) is a mobile function, a very basic and necessary function on a mobile device. This functionality is considered as one of the common types of services because it is very cheap and easy to use for all mobile users, but among all the advantages of SMS, there is a very annoying activity called spam (unsolicited commercial advertisements). Spam is a type of message sending continuously through electronic devices, the message is considered annoying and unimportant without the permission of the recipient. People who send spam are called spammers. Spam emails are sent by advertisers with the lowest operating costs, so there are many spammers and the number of emails requested is very high, so many people feel disadvantaged and uncomfortable with this. Naive Bayes method is a classification method based on Bayes' theorem. The Naive Bayes method uses statistical and probability methods to generate categorical data to predict the future based on the past. The data set comes from <u>https://www.kaggle.com/shravan3273/sms-spam</u>.

In this study, the method to solve the problem of spam messages can use classification technology that groups between spam and ham messages when the message enters the user's mobile device. SMS classification uses the Naive Bayes method when viewing SMS content. The application of the naive bayes method in data mining is expected to be able to distinguish spam or ham messages.

Theoretical Framework

A. SMS (Short Message Service)

Short message service, or better known as SMS (Short Message Service), is a cellular telephone service that allows users to exchange written messages. SMS typically contain up to 160 characters and are used to send one-time messages, which can be sent wirelessly without the need for an Internet network.

Master of Informatics Engineering IBI Darmajaya Lampung, Indonesia

Email: nosiel.nosiel.2021211026@mail.darmajaya.ac.id

B. Spam SMS

There are two characteristics of spam messages: the first is unsolicited messages and the second is the large number of messages sent (sent in groups). The characteristics of SMS spam are very different from spam. The amount and content of spam is clearly different. In email, spam can be classified by subject, subject, sender, and usable characters. The capacity is greater than email. A message can only contain 160 characters.

¹⁻³ Master of Informatics Engineering IBI Darmajaya

^{*)} corresponding author

Nosiel

C. Data Mining

Data mining refers to the process of discovering previously unknown information from large data sets. In this process, data mining will extract valuable information by analyzing the presence of certain patterns or relationships in big data. Another definition of data mining is a set of processes that use one or more computer learning techniques to automatically analyze and extract knowledge, or a set of processes to explore additional value from data sets in the form of artificially unknown knowledge. Since data mining is a series of processes, data mining can be divided into several stages. This stage is interactive, with user participation directly or through a knowledge base. These stages are shown in Fig 1.

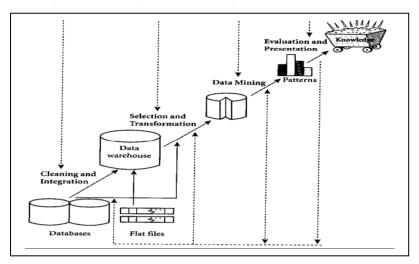


Fig 1. Data mining stages

D. Naive Bayes

Naive Bayes is a simple probability-based prediction technique, which is based on the application of Bayes' theorem (Bayes rule) and assumes strong independence. (Placettio, 2012).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}....(1)$$

Methods

Calculation of naive bayes classification using Orange tools/applications.

1. Dataset Preparation

The data used is a dataset available on the webset page <u>https://www.kaggle.com/shravan3273/sms-spam</u>.with the filename spam.tab

2. Corpus Viewers

Corpus Viewer to view the text file that we are using. The data is obtained after connecting the corpus with the corpus viewer as below.

- 3. *Data* Sampler
- At this stage select a subset of data from the input data set.

(2)

- 4. Text Preprocess
- Text preprocessing is the stage where we select data so that the data we process becomes more structured.
- 5. Bag of Words
- The function of the bag of words is to display text into a set of words that can be represented by their appearance in a document 6. Naive Bayes
 - Formulation using the nave Bayes method using the equations of Bayes' theorem:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Description:

X: unfinished class known H: Hypothesis data X is a specific class P(H|X): Probability of hypothesis H based on Condition X P(H): *Probability* hypothesis H P(X|H): *Probability* X based on condition On the hypothesis H P(X): *Probability* X process classification requires some hints for determine what label/class is match the sample Then the above formula is adjusted to be:

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$
(3)

- 7. Prediction
 - Forecasting is the result of estimating invisible data. AUC (Area Under the Curve)
 - a. The use of Prediction contained in the AUC model makes it easy to compare one model to another
 - b. CA (Classification Accuracy)
 - Calculate the classification accuracy level
 - c. F1

The average balance of precision and memory, the formula is:

- F1 = 2x (precision x ecall)/(precision + recall)(4)
- d. Precision
 - The classifier does not provide the ability to assign a positive label to a negative sample and vice versa. Precision has the following formula:
 - *Precision* = tp / (tp + fp) (5) Description:

Tp = number of true positives

Fp = number of false positives

e. *Recall*

Classifier has the ability to find and classify all samples with positive values.

8. Confusion Matrix

Performance measurement for data mining classification problems with two or more output classes.

9. ROC Analysis

Analyzing the results of the tool's performance measurement to classify the problem of determining model thresholds.

Results and Discussion

Orange is a data mining application that automatically calculates based on the widget that the researcher chooses.

1. Dataset Preparation

The data set used is called Spam.tab. This dataset has been prepared by orange application as training data...



Fig 2. New menu page for starting data mining calculations

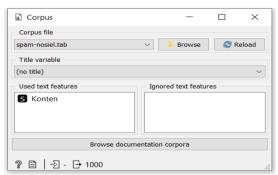


Fig 3. Selecting a dataset with the file name spam-nosiel.tab

2. Corpus Viewer

Corpus Viewerto see the text file we are using. The data is obtained after connecting the corpus with the corpus viewer as below.

File Edi	it View	Widget	Options	Help
Andrew College	ualize	_	_	
Constant Tree	CH2 Rule Induction	Calibrated Learner	kNN kNN	Corpus View
Linear	Forest Logistic Regression	Nalve Bayes	SVM SCA	Corpus
Neural Network	Stochastic Gradient D.	Stacking	Save Model	
Load Model			Fig 4	. Corpus Viewer

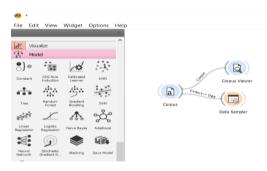
Double click then the corpus viewer form displays info, search features and display features.

Corpus Viewer						-	\times
	RegExp F	Filter:					
Preprocessed: False • Tokens: n/a	-	Document 2	^	Kategori: no:	spam 13.0		
- Types: n/a	3	Document 3		Konten:	URGENT! You have won membership in our å£100		Txt th
POS tagged: False	4	Document 4			word: CLAIM to No: 8101 LCCLTD POBOX 4403LD		et
V-grams range: 1-1 Matching: 1000/1000	5	Document 5					
Search features	6	Document 6					
C Kategori	7	Document 7					
S Konten	8	Document 8					
Display features	9	Document 9					
C Kategori	10	Document 10					
N no S Konten	11	Document 11					
Konten	12	Document 12					
Show Tokens & Tags	13	Document 13					
Auto send is on	14	Document 14	~				

Fig 5. Displaying info, search features and display features.

3. Data Sampler

At this stage select a subset of data from the input data set.





Double click then the sampler data form displays the sampling type

🔜 Data Sampler	?	\times								
Sampling Type Fixed proportion of	data:									
Fixed sample size		100 %								
Instances:	46	\$								
Sample with re	Sample with replacement									
 Cross validation 										
Number of subsets: 10										
Unused subset: 1										
O Bootstrap										
Options										
Replicable (determined)	inistic) s	ampling								
Stratify sample (wh	nen poss	sible)								
Sample Data										
? 🖹 → 1000 🕞 1000										
Fig 7. Showing the	samp	ling typ								

4. Preprocess Text

Text Preprocessing is the stage where we select the data so that the data that we will process becomes more structured.

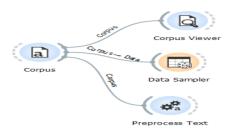


Fig 8. Preprocess Text

5. *Bag of Words* The function of the bag of words is to display text into a set of words that can be represented by their appearance in a document



Fig 9. Bag of Words

6. Naive Bayes

Enter and produce models and learners in the form of data and preprocessors.



Fig 10. Nave Bayes

7. *Prediction* Prediction is estimating the outcome for invisible data.

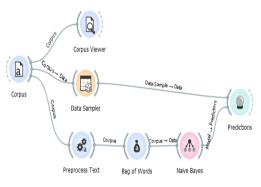


Fig 11. Prediction

Show protocolities for	-	1.2010.000				-
um		Nalve Bayes	Kategori	10	Korten	
span.	1		ham :		That's very rude, you on campus?	
	Z	1.00:0.00 - tam			Hata maybule rite. U know me well. Da feeling of being liked by someone is gd lor. U faster go find one then a.	
	3	1.00:0.00 - ham			, And got up, have to be out of the norm very soon, SuQ, i hadn't put the clode back til at 8 i shouled at everyone .	
	4	1.00:0.00 - hem			Sony me going home find Daddy come fetch (_ later	
	5	1.00:0.00 - ham			Bring home some Wendy =D	
	6	1.00:0.00 - ham			What is this her place you talk of? Esplain!	
	7	000:100-sp			New TEXTBLIDDY Chat 2 homy gaps in ur area 4 just 25p Free 2 receive Search portcode or at gaptextbuddy.com	
	8	1.00 : 0.00 - ham			Sine competition for him.	
	9.	1.00:0.00 - ham			Your brother is a genius	
	10	1.00:0.00 - hern			He is there. Nou call and meet film	
	11	1.00 : 0.00 - ham			I borrow ur bag ok.	
	12	1.00 : 0.00 - tam	han	77.0	Thats cool i am a gentieman and will treat you with dignity and respect.	
	3 13	1.00 : 0.00 - ham	hers	637.0	What all we do in the shower, baby?	
	14	1.00:0.00 - ham	tan		Hey no I ad a crap rite was born without ya 2 boggy with me u boring biaschi Thank but u wait til net time if ave y	ł
	15	1.00 0.00 - ham	han	939.0	No we sell it all so we'll have tons if coins. Then sell our coins to someone thru paypal. Volal Money back in Me	
	15	1.00 : 0.00 ~ ham			How is it possible to teach you. And where,	
	17	1.00:0.00 - ham	hers.	281.0	You got called a tool?	
	18	1.00 : 0.00 - tam	han	884.0	Gbbs unsold mile tussey	
	19	1.00:0.00 - ham	tan	762.0	Nothing but we just tot u would ask cos u ba gua But we went not faber yest Yest jus went out already main so	
	20	1.00 : 0.00 - ham	han	320.0	Coffee cake, i guess.	
	21	1.00:0.00 - ham	here	550.0	Huny up. fve been weed-deficient for like three-days	
	22	1.00:0.00 - harh	hart.	175.0	Well, I'm gonna finish my bath now. Have a goodfine night.	
			(c)			
	N	lodel AUC CA	FT Precisi	on Reca	1	
	Nai	e Baves 0.999 0.983	0.963 0.964	0.983		
Restane Original Order						

Fig 12. Output Prediction

Based on the calculation results can be:

- 1. AUC = 0.999
- 2. CA = 0.983
- 3. F1 = 0.983
- 4. Precision = 0.984
- 5. Recall = 0.983
- 8. Confusion Matrix

Based on the prediction results, the confusion matrix produces selected data and data.

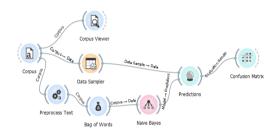


Fig 13. Confusion Matrix

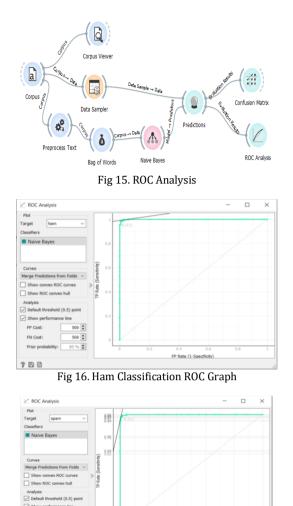
# Confusion Matrix								- 0	Х
Learners							Show:	Number of instance	es v
Naive Bayes				Pi	redicted				
				ham	spam	Σ			
		_	ham	831	16	847			
	>	Actual	spam	1	152	153			
			Σ	832	168	1000			
Predictions Probabilities									
Apply Automatically			Select Corre	et		Select Misclassified		Clear Selection	
? B B -									

Fig 14. Output Confusion Matrix

From the Confusion Matrix, there are data results, including:

- a. From a total of 832 ranked human rights data, there are 831 human rights data whose predictions are consistent with the actual data.
- b. From a total of 168 ranking spam data, there are 152 spam data whose predictions are consistent with the actual data.
- 9. ROC Analysis

ROC analysis was obtained from the evaluation results. Describe the relationship between sensitivity and specification for a naive Bayes model in the form of a line graph.



For performance like
 Fr Cast:
 Fr Fake (1-Specifich)
 Fr Fake (1-Spe

After making predictions, we test the AUC of the Naive Bayes method to get accuracy from the smallest to the largest, we use the Data Sampler and change the Sampling Type, such as Fixed Proportion Of data to 100%, we get AUC = 0.998, CA = 0.971, F1 = 0.972, Precision = 0.974, Recall = 0.971, then we change the Cross Validation number of subsets to 10 and unused subsets to 1 in getting AUC = 0.999, CA = 0.982, F1 = 0.983, Precision = 0.984, Recall = 0.982, then we change Bootstrap got AUC = 0.998, CA = 0.971, F1 = 0.972, Precision = 0.974, Recall = 0.971, in table 1 the comparison changed the Sampling Type as shown in the table below.

TABLE 1. Comparison of AUC changing Sampling Type

No.	Sampling Type	AUC	CA	F1	Precision	Recall
1.	Fixed Proportion Of data menjadi 100%	0.998	0.971	0.972	0.974	0.971
2.	Cross Validation number of subsets menjadi 10 dan unused subsets menjadi 1	0.999	0.982	0.983	0.984	0.982
3.	Bootstrap	0.998	0.971	0.972	0.974	0.971

Conclusions

The results of the research that have been carried out, it can be concluded that the naive Bayes method can correctly classify 1000 unwanted text messages contained in the spam.tab dataset file. The results of the classification precision are AUC = 0.999, CA = 0.982, F1 = 0.983, Precision = 0.984, Recall = 0.982, orange applications can use ROC analysis to visualize data in data mining through the naive Bayes method using line graphs. According to my research based on cross-validation of sampling type, the highest precision is 0.999%

Acknowledgment

We would like to express our gratitude to the lecturers, colleagues and the entire academic community of the Master of Informatics Engineering IBI Darmajaya.

References

- [1] Algorithms, Comparison, Nave Bayes, and DAN Decision Trees, 'Comparison of Nave Bayes, Svm, and Decision Tree Algorithms For Classification of Sms Spam', 05.02 (2020), 167–74
- [2] Apandi, Tri Herdiawan, and Castaka Agus Sugianto, 'COMPARATIVE ANALYSIS OF MACHINE LEARNING ON SMS SPAM DATA', 12.1 (2018)
- [3] Irmayani, Windi, and Keywords, 'DATA VISUALIZATION IN DATA MINING USING CLASSIFICATION METHOD Accepted: Published':, IX.I (2021), 68–72
- [4] Jaya, University of Banten, Study Program, Informatics Engineering, Faculty of Computer Science, and University of Banten Jaya, 'COMPARATIVE OF NAVE BAYES ALGORITHM AND SUPPORT VECTOR MACHINE (SVM)', 3.2 (2019), 178–94
- [5] Nasution, Firizqy Ramadhana, and Moch Arif Bijaksana, 'SMS Classification Spam Detection Using Artificial Immune System Algorithm and Apriori Frequent Itemset', 2013, 1–9
- [6] Octora, Vero Arneal, Moch Arif Bijaksana, and M Tech, 'SMS Spam Filtering Using Artificial Immune System (AIS) Method And Tokenization With Vectors Algorithm SMS Spam Filtering Using Artificial Immune System (AIS) Method and Tokenization With Vectors Algorithm', 2.3 (2015), 7838–45
- [7] Pranata, Eko Ardian, and Go Frendi Gunawan, 'Application of the Naïve Bayes Method for Classification of SMS Spam Using Java Programming', 07 (2019), 104–8
- [8] Pratama, Rio, and Ibnu Asror, 'SMS Spam Detection Using Vector Space Model Method With K-Means Clustering', 5.2 (2018)
- [9] Setiyono, Agus, Hilman F Pardede, and Masters in Computer Science, 'Classification of SMS Spam Using a Support Vector Machine', 15.2 (2019), 275–80 https://doi.org/10.33480/pilar.v15i2.693 >