IPB University
Bogor Indonesia

Department of Computer Science
http://cs.ipb.ac.id/

Data Mining Capstone Project Presentation
Genap 2021/2022, 3rd June 2022

Perbandingan Metode Support Vector Machine (SVM), Logistic Regression, dan XGBoost Classifier dalam Mengklasifikasikan Pesan Spam

Alvin Christian, Andreas Prananda Putra, Perisai Zidane Hanapi, Rahmat Qodri

Department of Computer Science, IPB University, Bogor, Indonesia

andreaspranandap@gmail.com, apeirodox@gmail.com, rahmat.qodri1812@gmail.com, pzidaneh@gmail.com.

Anggota Kelompok:

Alvin Christian
G14190047

Rahmat Qodri
G14190029

Perisai Zidane
G14190055

Andreas Prananda P
G14190073

Latar Belakang

- Manusia selalu berkomunikasi, secara lisan maupun tulisan.
- Salah satu media komunikasi adalah SMS.
- Pesan masuk yang tersimpan akan menumpuk: memenuhi penyimpanan & menenggelamkan pesan penting
- Perlu ada penyaring
- Diperlukan metode klasifikasi berbasis teks untuk memecahkan masalah ini

Dataset yang digunakan

Dataset yang digunakan dalam penelitian kali ini adalah:

Dataset SMS Spam yang terdiri dari 2 kolom (teks dan label) dan 1.143 instans. Kolom teks berisi pesan yang diterima dan kolom label menjelaskan apakah pesan tersebut merupakan pesan normal, promo, atau penipuan. Tidak terdapat *missing value*.

Sumber:
https://gist.github.com/agtbaskara/a1a7017027cc1df9d35cf06e1e5575b7

Dataset yang digunakan

Karena fokus penelitian kali ini hanyalah untuk mengklasifikasikan apakah pesan tersebut spam atau bukan, maka label pesan “promosi” dan “penipuan” akan digabungkan menjadi “spam”. Kasus ini menjadi masalah *binary classification*.

Kelas pesan “normal” dan “spam” sudah seimbang dengan jumlah:

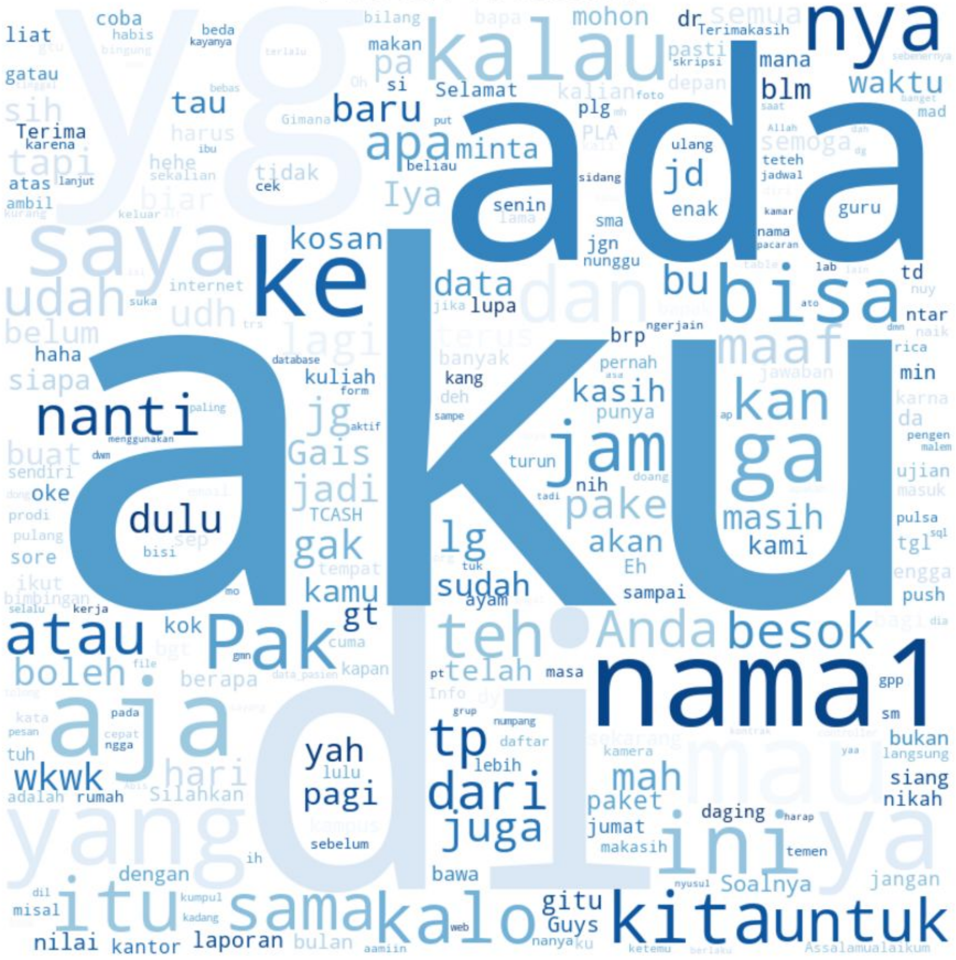
- spam = 574 (50.22%)
- normal = 569 (49.78%)

Tahap Praproses Teks

- konversi seluruh karakter menjadi huruf kecil
- menghilangkan tanda baca dan angka
- mengeluarkan kata sambung (*stopwords*)
- normalisasi bahasa gaul/alay (*slang*)
- *stemming* (mengubah kata berimbuhan menjadi kata dasar)
- menghilangkan *whitespace*

Word cloud sebelum praproses teks

Pesan Normal



Pesan Spam

Word cloud setelah praposes teks

Pesan Normal

Pesan Spam

Data splitting

- Dataset tersebut kemudian dibagi menjadi 80% data latih (*training*) dan 20% data uji (*testing*) secara acak.
- Selanjutnya data latih akan digunakan dalam validasi silang untuk mencari parameter paling baik (*hyperparameter tuning*)
- Label yang sebelumnya “normal” dan “spam” diubah menjadi 0 (normal) dan 1 (spam)

Vektorisasi teks: TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) merupakan metode yang digunakan menentukan seberapa jauh keterhubungan kata (term) terhadap dokumen dengan memberikan bobot setiap kata (Herwijayanti et al. 2018)

	R1	R2	R3	TF1	TF2	TF3	IDF	TFIDF1	TFIDF2	TFIDF3
makan	1	1	1	0.2	0.25	0.167	0	0	0	0
disini	1	1	1	0.2	0.25	0.167	0	0	0	0
gurih	1	0	0	0.2	0	0.000	0.48	0.095	0	0
dan	1	0	1	0.2	0	0.167	0.18	0.035	0	0.0293
enak	1	0	1	0.2	0	0.167	0.18	0.035	0	0.0293
biasa	0	1	0	0	0.25	0.000	0.48	0	0.119	0
saja	0	1	0	0	0.25	0.000	0.48	0	0.119	0
hambar	0	0	1	0	0	0.167	0.48	0	0	0.080
tidak	0	0	1	0	0	0.167	0.48	0	0	0.080

Model klasifikasi spam yang digunakan

Model yang digunakan:

- Support Vector Machine (SVM). Pendekatan berdasarkan properti geometris dari data.
- Logistic Regression. Pendekatan secara statistik.
- Extreme Gradient Boosting (XGB). Metode ensemble.

Hyperparameter tuning

Proses *hyperparameter tuning* dilakukan dengan menggunakan *grid search*. *Grid search* merupakan algoritma *brute force* yang mencoba seluruh kombinasi parameter yang ditentukan dengan validasi silang dari data latih (*training*).

SVM vs SVM tuned

	Predicted	
Actual	**Spam**	**Normal**
Spam	103	0
Normal	69	57

	Predicted	
Actual	**Spam**	**Normal**
Spam	100	3
Normal	3	123

Logistic Regression vs Logit tuned

	Predicted	
Actual	**Spam**	**Normal**
Spam	101	2
Normal	4	122

	Predicted	
Actual	**Spam**	**Normal**
Spam	101	2
Normal	3	123

XGB vs XGB tuned

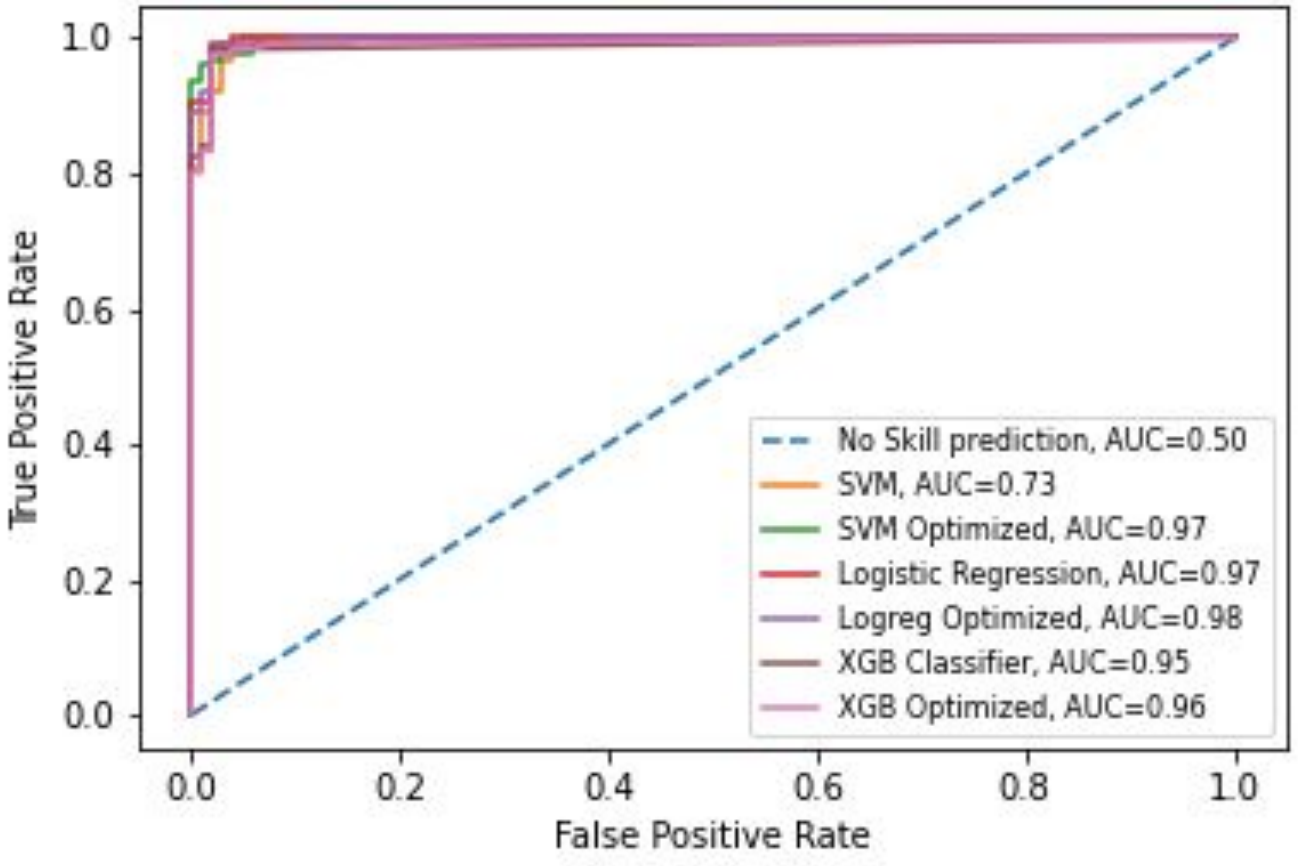
	Predicted	
Actual	**Spam**	**Normal**
Spam	101	2
Normal	9	117

	Predicted	
Actual	**Spam**	**Normal**
Spam	101	2
Normal	7	119

Perbandingan akurasi model

Model	accuracy	f1_score	auc_roc_score	running_time_second
SVM	0.6987	0.6230	0.7262	0.08
SVM tuned	0.9738	0.9762	0.9735	24.98
Logistic Regression	0.9738	0.9760	0.9744	0.02
Logistic Regression tuned	0.9782	0.9801	0.9784	1.86
XGBClassifier	0.9520	0.9551	0.9546	0.54
XGBClassifier tuned	0.9607	0.9636	0.9625	172.84

Perbandingan akurasi model



Indeks pesan yang salah diprediksi

```
[77] wrong_pred_index_svm

     array([  0,   5,   8,  12,  15,  20,  25,  26,  33,  39,  41,  42,  45,
             49,  50,  54,  56,  57,  60,  74,  76,  78,  79,  84,  85,  86,
             87,  92,  94, 104, 108, 110, 114, 120, 124, 126, 127, 128, 129,
            140, 143, 146, 151, 152, 154, 155, 159, 163, 165, 166, 170, 171,
            178, 186, 188, 193, 194, 195, 196, 200, 201, 210, 211, 213, 214,
            217, 222, 223, 225])

[78] wrong_pred_index_svm_cv

     array([117, 121, 128, 174, 181, 193])

[64] wrong_pred_index_lgr

     array([ 42,  60, 128, 174, 181, 217])

[65] wrong_pred_index_lgr_cv

     array([ 60, 128, 174, 181, 217])
```

```
[79] wrong_pred_index_xgb

     array([ 15,  42,  60,  76, 110, 121, 128, 146, 174, 181, 194])

     wrong_pred_index_xgb_cv

     array([ 15,  42,  60, 121, 128, 146, 174, 181, 194])
```

Pesan yang banyak salah diprediksi oleh ketiga model:

- 42
- 60
- 128
- 174
- 181

Isi pesan yang salah diprediksi

- brminat cash kredit mtor scond istmwa tipe merek mnyediakn unit dediktp yk
- main gamesmu beli banyak apps apps google play store
- pt pertamina persero karyawan i lulus smk sih sih kirim lamar cv ijazah photox notip email pertaminareckrutment ymail com
- pin tcash sila pin nikmat layan tcash telkomsel
- atur telepon seluler kirim masuk pin terima kasih

Kesimpulan

Berdasarkan ketiga model klasifikasi yang digunakan (SVM, Logit, dan XGB) untuk melakukan klasifikasi pesan spam dari data yang sudah melalui praproses teks, regresi logistik memiliki nilai akurasi paling tinggi.

Regresi Logistik dengan parameter yang sudah di-*tuning* memiliki nilai akurasi sebesar 0.9782, skor f1 sebesar 0.98, dan nilai roc/auc sebesar 0.9784.

Saran

Tahap praproses teks dapat ditingkatkan dengan membuat kamus besar kumpulan kata sambung (*stopwords*), kumpulan bahasa non-formal, kumpulan singkatan, dan kumpulan kata yang tidak lengkap yang diharapkan dapat meningkatkan performa model klasifikasi.

Thank you

IPB University
Bogor Indonesia

Department of Computer Science
FMIPA-IPB Kampus Darmaga
Jl. Meranti Wing 20 Level V, Bogor, Indonesia
Phone/Fax: +62 251 8625584
http://cs.ipb.ac.id/