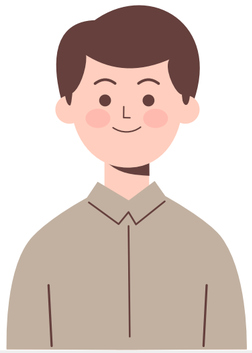
SPAM HAM EMAIL CLASSIFICATION

PEMROSESAN TEKS

KELOMPOK 6 (SAINS DATA 2022 A)

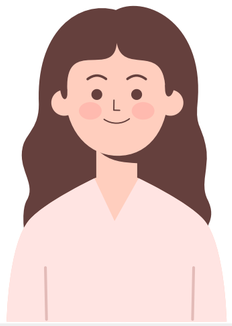
All About Team

Riva Dian Ardiansyah (043)



- Mengerjakan SVM (TD-IDF)
- Mengerjakan SVM (Word2Vec)
- Mengerjakan PPT Akhir
- Membuat User Guide

Analicia (007)

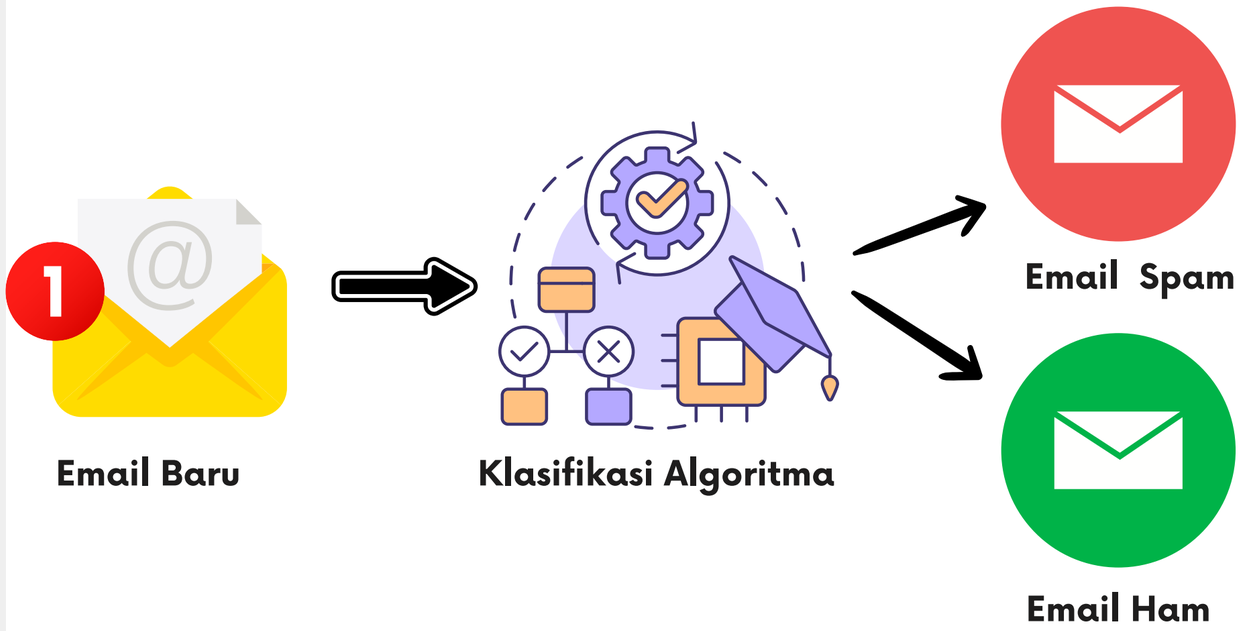


- Mengerjakan KNN (TD-IDF)
- Mengerjakan Naivve B (Word2Vec)
- Mengerjakan PPT Kemajuan Progres
- Membuat Laporan Akhir

Fadhilah Nuria Shinta (003)

- Mengerjakan Naive B (TD-IDF)
- Mengerjakan KNN (Word2Vec)
- Membuat GUI

About Project



Dalam projek kali ini, kami akan melakukan Klasifikasi Spam pada Email. Adapun beberapa tahap dalam Klasifikasi, yaitu Observing data, Data cleaning, Data preparation, Evaluation, Confusion matrix, dan Classification report

Projek kami tentang “Pengklasifikasian Spam pada E-mail”

Latar Belakang

Klasifikasi spam pada email adalah tindakan penting dalam menjaga keamanan komunikasi online. Dengan mengidentifikasi dan memblokir pesan spam, kita melindungi diri dari potensi penipuan, menjaga keamanan data pribadi, dan meningkatkan efisiensi dalam pengelolaan email. Namun nyatanya, seringkali pesan-pesan yang seharusnya bukan spam juga teridentifikasi sebagai spam, yang mana hal ini mengganggu pengalaman pengguna.

TUJUAN

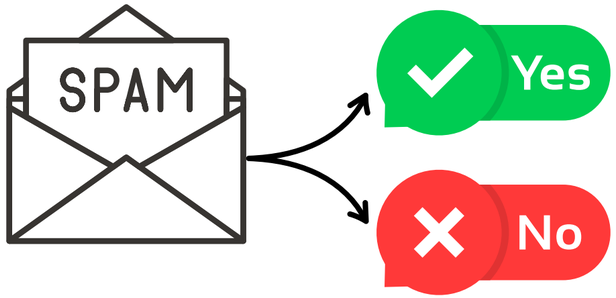
1. Membersihkan data email dari segala jenis noise, termasuk karakter khusus, tanda baca, atau informasi yang tidak relevan.
2. Mempersiapkan data email dengan cara yang memungkinkan model Word2Vec memahami konteks dan makna kata-kata.
3. Penggunaan model klasifikasi untuk mengkategorikan atau mengklasifikasikan email berdasarkan kriteria tertentu (spam dan non spam)
4. Menggambarkan hubungan antara kata-kata dalam ruang vektor, sehingga dapat digunakan untuk analisis lebih lanjut, termasuk pengelompokan atau pencarian pola.

Algoritma

KNN

SVM

Naive Bayes



Alasan memilih ketiga algoritma tersebut:

Pemilihan ketiga algoritma, KNN, SVM, dan Naive Bayes, dengan ekstraksi fitur TF-IDF dan Word2Vec untuk deteksi spam/ham email dapat disebabkan oleh keunggulan masing-masing metode dalam menangani permasalahan klasifikasi teks.

1. KNN (K-Nearest Neighbors): Cocok untuk tugas klasifikasi teks karena dapat menangani pola-pola lokal dalam data teks. Dengan TF-IDF atau Word2Vec, dapat memanfaatkan representasi vektor kata untuk menentukan kelas email.

2. SVM (Support Vector Machine): Efektif dalam menangani masalah klasifikasi biner dan dapat berkinerja baik dengan ruang fitur berdimensi tinggi seperti yang dihasilkan oleh TF-IDF atau Word2Vec.

3. Naive Bayes: Metode yang cepat dan sederhana, sering kali efektif untuk klasifikasi teks. Dengan TF-IDF atau Word2Vec, dapat memberikan estimasi probabilitas kelas berdasarkan kemunculan kata-kata.

Refernsi Artikel

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 1, February 2019, pp. 245~254
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i1.pp245-254

Spam image email filtering using K-NN and SVM

Yasmine Khalid Zamil, Suhad A. Ali, Mohammed Abdullah Naser
Department of Computer Science, College of Science for Women, University of Babylon, Iraq

Article Info

Article history:

Received Apr 19, 2018
Revised Sep 18, 2018
Accepted Okt 1, 2018

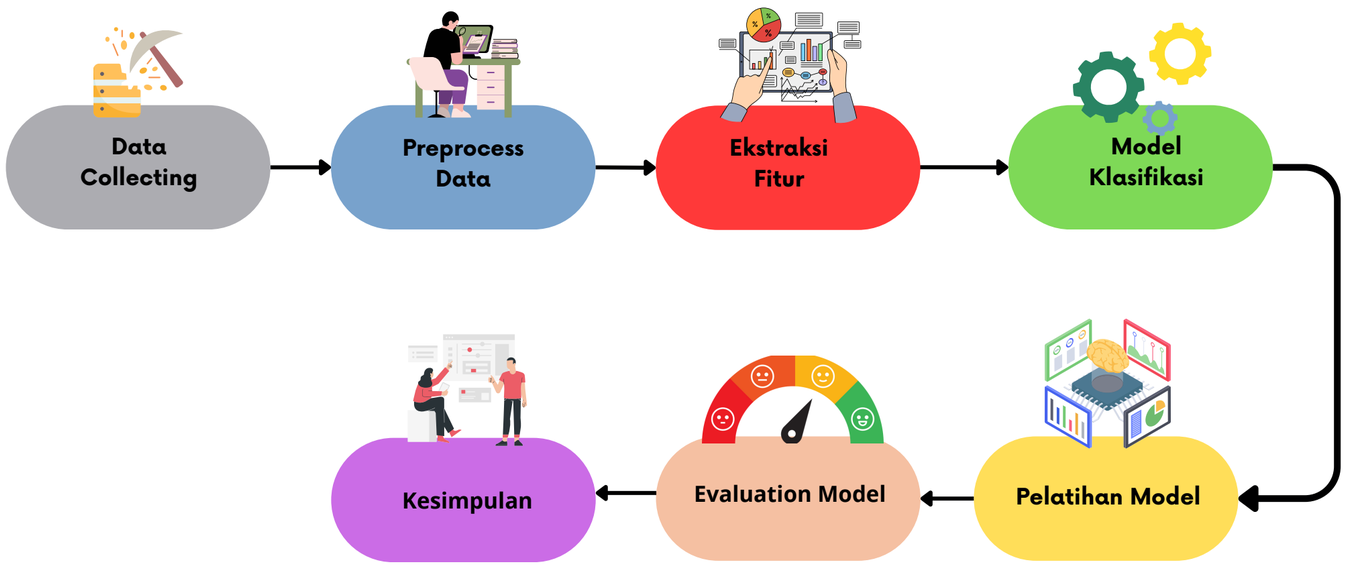
Keywords:

KNN
Spam filtering techniques
Spam image
SVM

ABSTRACT

The developing utilization of web has advanced a simple and quick method for e-correspondence. The outstanding case for this is e-mail. Presently days sending and accepting email as a method for correspondence is prominently utilized. Be that as it may, at that point there stand up an issue in particular, Spam mails. Spam sends are the messages send by some obscure sender just to hamper the improvement of Internet e.g. Advertisement and many more. Spammers introduced the new technique of embedding the spam mails in the attached image in the mail. In this paper, we proposed a method based on combination of SVM and KNN. SVM tend to set aside a long opportunity to prepare with an expansive information set. On the off chance that "excess" examples are recognized and erased in pre-handling, the preparation time could be diminished fundamentally. We propose a k-nearest neighbor (k-NN) based example determination strategy. The strategy tries to select the examples that are close to the choice limit and that are effectively named. The fundamental thought is to discover close neighbors to a question test and prepare a nearby SVM that jelly the separation work on the gathering of neighbors. Our experimental studies based on a public available dataset (Dredze) show that results are improved to approximately 98%.

METODE



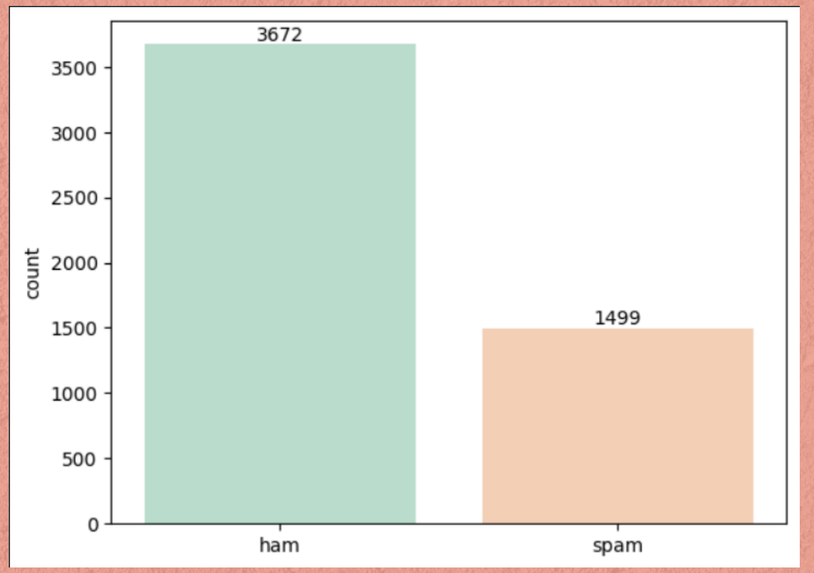
BEBERAPA METODE

Data Collecting

#	label	text	label_num
	Labels of Emails which can be either Spam or Ham	Emails data	if spam it's 1, or else it's 0
0 5170	ham 71% spam 29%	4993 unique values	0 1
605	ham	Subject: enron methanol ; meter # : 988291 this is a follow up to the note i gave you on monday , 4...	0
2349	ham	Subject: hpl nom for january 9 , 2001 (see attached file : hplnol 09 . xls) - hplnol 09 . xls	0
3624	ham	Subject: neon retreat ho ho ho , we ' re around to that most wonderful time of the year - - - neon ...	0

https://www.kaggle.com/code/nagasai524/spam-email-classification-using-word2vec

Column1	label	text	label_num
0	ham	Subject: christmas tree farm pictures	0
1	ham	Subject: vastar resources , inc .gary , production from the high island larger block a - 1 # 2 commenc	0
2	ham	Subject: calpine daily gas nomination- calpine daily gas nomination 1 . doc	0
3	ham	Subject: re : issuefyi - see note below - already done .stella- forwarded	0
4	ham	Subject: meter 7268 nov allocationfyi .- forwarded by lauri a allen / hou	0
5	ham	Subject: mcmullen gas for 11 / 99jackie ,since the inlet to 3 river plant is shut in on 10 / 19 / 99 (the	0
6	ham	Subject: meter 1517 - jan 1999george ,i need the following done :jan 13zero out 012 - 27049 - 02 - 0	0
7	ham	Subject: duns number changesfyi- forwarded by gary l payne / hou / ect	0
8	ham	Subject: king ranchthere are two fields of gas that i am having difficulty with in the unifysystem .1 . c	0
9	ham	Subject: re : entex transistionthanks so much for the memo . i would like to reiterate my support on	0
10	ham	Subject: entex transistionthe purpose of the email is to recap the kickoff meeting held on yesterdayv	0
11	ham	Subject: lst rev dec . 1999 josey ranch nomfyi- forwarded by susan d tre	0
12	ham	Subject: 2 nd rev dec . 1999 josey ranch nom- forwarded by susan d tre	0
13	ham	Subject: unify close schedulethe following is the close schedule for this coming month (year - end .)	0
14	ham	Subject: meter 1431 - nov 1999aimee ,sitara deal 92943 for meter 1431 has expired on oct 31 , 1999	0
15	ham	Subject: meter 1431 - nov 1999daren -could you please resolve this issue for howard ? i will be out c	0
16	ham	Subject: y 2 k - texas logname home pagergeorge grant 281 - 282 - 9084 713 - 764 - 5128charlotte ha	0
17	ham	Subject: re : lyondell citgomy latest understanding is the buyback will be in place through 12 / 31 / 99	0
18	ham	Subject: hpl fuel gas buy - back for december 1999fyi :- forwarded by g	0
19	ham	Subject: ua 4 - meter 1441 for 11 / 97 - falfurriasdaren - i need your help in resolving this issue . ther	0
20	ham	Subject: ua 4 for meter 8608 - 6 / 98 - deal 96731daren - deal 96731 is not in cpr for 6 / 98 or oss . p	0
21	ham	Subject: january spot ticketsdaren ,hplc is purchasing gas from the following list of producers . thispr	0
22	ham	Subject: pennzenergy property details- forwarded by ami chokshi / corp	0
23	ham	Subject: miscellaneous- forwarded by ami chokshi / corp / enron on 12	0
24	ham	Subject: re : purge of old contract _ event _ statusfyi - what do you all think ?- - - - - - - - - - - - - - - - - -	0
25	ham	Subject: out on vacationfyi ,i will be out of the office (actually out of the country) starting from22 r	0



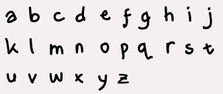
Dataset berisi 5.170 baris dan 4 kolom

Data Preprocessing

DATA Preprocessing

Pada bagian ini, file email mentah akan diubah menjadi teks biasa. Transformasi seperti menghapus header email, menghapus tanda baca, stemming kata-kata (lebih lanjut tentang ini nanti) dll... akan dilakukan

Lowercase



↓

Penghapusan karakter yang tidak perlu

↓

Penghapusan hashtag dan tag akun

↓

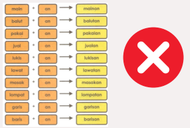
Penghapusan karakter selain angka dan huruf

↓

Penghapusan kata-kata sering muncul dan tidak bermakna

↓

Penghapusan imbuhan setiap kata



Sebelum di preprocessing

2	3624	ham	Subject: neon retreat ho ho ho , we ' re around to that most wonderful time of the year - - - neon leaders retreat time ! i know that this time of year is extremely hectic , and that it ' s tough to think about anything past the holidays , but life does go on past the week of december 25 through january 1 , and that ' s what i ' d like you to think about for a minute . on the calender that i handed out at the beginning of the fall semester , the retreat was scheduled for the weekend of january 5 - 6 . but because of a youth ministers conference that brad and dustin are connected with that week , we ' re going to change the date to the following weekend , january 12 - 13 . now comes the part you need to think about . i think we all agree that it ' s important for us to get together and have some time to recharge our batteries before we get to far into the spring semester , but it can be a lot of trouble and difficult for us to get away without kids , etc . so , brad came up with a potential alternative for how we can get together on that weekend , and then you can let me know which you prefer . the first option would be to have a retreat similar to what we ' ve done the past several years . this year we could go to the heartland country inn (www . . com) outside of brenham . it ' s a nice place , where we ' d have a 13 - bedroom and a 5 - bedroom house side by side . it ' s in the country , real relaxing , but also close to brenham and only about one hour and 15 minutes from here . we can golf , shop in the antique and craft stores in brenham , eat dinner together at the ranch , and spend time with each other . we ' d meet on saturday , and then return on sunday morning , just like what we ' ve done in the past . the second option would be to stay here in houston , have dinner together at a nice restaurant , and then have dessert and a time for visiting and recharging at one of our homes on that saturday evening . this might be easier , but the trade off would be that we wouldn ' t have as much time together . i ' ll let you decide . email me back with what would be your preference , and of course if you ' re available on that weekend . the democratic process will prevail - - majority vote will rule ! let me hear from you as soon as possible , preferably by the end of the weekend . and if the vote doesn ' t go your way , no complaining allowed (like i tend to do !) have a great weekend , great golf , great fishing , great shopping , or whatever makes you happy ! bobby	0

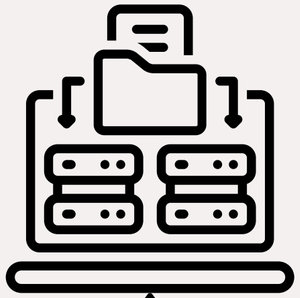
Sesudah di preprocessing

2	subject neon retreat around wonderful time year neon leaders retreat time know time year extremely hectic tough think anything past holidays life past week december january like think minute calender handed beginning fall semester retreat scheduled weekend january youth ministers conference brad dustin connected week going change date following weekend january comes part need think think agree important get together time recharge batteries get far spring semester lot trouble difficult get away without kids etc brad came potential alternative get together weekend let know prefer first option would retreat similar done past several years year could heartland country inn www com outside brenham nice place bedroom bedroom house side side country real relaxing also close brenham one hour minutes golf shop antique craft stores brenham eat dinner together ranch spend time meet saturday return sunday morning like done past second option would stay houston dinner together nice restaurant dessert time visiting recharging one homes saturday evening might easier trade would much time together let decide email back would preference course available weekend democratic process prevail majority vote rule let hear soon possible preferably end weekend vote way complaining allowed like tend great weekend great golf great fishing great shopping whatever makes happy bobby	0

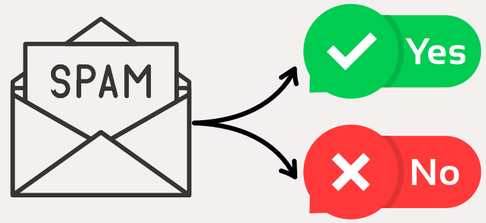
Ekstraksi Fitur

EKSTRAKSI FITUR

Proses mengubah data dokumen kedalam bentuk tertentu biasanya berbentuk numerik yang menampilkan bobot dari kata dalam suatu dokumen



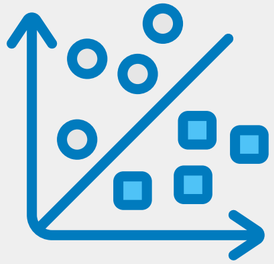
Ekstraksi fitur dengan Word2Vec dan TF-IDF



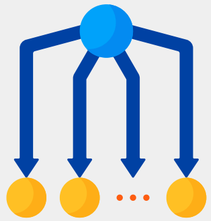
Model Klasifikasi

MODEL KLASIFIKASI

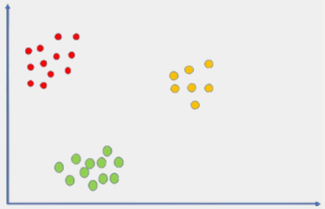
Modell klasifikasi yang sesuai dengan projek yaitu termasuk Naive Bayes, Support Vector Machines (SVM), dan KNN.



SVM



NAIVE BAYES

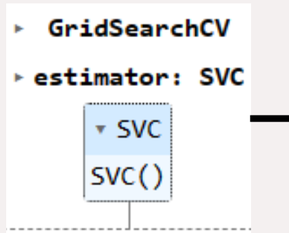


KNN

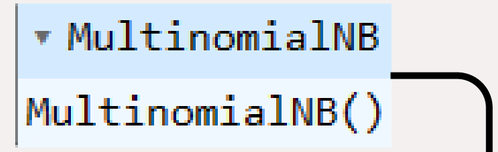
Pelatihan Model

Pelatihan Model

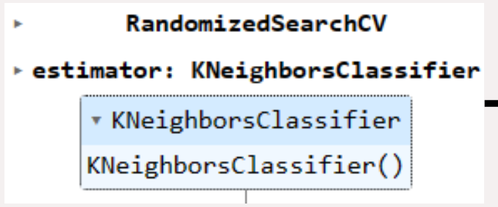
Parameter optimal melalui validasi silang atau pencarian grid (Gridsearch), MultinomialNB, RSCV.



SVM dengan SVC



Naive Bayes dengan MultinomialNB



KNN dengan RandomSearchCV

Evaluasi Model

Evaluasi Model

Evaluasi kinerja model menggunakan set pengujian dengan menggunakan metrik seperti presisi, recall, F1-score, dan akurasi. Pertimbangkan confusion matrix untuk lebih memahami kinerja model.

Komparasi Model

Akurasi KNN dengan ekstraksi fitur TF-IDF

KNN (TF-IDF)	0	1
PRECISSION	1.00	0.53
RECALL	0.65	0.99
ACCURACY	0.74	0.74
F1-SCORE	0.78	0.69
SUPPORT	742	293

Hasil dari KNN dengan Ekstraksi fitur menggunakan TF-IDF didapatkan akurasi 74%

Perbandingan Model KNN dengan Ekstraksi Fitur Word2Vec

KNN (Word2Vec)	0	1
PRECISSION	0.97	0.79
RECALL	0.90	0.93
ACCURACY	0.91	0.91
F1-SCORE	0.93	0.85
SUPPORT	735	300

Didapatkan best akurasi KNN Word2vec terbaik yakni K = 1, metric : euclidean dengan akurasi 91%

Komparasi Model

Perbandingan Model Naive Bayes dengan Ekstraksi Fitur TF-IDF

NAIVE BAYES (TF-IDF)	0	1
PRECISSION	0.90	1.00
RECALL	1.00	0.72
ACCURACY	0.92	0.92
F1-SCORE	0.95	0.84
SUPPORT	741	294

Hasil model Naive Bayes dengan Ekstraksi fitur menggunakan TF-IDF. menghasilkan akurasi 92%

Perbandingan Model Naive Bayes dengan Ekstraksi Fitur Word2Vec

NAIVE BAYES (Word2Vec)	0	1
PRECISSION	0.98	0.97
RECALL	0.99	0.94
ACCURACY	0.97	0.97
F1-SCORE	0.98	0.95
SUPPORT	732	303

Didapatkan akurasi Naive Bayes Word2vec terbaik yakni dengan akurasi 97%

Komparasi Model

Perbandingan Model SVM dengan Ekstraksi Fitur TF-IDF

SVM (TF-IDF)	0	1
PRECISSION	0.99	0.96
RECALL	0.98	0.98
ACCURACY	0.98	0.98
F1-SCORE	0.99	0.97
SUPPORT	735	300

Hasil model SVM dengan Ekstraksi fitur menggunakan TF-IDF didapatkan akurasi 98%

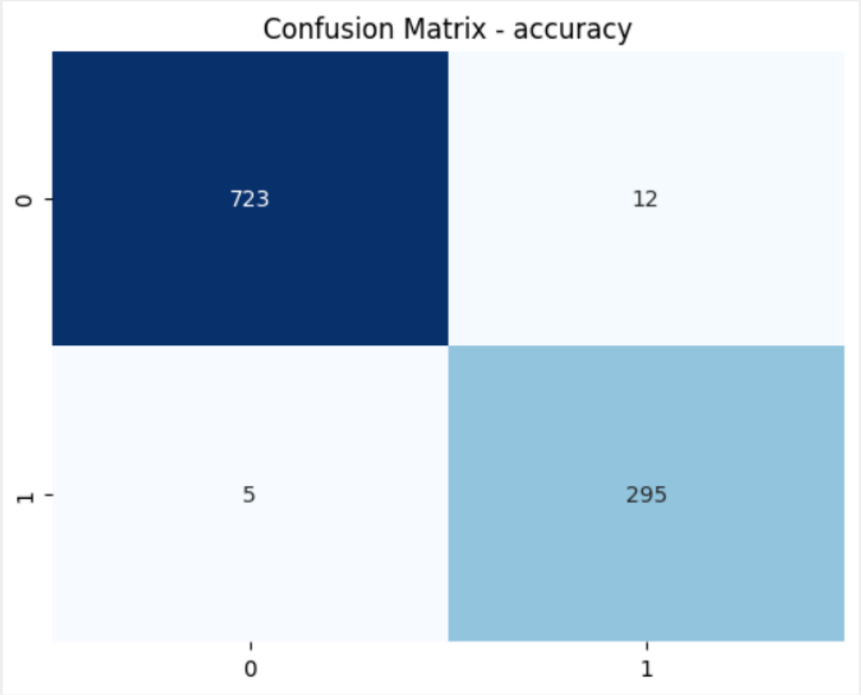
Perbandingan Model SVM dengan Ekstraksi Fitur Word2Vec

SVM (Word2Vec)	0	1
PRECISSION	0.97	0.73
RECALL	0.86	0.93
ACCURACY	0.88	0.88
F1-SCORE	0.91	0.82
SUPPORT	735	300

Didapatkan best akurasi SVM Word2vec yakni dengan akurasi 88%

KESIMPULAN

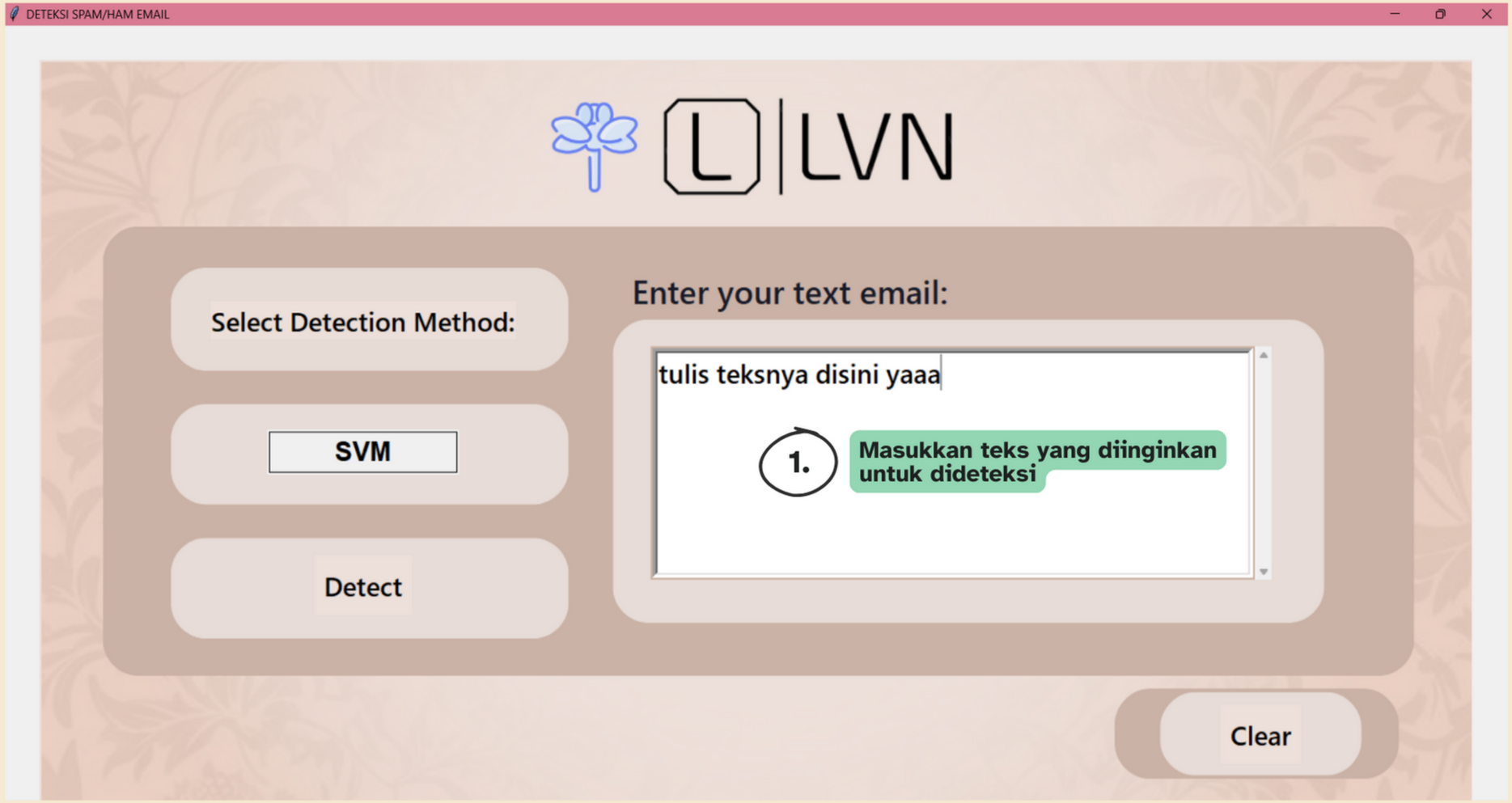
Kesimpulan



Berdasarkan hasil keseluruhan yang telah kami lakukan. Pada klasifikasi data spam dan ham kami menggunakan 3 algoritma diantaranya SVM, KNN, Naive Bayes dengan ekstraksi fitur menggunakan TFIDF dan Word2Vec. Kami mengkomparasi model ketiga algoritma dengan ekstraksi fitur yang berbeda, sehingga menunjukkan performa yang terbaik dari ketiga model algoritma. Dari ketiga algoritma didapatkan hasil terbaik dari SVM dengan ekstraksi fitur TFIDF yang menghasilkan akurasi sebesar 98%

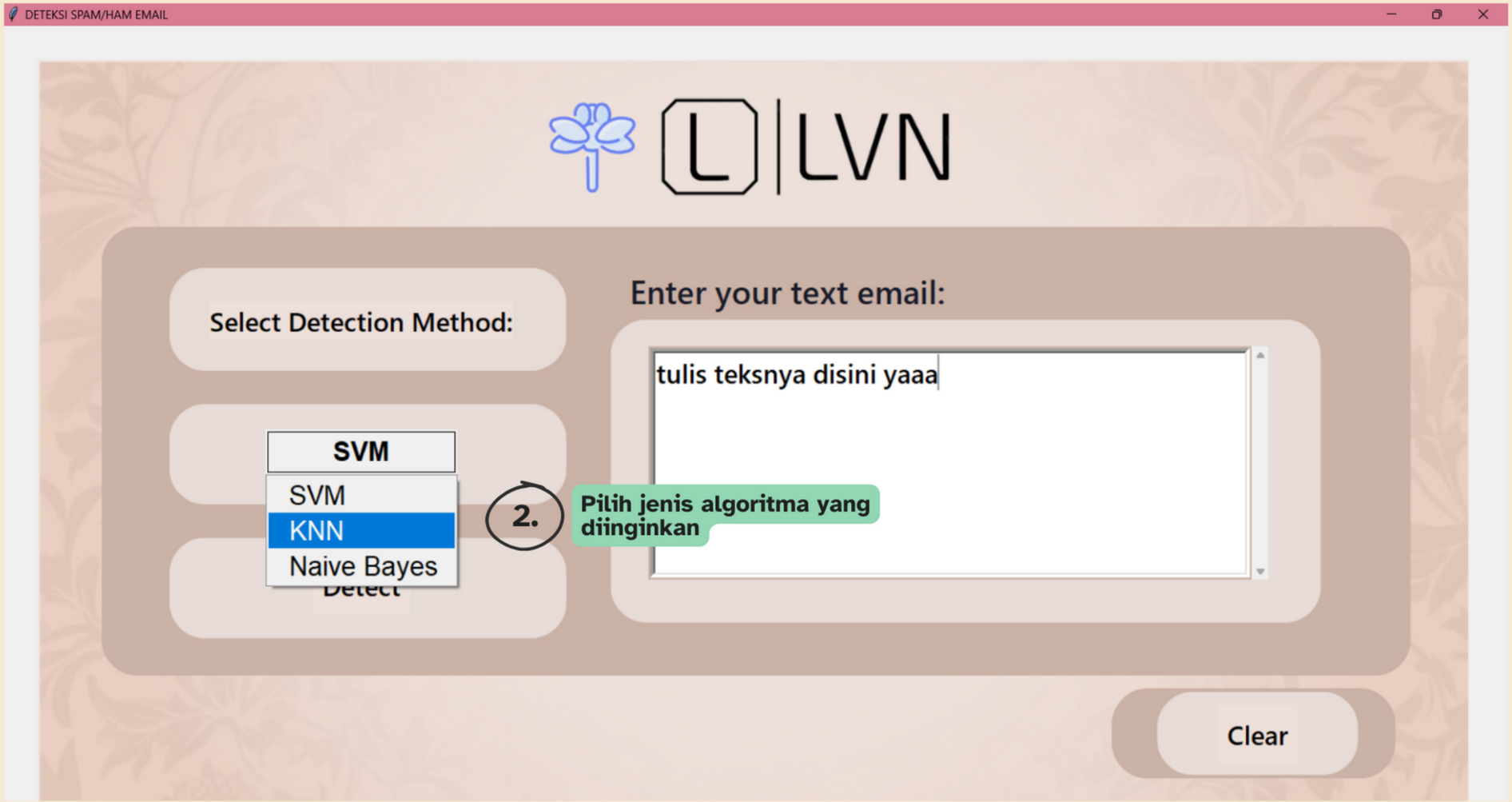
User Manual Guide

GUI



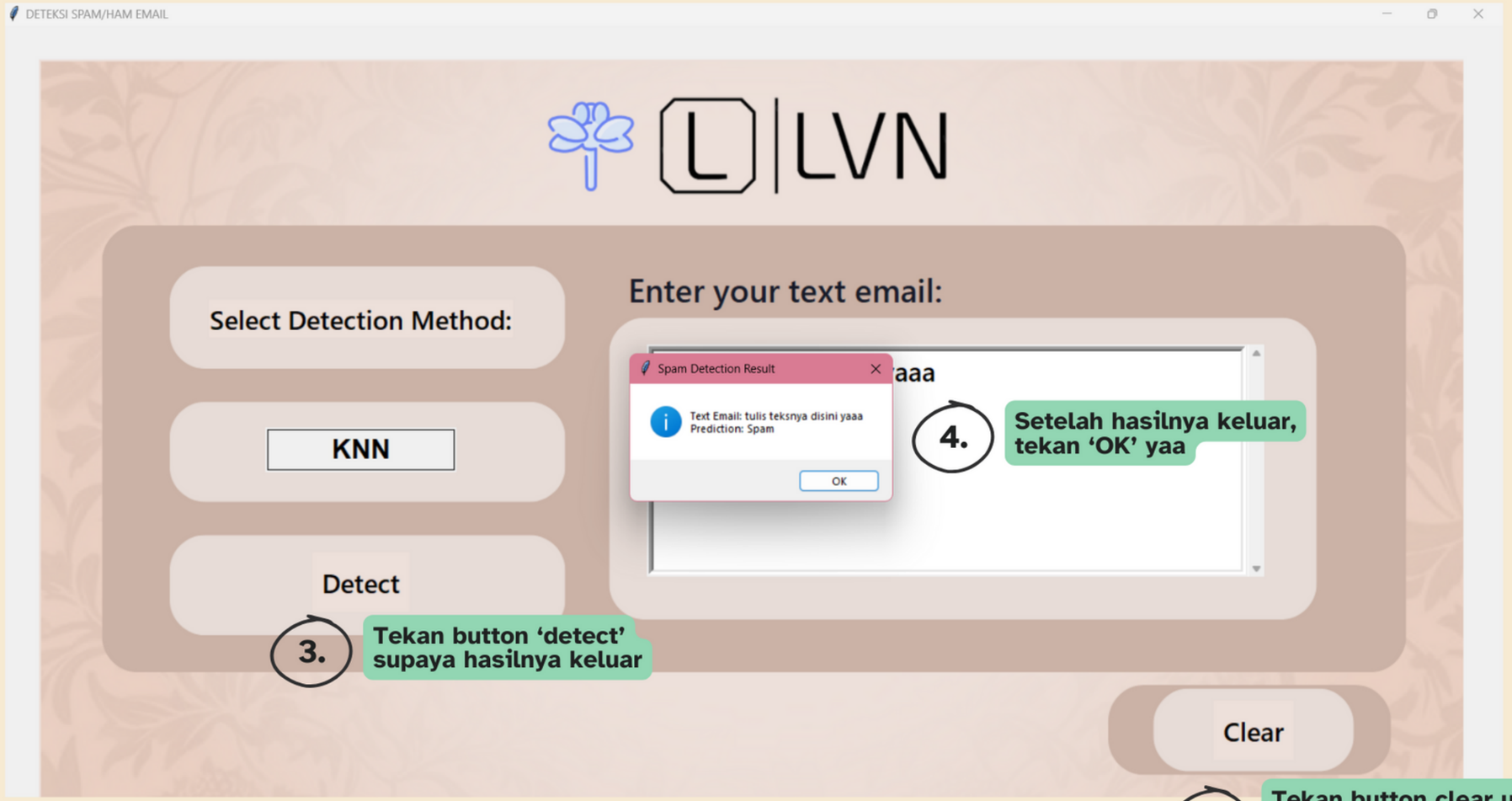
User Manual Guide

GUI



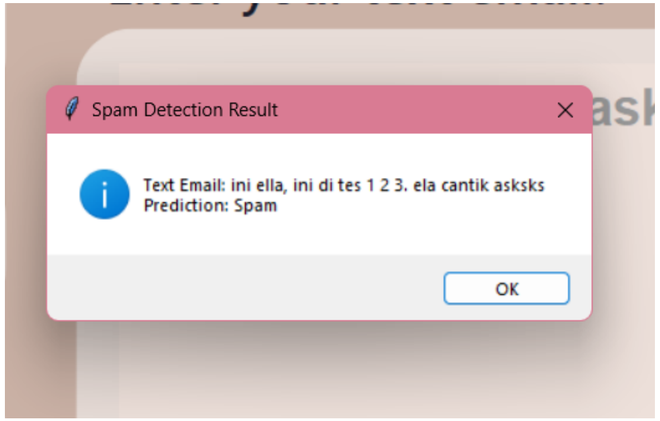
User Manual Guide

GUI

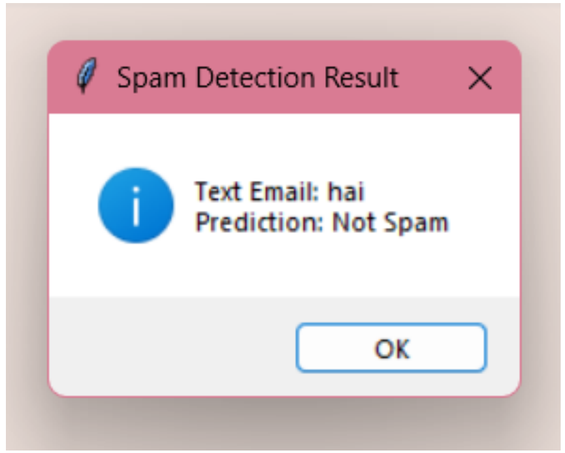


SPAM/HAM

SPAM



HAM



GUI akan memprediksi dari teks yang telah diinput oleh user karena telah mempelajari model dari pembuat

Perbedaan suatu notifikasi pada GUI dimana yang satu terdeteksi SPAM dan sebelahnya terdeteksi HAM atau NOT SPAM