Klasifikasi SMS (Model NB)[¶](#gjdgxs)

Contoh Deloyment untuk Natural Language Processing (NLP)

Orbit Future Academy - AI Mastery - KM Batch 3

Tim Deployment dan Tim NLP

2022

Hari ini, kita akan menyelesaikan kasus klasifikasi SMS spam menggunakan skema seperti gambar di bawah ini. 

Langkah-langkah penyelesaian melitputi:

1. Data Acquisition
2. Exploratory Data Analysis (EDA)
3. Pre-processing
   1. Case Folding
   2. Filtering
   3. Stopword
   4. Stemming
4. Feature Engineering
   1. Feature Extraction - TF IDF
   2. Feature Selection - Chi-Square
5. Modelling (Machine Learning)
6. Prediction
7. Model Evaluation

Pada kasus ini, kita menggunakan data set berbahasa indonesia, sehingga kita membutuhkan *Library Sastrawi* untuk menyelesaikan beberapa permasalahan seperti *stemming*.

Modules dan Packages[¶](#30j0zll)

In [ ]:

**!**pip -q install sastrawi

In [ ]:

**!**pip -q install nltk

In [1]:

**import** pandas **as** pd  
**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
  
**import** pickle  
  
**import** re  
**import** nltk  
nltk**.**download('stopwords')  
  
**from** nltk.tokenize **import** sent\_tokenize, word\_tokenize  
**from** nltk.corpus **import** stopwords  
  
**from** Sastrawi.Stemmer.StemmerFactory **import** StemmerFactory  
  
**from** sklearn.feature\_extraction.text **import** TfidfVectorizer  
  
**from** sklearn.feature\_selection **import** SelectKBest   
**from** sklearn.feature\_selection **import** chi2   
  
**import** random  
**from** joblib **import** dump  
  
**from** sklearn.naive\_bayes **import** MultinomialNB  
  
**from** sklearn.model\_selection **import** train\_test\_split  
  
**from** sklearn.metrics **import** accuracy\_score  
**from** sklearn.metrics **import** confusion\_matrix  
**from** sklearn.metrics **import** classification\_report  
  
**%matplotlib** inline

[nltk\_data] Downloading package stopwords to  
[nltk\_data] C:\Users\newbie\AppData\Roaming\nltk\_data...  
[nltk\_data] Package stopwords is already up-to-date!

Data Acquisition[¶](#1fob9te)

Penjelasan Label

* 0: SMS normal
* 1: SMS fraud atau penipuan
* 2: SMS promo

In [2]:

data **=** pd**.**read\_csv('dataset\_sms\_spam\_v1.csv')  
data**.**head()

Out[2]:

|  | **teks** | **label** |
| --- | --- | --- |
| **0** | [PROMO] Beli paket Flash mulai 1GB di MY TELKO... | 2 |
| **1** | 2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A... | 2 |
| **2** | 2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ... | 2 |
| **3** | 2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ... | 2 |
| **4** | 4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an... | 2 |

In [3]:

data**.**info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1143 entries, 0 to 1142  
Data columns (total 2 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 teks 1143 non-null object  
 1 label 1143 non-null int64   
dtypes: int64(1), object(1)  
memory usage: 18.0+ KB

Exploratory Data Analysis (EDA)[¶](#3znysh7)

In [4]:

print('Total Jumlah SMS:', data**.**shape[0], 'data\n')  
print('terdiri dari (label):')  
print('-- [0] SMS Normal\t:', data[data**.**label **==** 0]**.**shape[0], 'data')  
print('-- [1] Fraud / Penipuan\t:', data[data**.**label **==** 1]**.**shape[0], 'data')  
print('-- [2] Promo\t\t:', data[data**.**label **==** 2]**.**shape[0], 'data\n')

Total Jumlah SMS: 1143 data  
  
terdiri dari (label):  
-- [0] SMS Normal : 569 data  
-- [1] Fraud / Penipuan : 335 data  
-- [2] Promo : 239 data

In [5]:

height **=** data['label']**.**value\_counts()  
labels **=** ('SMS Normal', 'SMS Fraud / Penipuan', 'SMS Promo')  
y\_pos **=** np**.**arange(len(labels))  
  
plt**.**figure(figsize**=**(7,4), dpi**=**80)  
plt**.**ylim(0,600)  
plt**.**title('Distribusi Kategori SMS', fontweight**=**'bold')  
plt**.**xlabel('Kategori', fontweight**=**'bold')  
plt**.**ylabel('Jumlah', fontweight**=**'bold')  
plt**.**bar(y\_pos, height, color**=**['deepskyblue', 'royalblue', 'skyblue'])  
plt**.**xticks(y\_pos, labels)  
plt**.**show()

Text Preprocessing[¶](#2et92p0)

Case Folding[¶](#tyjcwt)

In [6]:

*# Buat fungsi untuk langkah case folding*  
**def** casefolding(text):  
 text **=** text**.**lower() *# Mengubah teks menjadi lower case*  
 text **=** re**.**sub(r'https?://\S+|www\.\S+', '', text) *# Menghapus URL*  
 text **=** re**.**sub(r'[-+]?[0-9]+', '', text) *# Menghapus angka*  
 text **=** re**.**sub(r'[^\w\s]','', text) *# Menghapus karakter tanda baca*  
 text **=** text**.**strip()  
 **return** text

In [7]:

raw\_sample **=** data['teks']**.**iloc[5]  
case\_folding **=** casefolding(raw\_sample)  
  
print('Raw data\t: ', raw\_sample)  
print('Case folding\t: ', case\_folding)

Raw data : 5 HARI LAGI ! EKSTRA Pulsa 50rb dg beli paket internet bulanan di MyTelkomsel utk pembelian pertama sjk 25Aug-25Sept. Cek detail promo di tsel.me/mytsel1  
Case folding : hari lagi ekstra pulsa rb dg beli paket internet bulanan di mytelkomsel utk pembelian pertama sjk augsept cek detail promo di tselmemytsel

Word Normalization[¶](#3dy6vkm)

In [8]:

key\_norm **=** pd**.**read\_csv('key\_norm.csv')  
  
**def** text\_normalize(text):  
 text **=** ' '**.**join([key\_norm[key\_norm['singkat'] **==** word]['hasil']**.**values[0] **if** (key\_norm['singkat'] **==** word)**.**any() **else** word **for** word **in** text**.**split()])  
 text **=** str**.**lower(text)  
 **return** text

Filtering (Stopword Removal)[¶](#1t3h5sf)

In [9]:

stopwords\_ind **=** stopwords**.**words('indonesian')

In [10]:

len(stopwords\_ind)

Out[10]:

758

In [11]:

*# Lihat daftar stopword yang disediakan NLTK*  
stopwords\_ind

Out[11]:

['ada',  
 'adalah',  
 'adanya',  
 'adapun',  
 'agak',  
 'agaknya',  
 'agar',  
 'akan',  
 'akankah',  
 'akhir',  
 'akhiri',  
 'akhirnya',  
 'aku',  
 'akulah',  
 'amat',  
 'amatlah',  
 'anda',  
 'andalah',  
 'antar',  
 'antara',  
 'antaranya',  
 'apa',  
 'apaan',  
 'apabila',  
 'apakah',  
 'apalagi',  
 'apatah',  
 'artinya',  
 'asal',  
 'asalkan',  
 'atas',  
 'atau',  
 'ataukah',  
 'ataupun',  
 'awal',  
 'awalnya',  
 'bagai',  
 'bagaikan',  
 'bagaimana',  
 'bagaimanakah',  
 'bagaimanapun',  
 'bagi',  
 'bagian',  
 'bahkan',  
 'bahwa',  
 'bahwasanya',  
 'baik',  
 'bakal',  
 'bakalan',  
 'balik',  
 'banyak',  
 'bapak',  
 'baru',  
 'bawah',  
 'beberapa',  
 'begini',  
 'beginian',  
 'beginikah',  
 'beginilah',  
 'begitu',  
 'begitukah',  
 'begitulah',  
 'begitupun',  
 'bekerja',  
 'belakang',  
 'belakangan',  
 'belum',  
 'belumlah',  
 'benar',  
 'benarkah',  
 'benarlah',  
 'berada',  
 'berakhir',  
 'berakhirlah',  
 'berakhirnya',  
 'berapa',  
 'berapakah',  
 'berapalah',  
 'berapapun',  
 'berarti',  
 'berawal',  
 'berbagai',  
 'berdatangan',  
 'beri',  
 'berikan',  
 'berikut',  
 'berikutnya',  
 'berjumlah',  
 'berkali-kali',  
 'berkata',  
 'berkehendak',  
 'berkeinginan',  
 'berkenaan',  
 'berlainan',  
 'berlalu',  
 'berlangsung',  
 'berlebihan',  
 'bermacam',  
 'bermacam-macam',  
 'bermaksud',  
 'bermula',  
 'bersama',  
 'bersama-sama',  
 'bersiap',  
 'bersiap-siap',  
 'bertanya',  
 'bertanya-tanya',  
 'berturut',  
 'berturut-turut',  
 'bertutur',  
 'berujar',  
 'berupa',  
 'besar',  
 'betul',  
 'betulkah',  
 'biasa',  
 'biasanya',  
 'bila',  
 'bilakah',  
 'bisa',  
 'bisakah',  
 'boleh',  
 'bolehkah',  
 'bolehlah',  
 'buat',  
 'bukan',  
 'bukankah',  
 'bukanlah',  
 'bukannya',  
 'bulan',  
 'bung',  
 'cara',  
 'caranya',  
 'cukup',  
 'cukupkah',  
 'cukuplah',  
 'cuma',  
 'dahulu',  
 'dalam',  
 'dan',  
 'dapat',  
 'dari',  
 'daripada',  
 'datang',  
 'dekat',  
 'demi',  
 'demikian',  
 'demikianlah',  
 'dengan',  
 'depan',  
 'di',  
 'dia',  
 'diakhiri',  
 'diakhirinya',  
 'dialah',  
 'diantara',  
 'diantaranya',  
 'diberi',  
 'diberikan',  
 'diberikannya',  
 'dibuat',  
 'dibuatnya',  
 'didapat',  
 'didatangkan',  
 'digunakan',  
 'diibaratkan',  
 'diibaratkannya',  
 'diingat',  
 'diingatkan',  
 'diinginkan',  
 'dijawab',  
 'dijelaskan',  
 'dijelaskannya',  
 'dikarenakan',  
 'dikatakan',  
 'dikatakannya',  
 'dikerjakan',  
 'diketahui',  
 'diketahuinya',  
 'dikira',  
 'dilakukan',  
 'dilalui',  
 'dilihat',  
 'dimaksud',  
 'dimaksudkan',  
 'dimaksudkannya',  
 'dimaksudnya',  
 'diminta',  
 'dimintai',  
 'dimisalkan',  
 'dimulai',  
 'dimulailah',  
 'dimulainya',  
 'dimungkinkan',  
 'dini',  
 'dipastikan',  
 'diperbuat',  
 'diperbuatnya',  
 'dipergunakan',  
 'diperkirakan',  
 'diperlihatkan',  
 'diperlukan',  
 'diperlukannya',  
 'dipersoalkan',  
 'dipertanyakan',  
 'dipunyai',  
 'diri',  
 'dirinya',  
 'disampaikan',  
 'disebut',  
 'disebutkan',  
 'disebutkannya',  
 'disini',  
 'disinilah',  
 'ditambahkan',  
 'ditandaskan',  
 'ditanya',  
 'ditanyai',  
 'ditanyakan',  
 'ditegaskan',  
 'ditujukan',  
 'ditunjuk',  
 'ditunjuki',  
 'ditunjukkan',  
 'ditunjukkannya',  
 'ditunjuknya',  
 'dituturkan',  
 'dituturkannya',  
 'diucapkan',  
 'diucapkannya',  
 'diungkapkan',  
 'dong',  
 'dua',  
 'dulu',  
 'empat',  
 'enggak',  
 'enggaknya',  
 'entah',  
 'entahlah',  
 'guna',  
 'gunakan',  
 'hal',  
 'hampir',  
 'hanya',  
 'hanyalah',  
 'hari',  
 'harus',  
 'haruslah',  
 'harusnya',  
 'hendak',  
 'hendaklah',  
 'hendaknya',  
 'hingga',  
 'ia',  
 'ialah',  
 'ibarat',  
 'ibaratkan',  
 'ibaratnya',  
 'ibu',  
 'ikut',  
 'ingat',  
 'ingat-ingat',  
 'ingin',  
 'inginkah',  
 'inginkan',  
 'ini',  
 'inikah',  
 'inilah',  
 'itu',  
 'itukah',  
 'itulah',  
 'jadi',  
 'jadilah',  
 'jadinya',  
 'jangan',  
 'jangankan',  
 'janganlah',  
 'jauh',  
 'jawab',  
 'jawaban',  
 'jawabnya',  
 'jelas',  
 'jelaskan',  
 'jelaslah',  
 'jelasnya',  
 'jika',  
 'jikalau',  
 'juga',  
 'jumlah',  
 'jumlahnya',  
 'justru',  
 'kala',  
 'kalau',  
 'kalaulah',  
 'kalaupun',  
 'kalian',  
 'kami',  
 'kamilah',  
 'kamu',  
 'kamulah',  
 'kan',  
 'kapan',  
 'kapankah',  
 'kapanpun',  
 'karena',  
 'karenanya',  
 'kasus',  
 'kata',  
 'katakan',  
 'katakanlah',  
 'katanya',  
 'ke',  
 'keadaan',  
 'kebetulan',  
 'kecil',  
 'kedua',  
 'keduanya',  
 'keinginan',  
 'kelamaan',  
 'kelihatan',  
 'kelihatannya',  
 'kelima',  
 'keluar',  
 'kembali',  
 'kemudian',  
 'kemungkinan',  
 'kemungkinannya',  
 'kenapa',  
 'kepada',  
 'kepadanya',  
 'kesampaian',  
 'keseluruhan',  
 'keseluruhannya',  
 'keterlaluan',  
 'ketika',  
 'khususnya',  
 'kini',  
 'kinilah',  
 'kira',  
 'kira-kira',  
 'kiranya',  
 'kita',  
 'kitalah',  
 'kok',  
 'kurang',  
 'lagi',  
 'lagian',  
 'lah',  
 'lain',  
 'lainnya',  
 'lalu',  
 'lama',  
 'lamanya',  
 'lanjut',  
 'lanjutnya',  
 'lebih',  
 'lewat',  
 'lima',  
 'luar',  
 'macam',  
 'maka',  
 'makanya',  
 'makin',  
 'malah',  
 'malahan',  
 'mampu',  
 'mampukah',  
 'mana',  
 'manakala',  
 'manalagi',  
 'masa',  
 'masalah',  
 'masalahnya',  
 'masih',  
 'masihkah',  
 'masing',  
 'masing-masing',  
 'mau',  
 'maupun',  
 'melainkan',  
 'melakukan',  
 'melalui',  
 'melihat',  
 'melihatnya',  
 'memang',  
 'memastikan',  
 'memberi',  
 'memberikan',  
 'membuat',  
 'memerlukan',  
 'memihak',  
 'meminta',  
 'memintakan',  
 'memisalkan',  
 'memperbuat',  
 'mempergunakan',  
 'memperkirakan',  
 'memperlihatkan',  
 'mempersiapkan',  
 'mempersoalkan',  
 'mempertanyakan',  
 'mempunyai',  
 'memulai',  
 'memungkinkan',  
 'menaiki',  
 'menambahkan',  
 'menandaskan',  
 'menanti',  
 'menanti-nanti',  
 'menantikan',  
 'menanya',  
 'menanyai',  
 'menanyakan',  
 'mendapat',  
 'mendapatkan',  
 'mendatang',  
 'mendatangi',  
 'mendatangkan',  
 'menegaskan',  
 'mengakhiri',  
 'mengapa',  
 'mengatakan',  
 'mengatakannya',  
 'mengenai',  
 'mengerjakan',  
 'mengetahui',  
 'menggunakan',  
 'menghendaki',  
 'mengibaratkan',  
 'mengibaratkannya',  
 'mengingat',  
 'mengingatkan',  
 'menginginkan',  
 'mengira',  
 'mengucapkan',  
 'mengucapkannya',  
 'mengungkapkan',  
 'menjadi',  
 'menjawab',  
 'menjelaskan',  
 'menuju',  
 'menunjuk',  
 'menunjuki',  
 'menunjukkan',  
 'menunjuknya',  
 'menurut',  
 'menuturkan',  
 'menyampaikan',  
 'menyangkut',  
 'menyatakan',  
 'menyebutkan',  
 'menyeluruh',  
 'menyiapkan',  
 'merasa',  
 'mereka',  
 'merekalah',  
 'merupakan',  
 'meski',  
 'meskipun',  
 'meyakini',  
 'meyakinkan',  
 'minta',  
 'mirip',  
 'misal',  
 'misalkan',  
 'misalnya',  
 'mula',  
 'mulai',  
 'mulailah',  
 'mulanya',  
 'mungkin',  
 'mungkinkah',  
 'nah',  
 'naik',  
 'namun',  
 'nanti',  
 'nantinya',  
 'nyaris',  
 'nyatanya',  
 'oleh',  
 'olehnya',  
 'pada',  
 'padahal',  
 'padanya',  
 'pak',  
 'paling',  
 'panjang',  
 'pantas',  
 'para',  
 'pasti',  
 'pastilah',  
 'penting',  
 'pentingnya',  
 'per',  
 'percuma',  
 'perlu',  
 'perlukah',  
 'perlunya',  
 'pernah',  
 'persoalan',  
 'pertama',  
 'pertama-tama',  
 'pertanyaan',  
 'pertanyakan',  
 'pihak',  
 'pihaknya',  
 'pukul',  
 'pula',  
 'pun',  
 'punya',  
 'rasa',  
 'rasanya',  
 'rata',  
 'rupanya',  
 'saat',  
 'saatnya',  
 'saja',  
 'sajalah',  
 'saling',  
 'sama',  
 'sama-sama',  
 'sambil',  
 'sampai',  
 'sampai-sampai',  
 'sampaikan',  
 'sana',  
 'sangat',  
 'sangatlah',  
 'satu',  
 'saya',  
 'sayalah',  
 'se',  
 'sebab',  
 'sebabnya',  
 'sebagai',  
 'sebagaimana',  
 'sebagainya',  
 'sebagian',  
 'sebaik',  
 'sebaik-baiknya',  
 'sebaiknya',  
 'sebaliknya',  
 'sebanyak',  
 'sebegini',  
 'sebegitu',  
 'sebelum',  
 'sebelumnya',  
 'sebenarnya',  
 'seberapa',  
 'sebesar',  
 'sebetulnya',  
 'sebisanya',  
 'sebuah',  
 'sebut',  
 'sebutlah',  
 'sebutnya',  
 'secara',  
 'secukupnya',  
 'sedang',  
 'sedangkan',  
 'sedemikian',  
 'sedikit',  
 'sedikitnya',  
 'seenaknya',  
 'segala',  
 'segalanya',  
 'segera',  
 'seharusnya',  
 'sehingga',  
 'seingat',  
 'sejak',  
 'sejauh',  
 'sejenak',  
 'sejumlah',  
 'sekadar',  
 'sekadarnya',  
 'sekali',  
 'sekali-kali',  
 'sekalian',  
 'sekaligus',  
 'sekalipun',  
 'sekarang',  
 'sekarang',  
 'sekecil',  
 'seketika',  
 'sekiranya',  
 'sekitar',  
 'sekitarnya',  
 'sekurang-kurangnya',  
 'sekurangnya',  
 'sela',  
 'selain',  
 'selaku',  
 'selalu',  
 'selama',  
 'selama-lamanya',  
 'selamanya',  
 'selanjutnya',  
 'seluruh',  
 'seluruhnya',  
 'semacam',  
 'semakin',  
 'semampu',  
 'semampunya',  
 'semasa',  
 'semasih',  
 'semata',  
 'semata-mata',  
 'semaunya',  
 'sementara',  
 'semisal',  
 'semisalnya',  
 'sempat',  
 'semua',  
 'semuanya',  
 'semula',  
 'sendiri',  
 'sendirian',  
 'sendirinya',  
 'seolah',  
 'seolah-olah',  
 'seorang',  
 'sepanjang',  
 'sepantasnya',  
 'sepantasnyalah',  
 'seperlunya',  
 'seperti',  
 'sepertinya',  
 'sepihak',  
 'sering',  
 'seringnya',  
 'serta',  
 'serupa',  
 'sesaat',  
 'sesama',  
 'sesampai',  
 'sesegera',  
 'sesekali',  
 'seseorang',  
 'sesuatu',  
 'sesuatunya',  
 'sesudah',  
 'sesudahnya',  
 'setelah',  
 'setempat',  
 'setengah',  
 'seterusnya',  
 'setiap',  
 'setiba',  
 'setibanya',  
 'setidak-tidaknya',  
 'setidaknya',  
 'setinggi',  
 'seusai',  
 'sewaktu',  
 'siap',  
 'siapa',  
 'siapakah',  
 'siapapun',  
 'sini',  
 'sinilah',  
 'soal',  
 'soalnya',  
 'suatu',  
 'sudah',  
 'sudahkah',  
 'sudahlah',  
 'supaya',  
 'tadi',  
 'tadinya',  
 'tahu',  
 'tahun',  
 'tak',  
 'tambah',  
 'tambahnya',  
 'tampak',  
 'tampaknya',  
 'tandas',  
 'tandasnya',  
 'tanpa',  
 'tanya',  
 'tanyakan',  
 'tanyanya',  
 'tapi',  
 'tegas',  
 'tegasnya',  
 'telah',  
 'tempat',  
 'tengah',  
 'tentang',  
 'tentu',  
 'tentulah',  
 'tentunya',  
 'tepat',  
 'terakhir',  
 'terasa',  
 'terbanyak',  
 'terdahulu',  
 'terdapat',  
 'terdiri',  
 'terhadap',  
 'terhadapnya',  
 'teringat',  
 'teringat-ingat',  
 'terjadi',  
 'terjadilah',  
 'terjadinya',  
 'terkira',  
 'terlalu',  
 'terlebih',  
 'terlihat',  
 'termasuk',  
 'ternyata',  
 'tersampaikan',  
 'tersebut',  
 'tersebutlah',  
 'tertentu',  
 'tertuju',  
 'terus',  
 'terutama',  
 'tetap',  
 'tetapi',  
 'tiap',  
 'tiba',  
 'tiba-tiba',  
 'tidak',  
 'tidakkah',  
 'tidaklah',  
 'tiga',  
 'tinggi',  
 'toh',  
 'tunjuk',  
 'turut',  
 'tutur',  
 'tuturnya',  
 'ucap',  
 'ucapnya',  
 'ujar',  
 'ujarnya',  
 'umum',  
 'umumnya',  
 'ungkap',  
 'ungkapnya',  
 'untuk',  
 'usah',  
 'usai',  
 'waduh',  
 'wah',  
 'wahai',  
 'waktu',  
 'waktunya',  
 'walau',  
 'walaupun',  
 'wong',  
 'yaitu',  
 'yakin',  
 'yakni',  
 'yang']

In [12]:

*# Buat fungsi untuk langkah stopword removal*  
  
more\_stopword **=** ['tsel', 'gb', 'rb'] *# Tambahkan kata dalam daftar stopword*  
stopwords\_ind **=** stopwords\_ind **+** more\_stopword  
  
**def** remove\_stop\_words(text):  
 clean\_words **=** []  
 text **=** text**.**split()  
 **for** word **in** text:  
 **if** word **not** **in** stopwords\_ind:  
 clean\_words**.**append(word)  
 **return** " "**.**join(clean\_words)

In [13]:

raw\_sample **=** data['teks']**.**iloc[5]  
case\_folding **=** casefolding(raw\_sample)  
stopword\_removal **=** remove\_stop\_words(case\_folding)  
  
print('Raw data\t\t: ', raw\_sample)  
print('Case folding\t\t: ', case\_folding)  
print('Stopword removal\t: ', stopword\_removal)

Raw data : 5 HARI LAGI ! EKSTRA Pulsa 50rb dg beli paket internet bulanan di MyTelkomsel utk pembelian pertama sjk 25Aug-25Sept. Cek detail promo di tsel.me/mytsel1  
Case folding : hari lagi ekstra pulsa rb dg beli paket internet bulanan di mytelkomsel utk pembelian pertama sjk augsept cek detail promo di tselmemytsel  
Stopword removal : ekstra pulsa dg beli paket internet bulanan mytelkomsel utk pembelian sjk augsept cek detail promo tselmemytsel

Stemming[¶](#4d34og8)

In [14]:

factory **=** StemmerFactory()  
stemmer **=** factory**.**create\_stemmer()  
  
*# Buat fungsi untuk langkah stemming bahasa Indonesia*  
**def** stemming(text):  
 text **=** stemmer**.**stem(text)  
 **return** text

In [15]:

raw\_sample **=** data['teks']**.**iloc[5]  
case\_folding **=** casefolding(raw\_sample)  
stopword\_removal **=** remove\_stop\_words(case\_folding)  
text\_stemming **=** stemming(stopword\_removal)  
  
print('Raw data\t\t: ', raw\_sample)  
print('Case folding\t\t: ', case\_folding)  
print('Stopword removal\t: ', stopword\_removal)  
print('Stemming\t\t: ', text\_stemming)

Raw data : 5 HARI LAGI ! EKSTRA Pulsa 50rb dg beli paket internet bulanan di MyTelkomsel utk pembelian pertama sjk 25Aug-25Sept. Cek detail promo di tsel.me/mytsel1  
Case folding : hari lagi ekstra pulsa rb dg beli paket internet bulanan di mytelkomsel utk pembelian pertama sjk augsept cek detail promo di tselmemytsel  
Stopword removal : ekstra pulsa dg beli paket internet bulanan mytelkomsel utk pembelian sjk augsept cek detail promo tselmemytsel  
Stemming : ekstra pulsa dg beli paket internet bulan mytelkomsel utk beli sjk augsept cek detail promo tselmemytsel

Text Preprocessing Pipeline[¶](#2s8eyo1)

In [16]:

*# Buat fungsi untuk menggabungkan seluruh langkah text preprocessing*  
**def** text\_preprocessing\_process(text):  
 text **=** casefolding(text)  
 text **=** text\_normalize(text)  
 text **=** remove\_stop\_words(text)  
 text **=** stemming(text)  
 **return** text

In [17]:

**%%time**  
data['clean\_teks'] **=** data['teks']**.**apply(text\_preprocessing\_process)  
  
*# Perhatikan waktu komputasi ketika proses text preprocessing*

Wall time: 7min 15s

In [18]:

data

Out[18]:

|  | **teks** | **label** | **clean\_teks** |
| --- | --- | --- | --- |
| **0** | [PROMO] Beli paket Flash mulai 1GB di MY TELKO... | 2 | promo beli paket flash my telkomsel app extra ... |
| **1** | 2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A... | 2 | rupiah ribu spesial pilih aktif promo sd novem... |
| **2** | 2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ... | 2 | pulang hormat sisa kuota flash kb download myt... |
| **3** | 2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ... | 2 | pulang hormat sisa kuota flash kb download myt... |
| **4** | 4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an... | 2 | rupiah ribu spesial pilih aktif buru skb |
| **...** | ... | ... | ... |
| **1138** | Yooo sama2, oke nanti aku umumin di grup kelas | 0 | yooo oke umumin grup kelas |
| **1139** | 😁 sebelumnya ga ad nulis kerudung. Kirain warn... | 0 | nulis kerudung kirain warna jins |
| **1140** | Mba mau kirim 300 ya | 0 | mbak kirim ya |
| **1141** | nama1 beaok bwrangkat pagi...mau cas atay tra... | 0 | nama beaok bwrangkat pagimau cas atay tranfer |
| **1142** | No bri atas nama kamu mana | 0 | nomor bri nama |

1143 rows × 3 columns

In [19]:

*# Simpan data yang telah melalui text preprocessing agar kita tidak perlu menjalankan proses tersebut mulai awal (Opsional)*  
data**.**to\_csv('clean\_data.csv')

Feature Engineering[¶](#17dp8vu)

In [20]:

*# Pisahkan kolom feature dan target*  
X **=** data['clean\_teks']  
y **=** data['label']

In [21]:

X

Out[21]:

0 promo beli paket flash my telkomsel app extra ...  
1 rupiah ribu spesial pilih aktif promo sd novem...  
2 pulang hormat sisa kuota flash kb download myt...  
3 pulang hormat sisa kuota flash kb download myt...  
4 rupiah ribu spesial pilih aktif buru skb  
 ...   
1138 yooo oke umumin grup kelas  
1139 nulis kerudung kirain warna jins  
1140 mbak kirim ya  
1141 nama beaok bwrangkat pagimau cas atay tranfer  
1142 nomor bri nama  
Name: clean\_teks, Length: 1143, dtype: object

In [22]:

y

Out[22]:

0 2  
1 2  
2 2  
3 2  
4 2  
 ..  
1138 0  
1139 0  
1140 0  
1141 0  
1142 0  
Name: label, Length: 1143, dtype: int64

Feature Extraction (TF-IDF & N-Gram)[¶](#3rdcrjn)

Proses mengubah teks menjadi vector menggunakan metode TF-IDF

In [23]:

tf\_idf **=** TfidfVectorizer(ngram\_range**=**(1,1))  
tf\_idf**.**fit(X)

Out[23]:

TfidfVectorizer()

In [24]:

*# Melihat Jumlah Fitur*  
print(len(tf\_idf**.**get\_feature\_names()))

3253

In [25]:

*# Melihat fitur-fitur apa saja yang ada di dalam corpus*  
print(tf\_idf**.**get\_feature\_names())

['aa', 'aamiiiin', 'aamiin', 'ab', 'abadi', 'abai', 'abbee', 'abdul', 'acara', 'acaratks', 'account', 'ada', 'adapromo', 'adi', 'adik', 'adison', 'admin', 'administrasi', 'adminlte', 'ado', 'adrian', 'adu', 'aduh', 'advertising', 'aea', 'aesthetic', 'afbe', 'affc', 'afr', 'afrika', 'agam', 'agen', 'agendain', 'agenpulsa', 'ags', 'agst', 'agsts', 'agt', 'agtskinfodlj', 'agua', 'agun', 'agus', 'agust', 'agustus', 'agustuskunjungi', 'ah', 'ahaha', 'ahub', 'aidzin', 'aigoo', 'air', 'aja', 'ajaa', 'ajaaa', 'ajabri', 'ajak', 'ajakin', 'ajar', 'ajeng', 'akademik', 'akang', 'akangteteh', 'akbar', 'akreditasi', 'akses', 'aksi', 'aktif', 'aktifasi', 'aktivasi', 'aktivitas', 'akucintaislam', 'akumulasi', 'akun', 'akurasi', 'akurat', 'ala', 'alaikum', 'alaikumsaya', 'alaiqum', 'alam', 'alamat', 'alami', 'alamsyah', 'alat', 'alesannya', 'algoritma', 'alhamdulillah', 'alhamdullilah', 'alhuda', 'ali', 'aliando', 'all', 'allah', 'allahaamiin', 'alphard', 'alquran', 'alur', 'aman', 'amanda', 'ambil', 'amin', 'ampuun', 'an', 'anabdullah', 'anak', 'ananda', 'and', 'anda', 'andaafr', 'andabch', 'andabfgs', 'andac', 'andadidinfo', 'andaerd', 'andafah', 'andafe', 'andafg', 'andahhh', 'andaijhk', 'andajf', 'andam', 'andamcxj', 'andarman', 'andasyaratcopy', 'andatb', 'andaxvf', 'andaxxxxx', 'andaytg', 'andi', 'andika', 'andipublishercompobnibank', 'andri', 'andrianto', 'android', 'androidmu', 'andromax', 'aneka', 'ang', 'angga', 'anggar', 'angka', 'angkot', 'angpaopoinsenyum', 'angsur', 'angunan', 'ani', 'animasi', 'anomali', 'anrasto', 'antar', 'antivirus', 'anto', 'anu', 'anwahyudi', 'anyar', 'anyonghaseyo', 'ap', 'apa', 'apaapa', 'apapa', 'apiapi', 'aplikasi', 'app', 'apple', 'appleplay', 'apply', 'appreciation', 'apps', 'appsinapps', 'april', 'aqila', 'aqua', 'ardiansa', 'ardiansah', 'ardiansyah', 'area', 'arengnya', 'argo', 'ari', 'arif', 'armada', 'artha', 'artisartis', 'arum', 'arya', 'aryadhi', 'aryaxxx', 'as', 'asa', 'asalasalan', 'asam', 'asap', 'ashar', 'asing', 'asisten', 'asli', 'asmul', 'ass', 'assalamu', 'assalamualaikum', 'assmaaf', 'asssy', 'asumsi', 'asuransi', 'at', 'atas', 'atauhub', 'atay', 'atm', 'ato', 'atuh', 'atuhlah', 'atulah', 'atur', 'aty', 'au', 'aug', 'augsept', 'aunul', 'avanza', 'avanzadari', 'aware', 'awas', 'away', 'awet', 'axiata', 'axis', 'axismu', 'axisnet', 'ayah', 'ayam', 'ayena', 'ayeuna', 'ayo', 'ayook', 'ayu', 'ayyuuuukk', 'baa', 'bab', 'babakaran', 'baberque', 'baca', 'backup', 'bagi', 'bagus', 'bahagia', 'bahan', 'bahas', 'bahasa', 'bahaya', 'bahri', 'baik', 'bakar', 'bakarkremes', 'bakmi', 'bakti', 'bal', 'balai', 'balas', 'balitrims', 'ball', 'balon', 'bambang', 'bandar', 'bandara', 'bandros', 'bandung', 'bandungg', 'bandungmulai', 'bang', 'banget', 'bangun', 'bank', 'bankbni', 'banting', 'bantu', 'bantuin', 'banyak', 'bapa', 'bapak', 'barang', 'bareng', 'barokahmp', 'baru', 'baruqskrng', 'barusan', 'basdat', 'baskin', 'batam', 'batas', 'batasankami', 'batasfupkecepatan', 'bati', 'batin', 'batman', 'bawa', 'bawah', 'bawain', 'bayang', 'bayar', 'bayardi', 'bayu', 'bb', 'bbade', 'bbd', 'bbm', 'bc', 'bca', 'bch', 'bchuinfo', 'bddfd', 'beaok', 'beban', 'bebas', 'bebek', 'beberes', 'beda', 'begitusayang', 'bekal', 'bekas', 'bekerjasama', 'belah', 'belanja', 'beli', 'beliau', 'belikan', 'belikn', 'beliung', 'bella', 'belok', 'bener', 'beneran', 'benerin', 'bensin', 'bentar', 'bentuk', 'berangkat', 'berangkt', 'berangktnya', 'beranta', 'berapaa', 'beras', 'berat', 'beres', 'berfikir', 'berita', 'berkah', 'berkas', 'berkat', 'berlakupromo', 'berminatsy', 'berrybenka', 'bersih', 'besar', 'besok', 'besoook', 'besto', 'beu', 'bfgs', 'bget', 'bgi', 'bgmana', 'bh', 'bhw', 'bhwa', 'bi', 'biar', 'biasa', 'biasana', 'biaya', 'biber', 'bicara', 'bidang', 'big', 'bih', 'bikin', 'bilang', 'bima', 'bimatri', 'bimbing', 'bingung', 'bintang', 'bioskop', 'bip', 'bis', 'bisah', 'bisakami', 'bisanga', 'bisangga', 'bisnis', 'bisns', 'bitlytentangcigna', 'bjam', 'bl', 'bla', 'blackberry', 'blackberrydakota', 'blackberryiphone', 'blanja', 'blanjacom', 'bliin', 'bliung', 'bljcotcashrabu', 'bljrpenguatan', 'blog', 'blogspotcom', 'blokir', 'bmkg', 'bni', 'bns', 'bntar', 'bobo', 'bobok', 'bocor', 'boladpt', 'bolak', 'bom', 'bonus', 'bonusnyacekbns', 'booking', 'borma', 'bos', 'box', 'bpa', 'bpjs', 'bpkb', 'bpkbsertipikatproses', 'bpkdedy', 'bpkibu', 'bpkyogi', 'bpr', 'brati', 'brencana', 'brgkt', 'brh', 'bri', 'brianasri', 'brinorekanir', 'britama', 'brminat', 'brminatcocoktlong', 'brngktnya', 'bro', 'bronet', 'bronetjam', 'bronetspecial', 'brpnya', 'brsik', 'brunya', 'bsediasms', 'bt', 'btpn', 'bts', 'bu', 'buada', 'buat', 'buatau', 'bubuka', 'budget', 'budiantar', 'bug', 'bujuk', 'buka', 'bukalapak', 'bukti', 'buktikeun', 'bulan', 'buletdatar', 'bully', 'bumaaf', 'bumau', 'bumbu', 'bumenu', 'bumiputera', 'bunas', 'bunda', 'bunga', 'bungkus', 'burger', 'bursa', 'buru', 'business', 'button', 'butuh', 'buy', 'bwrangkat', 'by', 'cabai', 'cabut', 'cafe', 'cahaya', 'cair', 'callsms', 'calon', 'cam', 'camera', 'cantik', 'cantum', 'capai', 'capek', 'cappuccino', 'card', 'care', 'carecom', 'careselamat', 'cari', 'caringinsadang', 'carrier', 'cas', 'case', 'cashback', 'cashkredit', 'casino', 'caturgadgetymailcom', 'cc', 'ccfa', 'cckt', 'cckta', 'cd', 'cecer', 'cegah', 'cek', 'cekrpjuta', 'cektunai', 'celaka', 'celana', 'cell', 'celluler', 'cema', 'cenah', 'center', 'cepat', 'cerita', 'cetak', 'cewek', 'cfc', 'challenge', 'chandra', 'charged', 'chartered', 'chat', 'chatnya', 'chatnyaa', 'chatting', 'cheking', 'chicken', 'china', 'chip', 'chn', 'choice', 'ciateul', 'cigna', 'cilimus', 'cinta', 'cip', 'cipika', 'ciprat', 'city', 'ckp', 'ckup', 'client', 'clip', 'cliponyu', 'cmn', 'cnf', 'coba', 'cobacoba', 'cobagrabcar', 'coboy', 'cochochips', 'cocok', 'code', 'coins', 'cokelat', 'colok', 'com', 'combo', 'construct', 'controler', 'controller', 'controllernya', 'copy', 'copyin', 'corp', 'costumer', 'cover', 'cowok', 'cowonya', 'cp', 'cpy', 'cream', 'credit', 'creditcard', 'crispy', 'cs', 'csps', 'csv', 'cuci', 'cukup', 'cupu', 'curi', 'cuss', 'customer', 'cuti', 'da', 'dacota', 'dadak', 'dadi', 'daerah', 'daftar', 'daging', 'dahsyat', 'dakota', 'dakotajt', 'dakwah', 'dambil', 'dan', 'dana', 'danamon', 'dapat', 'dapet', 'dapetkan', 'daptkn', 'dasar', 'data', 'database', 'datang', 'daun', 'daves', 'davis', 'davisrb', 'daya', 'dayeuh', 'db', 'dblkg', 'dcs', 'dddss', 'ddmmyyyy', 'de', 'dear', 'debat', 'dediktp', 'dee', 'deh', 'dekat', 'dekstop', 'delete', 'demo', 'dengar', 'dengerin', 'dengna', 'denpasar', 'depaaan', 'depan', 'depilkom', 'deposit', 'dermawan', 'derr', 'des', 'desember', 'desi', 'desila', 'desilla', 'desjan', 'desy', 'detail', 'deteksi', 'detik', 'develop', 'device', 'dgninternetgaul', 'dhi', 'diagnosa', 'dian', 'diantapani', 'dianterin', 'diapain', 'dibalikin', 'dibandingin', 'dibarengin', 'dibawain', 'dibutuhin', 'dicontroler', 'dicotroller', 'did', 'didr', 'diem', 'digantiin', 'diii', 'diiket', 'diil', 'diinvite', 'dijadiin', 'dijmput', 'dika', 'dikosan', 'dikostan', 'dikpad', 'dikti', 'dil', 'dila', 'diload', 'dimasukin', 'dinamic', 'dinas', 'dinding', 'dine', 'diomongin', 'diperpanjangnikmati', 'diprog', 'direct', 'direktur', 'direpository', 'diri', 'disatuin', 'disc', 'discond', 'discound', 'disesuaikanlebih', 'diskon', 'diskusi', 'disurvey', 'dit', 'ditelfon', 'ditempatin', 'diteruskanbayar', 'ditks', 'ditravel', 'ditungguuu', 'diturunin', 'diy', 'djunaedi', 'dki', 'dkrm', 'dlk', 'dllinfo', 'dluuntuk', 'doa', 'dokter', 'dokumen', 'dolar', 'domestic', 'domestik', 'dominos', 'donasi', 'donut', 'donuts', 'dooong', 'dosa', 'double', 'doublescoop', 'down', 'download', 'doyan', 'dp', 'dptkan', 'dptkn', 'draftnya', 'drakor', 'drhdarmawan', 'drhirawan', 'drhjsopian', 'drive', 'driver', 'dropdown', 'drpengundian', 'drptindosat', 'drrimbawandirmawa', 'drsh', 'drshmulyadi', 'drshnaldisuparman', 'drshsrianto', 'dscn', 'dss', 'dt', 'dtggu', 'dtgu', 'dtng', 'du', 'dua', 'duaduanya', 'duduk', 'duh', 'duit', 'dukung', 'duluan', 'dulumaaf', 'dummy', 'dunkin', 'dwi', 'dwm', 'dyhaha', 'eastern', 'eatboss', 'edan', 'edisi', 'edit', 'efgr', 'egi', 'eh', 'eka', 'eks', 'eksekusi', 'ekspresi', 'ekstra', 'elecktronik', 'elektrik', 'elektronik', 'email', 'emang', 'emasasia', 'emil', 'emlb', 'enak', 'enakaa', 'end', 'engkau', 'english', 'enhaii', 'ente', 'entenipenteng', 'er', 'error', 'errorhrus', 'ertiga', 'es', 'esok', 'essay', 'estimasi', 'eta', 'eticket', 'etik', 'etiket', 'euy', 'evaluasi', 'ewu', 'excellent', 'excelso', 'exhibition', 'experience', 'external', 'extra', 'facebook', 'fany', 'farida', 'fashion', 'fasilitas', 'fatih', 'fatima', 'favorit', 'fb', 'fc', 'fcopy', 'feb', 'feedback', 'felix', 'fg', 'fhjtl', 'fhx', 'field', 'file', 'filenya', 'film', 'finance', 'fip', 'fir', 'firli', 'first', 'fitutnya', 'fix', 'flash', 'flashspasiinfo', 'flashspasiya', 'flat', 'flatbln', 'flavour', 'folder', 'followloop', 'food', 'forest', 'form', 'format', 'formnya', 'fotmms', 'foto', 'fotomms', 'fpmipa', 'free', 'freed', 'freedom', 'freedoom', 'freefor', 'freza', 'fried', 'fries', 'friska', 'frown', 'ft', 'full', 'fungsi', 'further', 'futsal', 'gaaa', 'gaaaa', 'gaaais', 'gaaak', 'gabawa', 'gabung', 'gabungken', 'gadget', 'gadiabsen', 'gaes', 'gaga', 'gagal', 'gagas', 'gaib', 'gais', 'gajah', 'gaji', 'galau', 'galaxy', 'galeri', 'gallery', 'game', 'gamegame', 'gamepro', 'gameprog', 'games', 'gamesmu', 'gampang', 'gang', 'ganti', 'gantiinpenting', 'gara', 'garansi', 'garasi', 'garsindo', 'gas', 'gateway', 'gatis', 'gatot', 'gaya', 'gbberlaku', 'gbg', 'gbgb', 'gbgbg', 'gbgbonus', 'gbggb', 'gbhr', 'gbjt', 'gbkonten', 'gbnelpon', 'gdg', 'ge', 'gebyar', 'gedorin', 'gedung', 'gelang', 'gelas', 'geleh', 'gemastik', 'gemini', 'geminirbq', 'gempa', 'geng', 'gening', 'genuine', 'gera', 'gerai', 'gerlong', 'get', 'getlucky', 'gf', 'ghari', 'gi', 'gift', 'gigi', 'gik', 'gila', 'gilisms', 'git', 'gita', 'github', 'gituch', 'gitulah', 'glte', 'gm', 'gmana', 'gmau', 'gmmu', 'gnkan', 'gobigococ', 'gojek', 'gojekgotmehere', 'gokana', 'gokil', 'goldjt', 'gombal', 'google', 'goreng', 'gorengin', 'grab', 'grabcar', 'grabtaxi', 'grabtravel', 'graha', 'grand', 'grapari', 'gratis', 'group', 'groupwhere', 'groupwherenya', 'grup', 'gsm', 'gsmcdmavvutk', 'gtau', 'gtshop', 'gudang', 'gum', 'guna', 'gunakn', 'guntur', 'guru', 'gusti', 'guy', 'guys', 'gymnas', 'haagendasz', 'habis', 'hadap', 'hadiah', 'hadiahkunjungi', 'hadiahlik', 'hadiahrp', 'hadian', 'hadih', 'hadir', 'haha', 'hahaha', 'hahahaha', 'hai', 'haji', 'hak', 'halang', 'hallo', 'haluan', 'handphone', 'handy', 'handycam', 'hangat', 'hanum', 'hapapa', 'har', 'harap', 'hard', 'hardcover', 'harddisk', 'harga', 'hari', 'harijika', 'harikuota', 'harino', 'harirprb', 'haris', 'haritgl', 'hartanto', 'harum', 'hasan', 'hasil', 'hati', 'hatta', 'hatur', 'hayu', 'hayuu', 'hbngi', 'hd', 'hdr', 'hdrjn', 'hebat', 'heboh', 'hehe', 'hehehe', 'hehehehe', 'helmkuuuu', 'hemat', 'hendra', 'henti', 'hermawati', 'heroes', 'heula', 'hg', 'hggmbpstekan', 'hhb', 'hher', 'hi', 'hibur', 'hidup', 'high', 'hihi', 'hihip', 'hihu', 'hijau', 'hirenya', 'hit', 'hitam', 'hits', 'hitung', 'hj', 'hjsuri', 'hk', 'hmm', 'hmmm', 'hobi', 'home', 'honda', 'hongkongcarapake', 'hormat', 'host', 'hot', 'hotice', 'hp', 'hpjam', 'hpnya', 'hpterima', 'hquarters', 'hrap', 'hrian', 'hsulaiman', 'htl', 'httpgebyartelkomseljuplocom', 'hubagen', 'hubcs', 'hubcustumer', 'hubirhjmuliyadi', 'hubirhsurianto', 'hubjasa', 'hubkunjungi', 'hublolyta', 'hubsms', 'hubtks', 'hubung', 'huhuhu', 'hujan', 'hunting', 'hut', 'hypermind', 'ibadah', 'ibu', 'ibuanita', 'ice', 'id', 'ide', 'identitas', 'ieu', 'ih', 'ii', 'ijazah', 'ijhk', 'ik', 'ikan', 'ikhlas', 'iklan', 'ikut', 'ilkom', 'ilkomif', 'ilmu', 'im', 'iman', 'in', 'inap', 'inch', 'indeks', 'indofood', 'indomarcocell', 'indomaret', 'indomie', 'indomobil', 'indonesia', 'indonesiabunga', 'indosat', 'indosatooredoo', 'indosatooredoocomfreedom', 'indosatooredoointernet', 'indosatpoin', 'indosy', 'indovison', 'infohubsms', 'infojk', 'infokunjungi', 'infold', 'infomg', 'informasi', 'informasinyharap', 'infosms', 'infosmspin', 'ingat', 'ingetin', 'inih', 'inimandirino', 'inipromo', 'iniselama', 'iniselamatnomor', 'input', 'inspirasi', 'instagram', 'instal', 'install', 'installernya', 'instan', 'instances', 'instansi', 'instansiasi', 'instink', 'insya', 'insyaaallah', 'intake', 'intan', 'intermediate', 'internasional', 'international', 'internet', 'internetan', 'internetas', 'internetmu', 'interview', 'inti', 'intl', 'intrviewrekom', 'investasi', 'iot', 'iotak', 'ip', 'ipa', 'ipad', 'ipadsamsung', 'ipbythjajang', 'ipc', 'iphone', 'iphonipad', 'ir', 'irawan', 'ird', 'iring', 'irit', 'irj', 'irwardoyo', 'iseng', 'isi', 'isiin', 'isiulang', 'iskandarse', 'islam', 'ismawati', 'isti', 'istmwa', 'isya', 'it', 'items', 'iti', 'iur', 'iwan', 'ixindo', 'iya', 'iyaa', 'iyaaa', 'iyaaaa', 'izin', 'jabat', 'jadwal', 'jaga', 'jagain', 'jaguar', 'jahaaat', 'jahe', 'jajan', 'jajar', 'jakarta', 'jakartainfo', 'jalan', 'jalanken', 'jam', 'jamaah', 'jamber', 'jambi', 'jamgb', 'jamin', 'jan', 'janji', 'januari', 'japri', 'jarak', 'jarang', 'jaring', 'jarkom', 'jatuh', 'jauh', 'jawa', 'jaya', 'jazz', 'jeans', 'jeblok', 'jebol', 'jelas', 'jelek', 'jemari', 'jember', 'jemput', 'jend', 'jendral', 'jenis', 'jepang', 'jepret', 'jeruk', 'jf', 'jica', 'jilbab', 'jilid', 'jins', 'jitu', 'jiwa', 'jljend', 'jlpatuha', 'jne', 'jobfairactive', 'jodoh', 'jogja', 'joko', 'jokodi', 'jquery', 'jr', 'js', 'jt', 'jta', 'jtayo', 'jtbila', 'jtbln', 'jual', 'jualsy', 'juang', 'juiaug', 'juke', 'jukeayo', 'jul', 'juli', 'jumat', 'jun', 'juni', 'junior', 'junitukarkan', 'junketik', 'jurus', 'jusman', 'juta', 'jutasy', 'ka', 'kabar', 'kabarin', 'kabel', 'kacamata', 'kadang', 'kadisdik', 'kado', 'kagi', 'kah', 'kait', 'kak', 'kaka', 'kalah', 'kalahmau', 'kalapa', 'kaldaluarsa', 'kalem', 'kali', 'kaliinfo', 'kalimat', 'kalo', 'kamar', 'kamera', 'kamibfi', 'kamis', 'kampuang', 'kampung', 'kampus', 'kamrin', 'kang', 'kantin', 'kanto', 'kantor', 'kaprodi', 'kar', 'karakter', 'karet', 'karimbeliau', 'kariyawan', 'karso', 'kartu', 'kartuhalo', 'karyawani', 'kas', 'kashnbtlng', 'kasih', 'kasihin', 'kasur', 'katalog', 'kategori', 'kau', 'kav', 'kawal', 'kaya', 'kayak', 'kb', 'kbdapatkan', 'kdgkan', 'kebapain', 'kebaperan', 'kebebasa', 'keburu', 'kebut', 'kecewa', 'kecuali', 'kedlmn', 'kei', 'keik', 'keisatinternet', 'kejar', 'keju', 'kejut', 'kekalapa', 'kelapa', 'kelas', 'kelik', 'keliling', 'kelompok', 'keluar', 'keluarga', 'kemakom', 'kemanamana', 'kemaren', 'kemarin', 'kembaliin', 'kembang', 'kena', 'kenal', 'kenan', 'kenyang', 'kep', 'kepadabpkibu', 'kepmentawaisumbar', 'kepung', 'kereg', 'keren', 'kering', 'kerja', 'kerudung', 'kesana', 'kesini', 'ketemu', 'ketik', 'ketikdaftar', 'ketikinfo', 'kewirausahaan', 'keyboard', 'kfc', 'kg', 'kghk', 'khawatir', 'khilaf', 'kholiknimedi', 'khusus', 'ki', 'kikuk', 'kilat', 'kindly', 'kini', 'kirain', 'kirim', 'kirimin', 'kirimmakasih', 'kisar', 'kita', 'kiwi', 'kke', 'kl', 'klasifikasi', 'klau', 'klick', 'klik', 'klikhttp', 'klikhttptriwebnodecom', 'kluarganya', 'kmblimaaf', 'kmiuntuk', 'ko', 'kocak', 'kode', 'kodeuniq', 'kodevrd', 'koesmaryonomsdi', 'koin', 'koja', 'kolam', 'koleksi', 'kolom', 'kolot', 'komar', 'komik', 'komisi', 'kompas', 'kompensasi', 'kompetisi', 'kompor', 'komunikasi', 'kondisiy', 'koneksi', 'konfirmasi', 'kontak', 'kontra', 'kontrak', 'kontrol', 'koper', 'kopi', 'koreksi', 'kos', 'kosan', 'kosannya', 'kosong', 'kota', 'kpnnnih', 'kpopers', 'kpopmu', 'kredit', 'kremes', 'krim', 'krimkn', 'ktp', 'ktpkkbuku', 'ktr', 'ku', 'kualiatasnya', 'kuasa', 'kuat', 'kudu', 'kuliah', 'kuliahhal', 'kuliner', 'kulkas', 'kumpul', 'kunci', 'kunjung', 'kuota', 'kuotabeli', 'kuotacuma', 'kuotaspasik', 'kupon', 'kur', 'kurban', 'kurus', 'kusir', 'kwkw', 'laa', 'laaah', 'lab', 'labmatakuliah', 'lada', 'lady', 'ladyticketwebscom', 'lagistok', 'lahir', 'lain', 'laksana', 'laku', 'lala', 'lamar', 'lamarancvijazahphotox', 'lampir', 'lampu', 'lancar', 'langgan', 'langka', 'langsung', 'lanjut', 'lanjutin', 'lantai', 'lapang', 'lapor', 'laptop', 'laptopcameradll', 'laptopdll', 'laptopkamera', 'lari', 'larissa', 'latih', 'latte', 'layan', 'layar', 'lazada', 'ld', 'learningnya', 'lebaranpaket', 'lebay', 'lebib', 'lebih', 'ledeng', 'legal', 'lejit', 'lelah', 'lemari', 'lembang', 'lembar', 'lengkap', 'lenkap', 'lepas', 'lestari', 'lets', 'level', 'lewatin', 'lgs', 'lho', 'lia', 'liar', 'libur', 'lieur', 'lihat', 'lihatin', 'limit', 'limpah', 'line', 'lingkung', 'lipat', 'lipstik', 'lisan', 'listrik', 'listrikspeedytelkom', 'lm', 'lnfo', 'load', 'lobi', 'localhostsistemrumahsakitgerbangrawatdarurat', 'logo', 'loh', 'lok', 'lokalnikmati', 'lokasi', 'lombok', 'longsor', 'loopcoid', 'loopers', 'lottemart', 'love', 'ls', 'lsg', 'ltaquakembalian', 'lte', 'luang', 'lucu', 'luculucu', 'lukis', 'lukman', 'lulu', 'lulus', 'luluuus', 'lumayan', 'lunas', 'lupa', 'lupaaa', 'lurah', 'lv', 'lychee', 'maaaaaaaaf', 'maacih', 'maaf', 'maafin', 'maafsaya', 'maapin', 'macbook', 'macet', 'machine', 'mad', 'mag', 'maghrib', 'magicomnya', 'mah', 'mahal', 'mahar', 'mahasiswa', 'mahesian', 'mahmud', 'main', 'makalah', 'makan', 'makasihh', 'makasihhh', 'maker', 'maks', 'maksimal', 'maksud', 'mal', 'malam', 'malas', 'malaysia', 'mall', 'mama', 'mamah', 'mana', 'manaada', 'mandi', 'mandiri', 'manfaat', 'mang', 'mangga', 'manis', 'mantapok', 'manunggal', 'mar', 'mara', 'marah', 'maranggi', 'maret', 'mari', 'mariadi', 'marketing', 'mas', 'masak', 'masalah', 'masingmasing', 'massal', 'master', 'mastertogel', 'masuk', 'masukin', 'mataharimall', 'materi', 'mati', 'mayani', 'maykunjungi', 'mazda', 'mbah', 'mbak', 'mbhr', 'mbmbbonus', 'mbntu', 'mcdonalds', 'mdapat', 'mdpat', 'mdpt', 'mdptakan', 'mdptkam', 'mdptkan', 'medan', 'medptkan', 'medrec', 'meet', 'mega', 'meganai', 'mei', 'meijuni', 'meja', 'mel', 'melanjutkantunggu', 'melapormhub', 'melulu', 'member', 'men', 'menang', 'mending', 'mendpatkn', 'mendpt', 'mendptkn', 'menggangu', 'menghub', 'meni', 'menit', 'menithr', 'menittlp', 'menitvideo', 'mentah', 'menu', 'merah', 'merchant', 'merchants', 'merdeka', 'meren', 'merhatin', 'merica', 'mesan', 'metro', 'mg', 'mgksegera', 'mgrib', 'mh', 'mhs', 'mhsi', 'mhub', 'mhubungi', 'middle', 'midnight', 'mido', 'mie', 'mifi', 'mifimodem', 'migrasi', 'miliard', 'milik', 'million', 'milyar', 'min', 'minat', 'mineral', 'minggu', 'minggudaftar', 'mini', 'minimal', 'minimum', 'minimun', 'mining', 'minna', 'minta', 'minum', 'miring', 'mk', 'mkios', 'mkshgratis', 'mlahkgsngtwnvtmmacau', 'mlmktkgratis', 'mmber', 'mnctv', 'mndpt', 'mndptkn', 'mngenai', 'mnghubungi', 'mningkattelepon', 'mnthr', 'mntsms', 'mnyediakn', 'mnyesuaikan', 'moal', 'mobcom', 'mobil', 'mobiltruck', 'mobprog', 'mock', 'modal', 'model', 'modelcontroller', 'modem', 'moga', 'mohon', 'monitor', 'more', 'motor', 'motorolie', 'movie', 'mpd', 'mpulsa', 'mraih', 'ms', 'msalh', 'msceis', 'msi', 'msk', 'mtronik', 'mtronil', 'mu', 'mubazir', 'muda', 'mudah', 'mudahaman', 'muka', 'mul', 'mulai', 'muliani', 'multiple', 'mulyadi', 'mumpung', 'muncul', 'murah', 'murdoyo', 'murid', 'murtiani', 'musik', 'my', 'mycare', 'mytelkomsel', 'myvplus', 'myxlcoid', 'na', 'nabung', 'nagamas', 'naha', 'nak', 'nama', 'namagraha', 'namakuuu', 'nang', 'nanggung', 'naonaridyah', 'narti', 'nasabah', 'nasehat', 'nasi', 'nasiayamlalapansambal', 'nasib', 'nasibberbagi', 'nasibnutrisi', 'ndisek', 'ndut', 'negara', 'negarano', 'nego', 'negosiasi', 'nelpsms', 'nenek', 'net', 'netmax', 'new', 'newdengan', 'nfo', 'ngajak', 'ngajakan', 'ngajakin', 'ngaku', 'ngan', 'ngapain', 'ngasih', 'ngasihinnya', 'nge', 'ngebatin', 'ngebuild', 'ngecc', 'ngedeteksi', 'ngeditnya', 'ngeformat', 'ngegrab', 'ngeh', 'ngehubungin', 'ngerangkul', 'ngeroko', 'ngerti', 'ngevape', 'ngidam', 'nginget', 'ngobrol', 'ngoding', 'ngolah', 'ngomong', 'ngomongin', 'ngubek', 'ngumpul', 'ngurus', 'ni', 'niagacell', 'niat', 'nih', 'nikah', 'nikmat', 'nila', 'nilai', 'ninggalin', 'nipis', 'nissan', 'nitip', 'nmrku', 'nnti', 'no', 'noah', 'noanda', 'nohp', 'noid', 'nokia', 'nomor', 'nomr', 'non', 'nonaktif', 'nondik', 'nongkrong', 'nonton', 'nopin', 'nopinbl', 'nopinfg', 'nopingjk', 'norek', 'norekan', 'norekening', 'normal', 'nosim', 'nosimcard', 'note', 'notip', 'nova', 'november', 'noviit', 'novitasari', 'noxv', 'noxxx', 'npwp', 'nrml', 'nsp', 'ntb', 'nu', 'nuhun', 'nuhuno', 'nuhuun', 'nuju', 'nulis', 'numpang', 'nunggu', 'nur', 'nusa', 'nuy', 'nxmvfue', 'nya', 'nyaah', 'nyakitin', 'nyala', 'nyalain', 'nyalira', 'nyaman', 'nyari', 'nyerahin', 'nyesuain', 'nyetak', 'nyetok', 'nyimper', 'nyoba', 'nyumbang', 'nyusul', 'obat', 'obor', 'obral', 'oct', 'odp', 'of', 'off', 'office', 'offline', 'oge', 'oh', 'oilgop', 'oiya', 'ok', 'okayyy', 'okcall', 'oke', 'okeee', 'okeyyy', 'okt', 'oktra', 'okya', 'oleh', 'olshop', 'one', 'onix', 'online', 'onyx', 'oo', 'ooh', 'ooredohanya', 'ooredoo', 'operator', 'opoerator', 'oprator', 'opsi', 'optimis', 'orang', 'order', 'orek', 'orlando', 'ortu', 'osnnya', 'otak', 'otomatis', 'otp', 'ouhhh', 'output', 'outputnya', 'overreact', 'overthink', 'pa', 'pacar', 'paedi', 'pagar', 'pagi', 'pagimau', 'paham', 'pai', 'pajuin', 'pakai', 'pakbeliau', 'paket', 'paketaktifkan', 'paketextra', 'pakettelkomselflash', 'pakooh', 'paksa', 'palasari', 'pan', 'pandang', 'pandu', 'panggil', 'pangmeserkeun', 'pangmeulikeun', 'panjang', 'panpizzacoid', 'pantai', 'pantesan', 'papa', 'parafrase', 'parah', 'pariksa', 'parkir', 'partner', 'pas', 'pasang', 'pasangbonus', 'pasar', 'pasca', 'pasien', 'pasir', 'paspor', 'passengers', 'paste', 'pasti', 'patah', 'payment', 'pc', 'pcs', 'pd', 'pdf', 'pdg', 'pedoman', 'pegang', 'pegat', 'pelaksanapt', 'pelaksanatelkomsel', 'peluang', 'pemakaianinfo', 'pemberitahuan', 'pemenangaiue', 'pemenangbl', 'pemenanger', 'pemenanglm', 'pemenangnyahati', 'pemuda', 'pen', 'penginputan', 'pengkol', 'pengundiang', 'penring', 'pentaho', 'penting', 'penuh', 'pepaya', 'per', 'perbln', 'percaya', 'pergi', 'perhati', 'perihal', 'periksa', 'perilaku', 'periode', 'perkedel', 'pernuh', 'persero', 'pertable', 'pertamina', 'pertaminareckrutmentymailcom', 'pertanggungjawaban', 'pesan', 'pesawat', 'pesta', 'ph', 'phone', 'pikir', 'piknikada', 'pilih', 'pilkada', 'pimpin', 'pin', 'pinab', 'pinach', 'pinaiue', 'pinbbaff', 'pinbbc', 'pinbchdi', 'pinbfgs', 'pinbg', 'pinbh', 'pinca', 'pindah', 'pindhi', 'pindhiutk', 'pinfcd', 'pinggir', 'pinggirmulai', 'pinhadiah', 'pinj', 'pinjam', 'pinjf', 'pinmcxj', 'pinmhr', 'pinmk', 'pinnf', 'pinpemenang', 'pinrwd', 'pinrwf', 'pintarkursus', 'pintblvg', 'pinv', 'pinxaf', 'pinxbv', 'pinyrkz', 'pisaaaan', 'pisaaan', 'pizza', 'pizzaroti', 'pizzarotinya', 'pkl', 'pla', 'plaassalambagi', 'pladishare', 'planetcare', 'plasatelkommlg', 'platempat', 'platihan', 'play', 'playapple', 'plaza', 'plenetcell', 'plgn', 'plgyth', 'pln', 'plntelkom', 'plus', 'plusplus', 'plz', 'pm', 'pmenan', 'pmnang', 'pnjaman', 'pnwrn', 'poin', 'point', 'pokemon', 'poko', 'pokok', 'polisi', 'polisijangan', 'pom', 'pon', 'ponsel', 'pool', 'poor', 'pop', 'popin', 'popmie', 'porsi', 'posisi', 'positif', 'poskamling', 'post', 'potensi', 'ppl', 'praktek', 'praktikum', 'praktis', 'prasetya', 'prasidang', 'prasodjo', 'prastia', 'pratama', 'prb', 'prediksi', 'prelaunchprofit', 'premium', 'presentasi', 'prestasi', 'presto', 'pret', 'pribadi', 'pribdi', 'prihal', 'prima', 'print', 'printer', 'priode', 'prize', 'prodi', 'prof', 'profdriryonny', 'prog', 'program', 'programmer', 'progres', 'project', 'promo', 'promosi', 'proposal', 'prosedur', 'proses', 'proseshari', 'pt', 'ptaxis', 'ptbank', 'ptbni', 'ptcare', 'pthm', 'pthutchison', 'ptindofood', 'ptindofoodtbk', 'ptindosat', 'ptm', 'ptmkios', 'ptmtronik', 'ptplanet', 'ptpos', 'pttelkomsel', 'pttri', 'pttricare', 'pttrisimcard', 'ptx', 'puas', 'puasa', 'publik', 'publish', 'pucakti', 'pucuk', 'puji', 'pukulwib', 'pul', 'pulang', 'pull', 'pulsa', 'pulsaall', 'pulsabayar', 'pulsacek', 'pulsalimit', 'pulsamunikmati', 'pulsapin', 'pulsarb', 'puluhanjt', 'purwakarta', 'purwanto', 'pus', 'pusat', 'pusenkav', 'push', 'pusing', 'put', 'putar', 'putera', 'putri', 'putus', 'pvj', 'qaqa', 'quota', 'quotes', 'rabu', 'radiologi', 'raditya', 'rafael', 'ragu', 'rahasia', 'rahmad', 'rahman', 'rahmat', 'raih', 'raja', 'rakyat', 'ram', 'ramadhan', 'ramai', 'ramen', 'rancak', 'random', 'randy', 'range', 'rapat', 'rapi', 'rapih', 'rasa', 'rasti', 'rastina', 'ratarata', 'rate', 'ratna', 'ratus', 'rawat', 'rawatdaruatsql', 'rawit', 'rawitinfold', 'rawosi', 'rbbulan', 'rbharga', 'rbisiulang', 'rbkana', 'rbminggu', 'rbnrml', 'rbsmsstop', 'rbu', 'rbutk', 'rcti', 'rctiinfo', 'rd', 'real', 'rebu', 'rebus', 'red', 'reda', 'redeem', 'referensi', 'referral', 'refgratis', 'reg', 'register', 'registrasi', 'regular', 'reguler', 'rehabilitasi', 'rejeki', 'rejekinya', 'rek', 'rekening', 'rel', 'reloadpasti', 'remed', 'remedial', 'renang', 'rencang', 'rendah', 'renti', 'repot', 'reservasi', 'residence', 'resin', 'reski', 'resky', 'resmi', 'respon', 'responnya', 'restart', 'resti', 'resto', 'restu', 'retuwal', 'reuni', 'revisi', 'reza', 'rezky', 'ri', 'riah', 'ribet', 'ribu', 'ributrip', 'rica', 'rich', 'ride', 'riiweh', 'rindu', 'ringan', 'risal', 'riung', 'rj', 'rla', 'rm', 'rmhtanahmobilprodukjasa', 'robbins', 'roi', 'rokok', 'roni', 'ros', 'route', 'roxi', 'rpjt', 'rpjtdari', 'rpjtdr', 'rpjtuinfo', 'rpjuta', 'rpjutadri', 'rpkunjungi', 'rpmggcs', 'rprb', 'rprbada', 'rprbberlaku', 'rprbbln', 'rprbhariinfo', 'rpribu', 'rpsms', 'rptukarkan', 'rs', 'rsdemikian', 'rsnopin', 'rsuinfo', 'rtx', 'ruang', 'rudi', 'rudy', 'ruma', 'rumah', 'rumahtanah', 'rumit', 'rumput', 'rundown', 'rupiah', 'rusak', 'rutin', 'rwd', 'sabar', 'sabtu', 'sadar', 'sadikin', 'safei', 'safi', 'safitri', 'saha', 'sahabat', 'sahkan', 'sakit', 'sal', 'salah', 'salam', 'salat', 'saldo', 'sale', 'samain', 'samarinda', 'samawa', 'sambl', 'sambung', 'samping', 'sampk', 'sampoerna', 'samsung', 'sana', 'sananha', 'sandwich', 'sapa', 'sapi', 'sate', 'satpam', 'satu', 'satusatunya', 'saudara', 'save', 'sayabpk', 'sayang', 'sayapnasiteh', 'sayappaha', 'sayur', 'sbagai', 'sbelumhya', 'sbsar', 'scond', 'scoop', 'score', 'scroll', 'sctv', 'sd', 'sdgkn', 'sdnov', 'sdr', 'sebab', 'sebar', 'sebenarny', 'sebenernya', 'sebentar', 'seberang', 'seda', 'sedia', 'sedih', 'see', 'seengganya', 'segan', 'segelehgelehnya', 'segimana', 'sehat', 'sejati', 'sekalulian', 'sekarangdan', 'sekolah', 'sekrang', 'selamat', 'selamatanda', 'selamatkepada', 'selamatmenyambutkemenangan', 'selamatno', 'selang', 'selasa', 'seleb', 'selempang', 'selesai', 'selesei', 'selisih', 'selling', 'semangat', 'semanggi', 'semarak', 'semarang', 'semaumaunya', 'sembuh', 'semester', 'seminar', 'semnasdikti', 'sempat', 'semuanyaaa', 'semuanyasemoga', 'senang', 'sendiiri', 'sendiriayo', 'senin', 'seninkamis', 'sensasi', 'sensor', 'senyum', 'sep', 'sepenuh', 'september', 'septiana', 'sepupu', 'serah', 'serang', 'serba', 'seri', 'serianto', 'serta', 'sertifikat', 'seru', 'seruseruan', 'services', 'servis', 'sesal', 'sesame', 'sesi', 'sesuai', 'sesuain', 'setia', 'setiabudhi', 'setiabudi', 'setor', 'setting', 'sf', 'sgp', 'sgr', 'shara', 'share', 'shd', 'sheet', 'shio', 'sholat', 'shop', 'shopinfo', 'si', 'siang', 'siap', 'siar', 'siaraan', 'siaw', 'sibuk', 'sidang', 'sidangdemikian', 'sidomuncul', 'sigindo', 'sigit', 'sih', 'sihaloho', 'sila', 'silah', 'silaturahmi', 'sim', 'simak', 'simcard', 'simpan', 'simpati', 'simpul', 'sin', 'sinar', 'sindo', 'singapur', 'singgung', 'singkat', 'single', 'sinonim', 'sioangka', 'sirup', 'sisa', 'sisi', 'sistem', 'situ', 'situs', 'sjdsini', 'sjk', 'sk', 'skb', 'skema', 'skip', 'skl', 'skor', 'skripsi', 'skripsigemastikkerjaan', 'slainya', 'sli', 'slide', 'slip', 'slm', 'slma', 'slmt', 'slmtno', 'slmtnosimcard', 'slot', 'sma', 'small', 'smangat', 'smartfren', 'smartfrenmu', 'smartphone', 'smk', 'smlt', 'sms', 'smsgift', 'smsin', 'smssblum', 'smstlp', 'smuanya', 'sn', 'snack', 'snbyr', 'snote', 'soalna', 'sob', 'soebroto', 'soedirmanpurwokerto', 'soekarno', 'sofiyah', 'softball', 'softfilenya', 'software', 'sok', 'solder', 'solo', 'solusi', 'something', 'soraya', 'sore', 'sospol', 'source', 'sp', 'spagetti', 'spam', 'spasi', 'special', 'spesial', 'sponsor', 'sportsbook', 'spreading', 'sql', 'sr', 'sri', 'ss', 'ssd', 'standart', 'stasiun', 'stasiunbandara', 'status', 'statusberhenti', 'stay', 'stickernya', 'stickershop', 'stiker', 'stop', 'store', 'streaming', 'stres', 'strip', 'struktur', 'studi', 'studio', 'suami', 'suamiisteri', 'suara', 'subang', 'subhanahu', 'subsidi', 'subuh', 'sudirman', 'sugiyono', 'suka', 'sukarnohatta', 'sukses', 'sulit', 'sumargani', 'sumber', 'summary', 'sumo', 'sundae', 'sunday', 'sungai', 'sunnytrx', 'super', 'surabaya', 'surat', 'surpai', 'suruh', 'survei', 'surveu', 'survey', 'suryani', 'susah', 'suyono', 'suzuki', 'swamedia', 'syamsudin', 'syarat', 'syariah', 'sync', 'syrt', 'sys', 'syukur', 'tabel', 'table', 'tablenya', 'tabugan', 'tabung', 'tagih', 'tahap', 'take', 'takut', 'tama', 'tambah', 'tampil', 'tanah', 'tanda', 'tangan', 'tanggal', 'tanggung', 'tanjung', 'tanp', 'tanyain', 'tap', 'tapdptkan', 'taplus', 'taqoballahu', 'tarif', 'tariff', 'tariflokasi', 'tarik', 'taro', 'tau', 'taut', 'tawar', 'taxi', 'tbk', 'tbs', 'tbsstlh', 'tbungn', 'tcash', 'tcashmu', 'tcpdf', 'tdftar', 'tdftr', 'tdn', 'tdr', 'tea', 'tear', 'tebak', 'tebal', 'teduh', 'tegal', 'teh', 'tehkalau', 'tekan', 'telat', 'telepon', 'televisi', 'teliti', 'telkom', 'telkomsel', 'telkomselpoin', 'telor', 'telpon', 'telponsms', 'tema', 'teman', 'temantemanmu', 'tembak', 'tembus', 'temen', 'temenin', 'tempat', 'tempe', 'template', 'tempo', 'temposelamat', 'tempuh', 'temu', 'tenggara', 'teori', 'terang', 'terbang', 'terbataspromo', 'terbatassd', 'terbatasskb', 'terdekathubungi', 'terdekattelf', 'terdftr', 'terima', 'terimakasih', 'terkadang', 'terkahir', 'terminal', 'tersmbung', 'tersrrah', 'terusin', 'tes', 'test', 'teteh', 'text', 'tf', 'tgguin', 'tglmaret', 'th', 'thailand', 'thank', 'the', 'themes', 'thing', 'this', 'thr', 'tia', 'tibatiba', 'ticket', 'tidak', 'tiga', 'tika', 'tiket', 'tilang', 'timezone', 'timur', 'tinggal', 'tinggi', 'tingkat', 'tipemerk', 'tipon', 'tipu', 'titaniadbl', 'titip', 'titipin', 'tkn', 'tlb', 'tlc', 'tld', 'tlpsms', 'toch', 'toefl', 'togel', 'token', 'toko', 'tokopedi', 'tolong', 'tombol', 'tong', 'tonton', 'top', 'topik', 'topup', 'torch', 'torchjt', 'torchrp', 'total', 'tourch', 'tourtravel', 'toyota', 'toyotanyasy', 'tpilih', 'tranfer', 'trans', 'transaksi', 'transfer', 'transformasi', 'transport', 'travel', 'travelokacom', 'treatment', 'tren', 'trend', 'trfr', 'tri', 'tribunjt', 'tricare', 'tricarepin', 'tricoid', 'tridaya', 'trima', 'trimksh', 'trimksi', 'trims', 'tripin', 'triple', 'triplle', 'trmh', 'trmks', 'trmksh', 'trms', 'trnsfer', 'tronik', 'trpilih', 'trsambung', 'trsambungjd', 'trsmbung', 'truck', 'trx', 'tselmeas', 'tselmefl', 'tselmeinternet', 'tselmemytsel', 'tselmetappromo', 'tselmetelkomselpoin', 'tselpoin', 'tsm', 'tst', 'tsu', 'ttg', 'tua', 'tuanai', 'tudiknas', 'tugas', 'tugasquiz', 'tugu', 'tuh', 'tuju', 'tuk', 'tukar', 'tulang', 'tulis', 'tulisanya', 'tumben', 'tunai', 'tunggal', 'tunggu', 'tunjuk', 'turbo', 'turun', 'tutup', 'twallet', 'twd', 'twitter', 'type', 'uang', 'uangnyapenting', 'uas', 'uasnya', 'ubah', 'ucek', 'udunan', 'ugabung', 'uimfo', 'uinfo', 'uinfohubpakh', 'ujannya', 'uji', 'ujiklyakankerjaberstifikat', 'ujual', 'ujug', 'ulanag', 'ulang', 'ulg', 'ulsa', 'ultima', 'umelihat', 'umembeli', 'umum', 'umumin', 'umur', 'unbearable', 'undang', 'undi', 'undianfbgs', 'undianmlengkapi', 'undur', 'ungkapin', 'unisba', 'unit', 'unlimited', 'unpad', 'unreg', 'unruk', 'unt', 'untk', 'untuk', 'untung', 'up', 'update', 'updatenya', 'upemesanan', 'upemesananorder', 'upi', 'upinet', 'upnyaa', 'urang', 'urangbandung', 'urg', 'urgent', 'urus', 'us', 'usa', 'usaha', 'usahain', 'use', 'using', 'ustad', 'usul', 'utama', 'utkinfo', 'uts', 'value', 'values', 'vas', 'vchr', 'vd', 'verifikasi', 'versi', 'very', 'via', 'video', 'view', 'vip', 'visa', 'visimisi', 'visit', 'voice', 'volume', 'vote', 'voucher', 'vrb', 'wa', 'waaah', 'waalaikumsalam', 'waalaikumsalamin', 'wacana', 'wahyudin', 'wahyumengenai', 'wajah', 'wajib', 'wakil', 'waktupenguji', 'walfaidin', 'wallppr', 'walopun', 'wamhn', 'waminkum', 'wanita', 'war', 'warna', 'warung', 'wassalamu', 'wastukencana', 'wata', 'wb', 'we', 'websete', 'websitenya', 'week', 'weekend', 'weekendharga', 'weka', 'wendys', 'whatsapp', 'wib', 'wibbagenda', 'wid', 'wida', 'wife', 'wifi', 'wihh', 'wilayah', 'windows', 'wisuda', 'wk', 'wkkw', 'wktnya', 'wkwk', 'wkwkw', 'wkwkwk', 'wkwkwkwk', 'wonderland', 'wooh', 'wow', 'wqwq', 'wr', 'wrwb', 'wwfindonesia', 'xjam', 'xl', 'xmgg', 'xnpdgbdodzzzz', 'xpress', 'xv', 'xx', 'xxi', 'xxx', 'xxxx', 'xxxxx', 'xxxxxx', 'xxxxxxx', 'xxxxxxxxxx', 'xyqd', 'ya', 'yaa', 'yaaa', 'yaaaa', 'yaaaaa', 'yaaah', 'yacc', 'yach', 'yah', 'yahama', 'yahh', 'yakalau', 'yamalem', 'yani', 'yank', 'yanti', 'yanuhun', 'yaok', 'yaris', 'yarispin', 'yawww', 'yellow', 'yesok', 'yk', 'yogya', 'yooo', 'your', 'ypac', 'yrkz', 'ytg', 'ythmiitra', 'ythmkios', 'ythmohon', 'ythno', 'ythrekning', 'ythsaya', 'ythtri', 'yu', 'yudisium', 'yuk', 'yuks', 'yuni', 'yunit', 'zalora', 'zarkasi', 'zjt', 'zona', 'ztkm']

In [26]:

*# Melihat matriks jumlah token*  
*# Data ini siap untuk dimasukkan dalam proses pemodelan (machine learning)*  
  
X\_tf\_idf **=** tf\_idf**.**transform(X)**.**toarray()  
X\_tf\_idf

Out[26]:

array([[0., 0., 0., ..., 0., 0., 0.],  
 [0., 0., 0., ..., 0., 0., 0.],  
 [0., 0., 0., ..., 0., 0., 0.],  
 ...,  
 [0., 0., 0., ..., 0., 0., 0.],  
 [0., 0., 0., ..., 0., 0., 0.],  
 [0., 0., 0., ..., 0., 0., 0.]])

In [27]:

*# Melihat matriks jumlah token menggunakan TF IDF, lihat perbedaannya dengan metode BoW*  
*# Data ini siap untuk dimasukkan dalam proses pemodelan (machine learning)*  
  
data\_tf\_idf **=** pd**.**DataFrame(X\_tf\_idf, columns**=**tf\_idf**.**get\_feature\_names())  
data\_tf\_idf

Out[27]:

|  | **aa** | **aamiiiin** | **aamiin** | **ab** | **abadi** | **abai** | **abbee** | **abdul** | **acara** | **acaratks** | **...** | **yudisium** | **yuk** | **yuks** | **yuni** | **yunit** | **zalora** | **zarkasi** | **zjt** | **zona** | **ztkm** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **2** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **4** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1138** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1139** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1140** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1141** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1142** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

1143 rows × 3253 columns

In [28]:

**with** open('tf\_idf\_feature.pickle', 'wb') **as** output:  
 pickle**.**dump(X\_tf\_idf, output)

Feature Selection[¶](#26in1rg)

In [29]:

*# Mengubah nilai data tabular tf-idf menjadi array agar dapat dijalankan pada proses seleksi fitur*  
X **=** np**.**array(data\_tf\_idf)  
y **=** np**.**array(y)

In [30]:

*# Ten features with highest chi-squared statistics are selected*   
chi2\_features **=** SelectKBest(chi2, k**=**1000)   
X\_kbest\_features **=** chi2\_features**.**fit\_transform(X, y)   
   
*# Reduced features*   
print('Original feature number:', X**.**shape[1])   
print('Reduced feature number:', X\_kbest\_features**.**shape[1])

Original feature number: 3253  
Reduced feature number: 1000

In [31]:

*# chi2\_features.scores\_ adalah nilai chi-square, semakin tinggi nilainya maka semakin baik fiturnya*  
data\_chi2 **=** pd**.**DataFrame(chi2\_features**.**scores\_, columns**=**['nilai'])  
data\_chi2

Out[31]:

|  | **nilai** |
| --- | --- |
| **0** | 0.835608 |
| **1** | 0.419698 |
| **2** | 1.558607 |
| **3** | 0.716455 |
| **4** | 0.800674 |
| **...** | ... |
| **3248** | 1.180239 |
| **3249** | 0.503162 |
| **3250** | 0.716455 |
| **3251** | 2.917381 |
| **3252** | 0.892022 |

3253 rows × 1 columns

In [32]:

*# Menampilkan fitur beserta nilainya*  
feature **=** tf\_idf**.**get\_feature\_names()  
data\_chi2['fitur'] **=** feature  
data\_chi2

Out[32]:

|  | **nilai** | **fitur** |
| --- | --- | --- |
| **0** | 0.835608 | aa |
| **1** | 0.419698 | aamiiiin |
| **2** | 1.558607 | aamiin |
| **3** | 0.716455 | ab |
| **4** | 0.800674 | abadi |
| **...** | ... | ... |
| **3248** | 1.180239 | zalora |
| **3249** | 0.503162 | zarkasi |
| **3250** | 0.716455 | zjt |
| **3251** | 2.917381 | zona |
| **3252** | 0.892022 | ztkm |

3253 rows × 2 columns

In [33]:

*# Mengurutkan fitur terbaik*  
data\_chi2**.**sort\_values(by**=**'nilai', ascending**=False**)

Out[33]:

|  | **nilai** | **fitur** |
| --- | --- | --- |
| **2088** | 49.062970 | paket |
| **1030** | 46.324101 | hadiah |
| **1549** | 45.593560 | kuota |
| **2178** | 39.438431 | pin |
| **1486** | 34.448410 | klik |
| **...** | ... | ... |
| **1520** | 0.044910 | kopi |
| **307** | 0.044468 | bca |
| **1694** | 0.031579 | maksimal |
| **531** | 0.013783 | cepat |
| **3121** | 0.008879 | via |

3253 rows × 2 columns

In [34]:

*# Menampilkan mask pada feature yang diseleksi*  
*# False berarti fitur tidak terpilih dan True berarti fitur terpilih*  
mask **=** chi2\_features**.**get\_support()  
mask

Out[34]:

array([False, False, True, ..., False, True, False])

In [35]:

*# Menampilkan fitur-fitur terpilih berdasarkan mask atau nilai tertinggi yang sudah dikalkulasi pada Chi-Square*  
new\_feature **=** []  
  
**for** bool, f **in** zip(mask, feature):  
 **if** bool:  
 new\_feature**.**append(f)  
 selected\_feature **=** new\_feature  
  
selected\_feature

Out[35]:

['aamiin',  
 'agam',  
 'agen',  
 'ags',  
 'agst',  
 'agsts',  
 'agt',  
 'agtskinfodlj',  
 'agun',  
 'agustuskunjungi',  
 'ajak',  
 'akses',  
 'aktif',  
 'aktivasi',  
 'alam',  
 'alamsyah',  
 'aliando',  
 'all',  
 'aman',  
 'amanda',  
 'an',  
 'anda',  
 'andabch',  
 'andika',  
 'andromax',  
 'angga',  
 'angpaopoinsenyum',  
 'angsur',  
 'anrasto',  
 'ap',  
 'apa',  
 'aplikasi',  
 'app',  
 'appleplay',  
 'apps',  
 'appsinapps',  
 'argo',  
 'armada',  
 'artha',  
 'artisartis',  
 'arya',  
 'as',  
 'asa',  
 'asli',  
 'ass',  
 'assalamualaikum',  
 'at',  
 'atas',  
 'ato',  
 'aug',  
 'augsept',  
 'avanza',  
 'axiata',  
 'axis',  
 'axisnet',  
 'ayah',  
 'ayam',  
 'ayo',  
 'bagi',  
 'baik',  
 'bakmi',  
 'balas',  
 'ball',  
 'bandung',  
 'bandungmulai',  
 'bang',  
 'banget',  
 'bank',  
 'banting',  
 'banyak',  
 'bapa',  
 'bareng',  
 'barokahmp',  
 'baru',  
 'baruqskrng',  
 'baskin',  
 'batas',  
 'batin',  
 'bawa',  
 'bayar',  
 'bb',  
 'bbm',  
 'bch',  
 'bebek',  
 'beda',  
 'belanja',  
 'beli',  
 'beliau',  
 'belikan',  
 'bentar',  
 'berangkat',  
 'beres',  
 'berkat',  
 'berlakupromo',  
 'berrybenka',  
 'besok',  
 'bfgs',  
 'bh',  
 'biar',  
 'biaya',  
 'biber',  
 'bicara',  
 'big',  
 'bikin',  
 'bilang',  
 'bimatri',  
 'bimbing',  
 'bingung',  
 'bioskop',  
 'bis',  
 'bitlytentangcigna',  
 'bl',  
 'blackberry',  
 'blanja',  
 'blanjacom',  
 'bljcotcashrabu',  
 'blog',  
 'bni',  
 'bolak',  
 'bonus',  
 'booking',  
 'borma',  
 'bpa',  
 'bpjs',  
 'bpkb',  
 'bpkibu',  
 'bpkyogi',  
 'brgkt',  
 'bri',  
 'brianasri',  
 'brinorekanir',  
 'britama',  
 'bro',  
 'bronet',  
 'bronetjam',  
 'bronetspecial',  
 'btpn',  
 'bu',  
 'buktikeun',  
 'bulan',  
 'bunga',  
 'buru',  
 'buy',  
 'callsms',  
 'camera',  
 'card',  
 'care',  
 'cashback',  
 'cek',  
 'cell',  
 'cfc',  
 'charged',  
 'chat',  
 'chatting',  
 'chicken',  
 'chn',  
 'cigna',  
 'cip',  
 'cipika',  
 'cliponyu',  
 'cobagrabcar',  
 'cocok',  
 'coins',  
 'cokelat',  
 'combo',  
 'controller',  
 'copy',  
 'cream',  
 'cs',  
 'cuci',  
 'curi',  
 'da',  
 'dacota',  
 'daging',  
 'dakota',  
 'dan',  
 'dana',  
 'dapat',  
 'data',  
 'database',  
 'datang',  
 'davis',  
 'deh',  
 'dekat',  
 'dengerin',  
 'dermawan',  
 'desember',  
 'detail',  
 'dhi',  
 'diagnosa',  
 'diil',  
 'dikti',  
 'dil',  
 'diri',  
 'disc',  
 'discound',  
 'diskon',  
 'dki',  
 'dlk',  
 'doa',  
 'domestic',  
 'domestik',  
 'dominos',  
 'donasi',  
 'donut',  
 'donuts',  
 'double',  
 'doublescoop',  
 'download',  
 'doyan',  
 'dptkan',  
 'dptkn',  
 'drhirawan',  
 'duit',  
 'dukung',  
 'dummy',  
 'dunkin',  
 'dwm',  
 'edisi',  
 'eh',  
 'eks',  
 'ekspresi',  
 'ekstra',  
 'email',  
 'emang',  
 'enak',  
 'english',  
 'er',  
 'ertiga',  
 'etiket',  
 'euy',  
 'ewu',  
 'extra',  
 'fashion',  
 'fatih',  
 'favorit',  
 'fc',  
 'feb',  
 'felix',  
 'fg',  
 'fhjtl',  
 'file',  
 'finance',  
 'flash',  
 'flashspasiya',  
 'flat',  
 'flavour',  
 'form',  
 'foto',  
 'fotomms',  
 'free',  
 'freedom',  
 'freedoom',  
 'gadget',  
 'gagal',  
 'gais',  
 'gajah',  
 'galeri',  
 'gallery',  
 'game',  
 'gamegame',  
 'games',  
 'gamesmu',  
 'gampang',  
 'gara',  
 'gatis',  
 'gbberlaku',  
 'gbg',  
 'gbgb',  
 'gbgbonus',  
 'gbggb',  
 'gbhr',  
 'gbkonten',  
 'gbnelpon',  
 'ge',  
 'gebyar',  
 'gemini',  
 'gerai',  
 'get',  
 'gift',  
 'gik',  
 'gila',  
 'gita',  
 'github',  
 'glte',  
 'gm',  
 'gmau',  
 'google',  
 'grab',  
 'grabcar',  
 'grabtaxi',  
 'grabtravel',  
 'grapari',  
 'gratis',  
 'grup',  
 'gudang',  
 'guru',  
 'guys',  
 'haagendasz',  
 'habis',  
 'hadiah',  
 'haha',  
 'hak',  
 'hallo',  
 'handphone',  
 'harap',  
 'harga',  
 'hari',  
 'harirprb',  
 'hatur',  
 'heboh',  
 'hehe',  
 'hemat',  
 'henti',  
 'heroes',  
 'heula',  
 'hg',  
 'hggmbpstekan',  
 'hihi',  
 'hit',  
 'hits',  
 'hobi',  
 'honda',  
 'hormat',  
 'hot',  
 'hp',  
 'hubagen',  
 'hubung',  
 'hujan',  
 'hut',  
 'ice',  
 'identitas',  
 'ieu',  
 'ih',  
 'ii',  
 'ijhk',  
 'iklan',  
 'im',  
 'indofood',  
 'indonesia',  
 'indosat',  
 'indosatooredoo',  
 'indosatooredoocomfreedom',  
 'infokunjungi',  
 'infold',  
 'informasi',  
 'infosms',  
 'inih',  
 'inipromo',  
 'instagram',  
 'instink',  
 'internasional',  
 'internet',  
 'internetan',  
 'internetas',  
 'internetmu',  
 'intl',  
 'intrviewrekom',  
 'iphone',  
 'ir',  
 'iring',  
 'isi',  
 'isiin',  
 'isiulang',  
 'islam',  
 'ismawati',  
 'iur',  
 'iya',  
 'jajan',  
 'jakarta',  
 'jam',  
 'jamgb',  
 'jamin',  
 'jan',  
 'janji',  
 'jaring',  
 'jatuh',  
 'jawa',  
 'jaya',  
 'jazz',  
 'jf',  
 'jitu',  
 'jlpatuha',  
 'jobfairactive',  
 'jr',  
 'jt',  
 'jual',  
 'juke',  
 'jul',  
 'juli',  
 'jumat',  
 'juni',  
 'juta',  
 'kadang',  
 'kak',  
 'kalah',  
 'kali',  
 'kamar',  
 'kamera',  
 'kamis',  
 'kampus',  
 'kang',  
 'kantor',  
 'kar',  
 'karimbeliau',  
 'kartu',  
 'kartuhalo',  
 'kav',  
 'kaya',  
 'kb',  
 'kbdapatkan',  
 'kebut',  
 'keju',  
 'kejut',  
 'kelik',  
 'keliling',  
 'kenal',  
 'kenyang',  
 'kesana',  
 'ketemu',  
 'ketik',  
 'khusus',  
 'kini',  
 'kirim',  
 'kl',  
 'klik',  
 'ko',  
 'kode',  
 'kodevrd',  
 'koin',  
 'komunikasi',  
 'kontra',  
 'kontrak',  
 'kosan',  
 'kosong',  
 'kredit',  
 'krim',  
 'ktp',  
 'kuasa',  
 'kuat',  
 'kuliah',  
 'kumpul',  
 'kunjung',  
 'kuota',  
 'kuotabeli',  
 'kuotaspasik',  
 'kupon',  
 'lab',  
 'lagistok',  
 'lain',  
 'laku',  
 'lancar',  
 'langgan',  
 'langka',  
 'lanjut',  
 'lanjutin',  
 'lapor',  
 'latte',  
 'layan',  
 'lazada',  
 'ld',  
 'legal',  
 'lenkap',  
 'lets',  
 'lgs',  
 'lho',  
 'limpah',  
 'line',  
 'listrik',  
 'listrikspeedytelkom',  
 'loh',  
 'lokasi',  
 'loopers',  
 'lottemart',  
 'lte',  
 'luculucu',  
 'lulus',  
 'lupa',  
 'lychee',  
 'maaf',  
 'mad',  
 'maghrib',  
 'mah',  
 'mahar',  
 'main',  
 'makan',  
 'maksud',  
 'malas',  
 'mama',  
 'mana',  
 'mandi',  
 'mandiri',  
 'mang',  
 'mar',  
 'mas',  
 'masak',  
 'masuk',  
 'mati',  
 'mayani',  
 'mbah',  
 'mbak',  
 'mbhr',  
 'mbmbbonus',  
 'mcdonalds',  
 'mdpt',  
 'mdptkan',  
 'medan',  
 'mega',  
 'meganai',  
 'mei',  
 'meijuni',  
 'mel',  
 'melapormhub',  
 'menang',  
 'mendpt',  
 'menit',  
 'menittlp',  
 'menitvideo',  
 'menu',  
 'merchants',  
 'merdeka',  
 'metro',  
 'mh',  
 'midnight',  
 'mie',  
 'mifi',  
 'mifimodem',  
 'migrasi',  
 'milik',  
 'min',  
 'minat',  
 'minggu',  
 'minggudaftar',  
 'mini',  
 'minimal',  
 'minimun',  
 'minum',  
 'miring',  
 'mkios',  
 'mlahkgsngtwnvtmmacau',  
 'mlmktkgratis',  
 'mnctv',  
 'mndpt',  
 'mndptkn',  
 'mnthr',  
 'mobil',  
 'model',  
 'moga',  
 'mohon',  
 'mtronik',  
 'mu',  
 'mudah',  
 'mumpung',  
 'murah',  
 'murdoyo',  
 'murtiani',  
 'musik',  
 'my',  
 'mycare',  
 'mytelkomsel',  
 'myxlcoid',  
 'nama',  
 'narti',  
 'nasabah',  
 'nasib',  
 'negara',  
 'nego',  
 'negosiasi',  
 'nelpsms',  
 'netmax',  
 'new',  
 'nge',  
 'ngomong',  
 'ngumpul',  
 'nikah',  
 'nikmat',  
 'nilai',  
 'nissan',  
 'nitip',  
 'noah',  
 'noanda',  
 'nokia',  
 'nomor',  
 'nonaktif',  
 'nonton',  
 'nopin',  
 'norek',  
 'norekan',  
 'normal',  
 'nosimcard',  
 'november',  
 'noxv',  
 'noxxx',  
 'nrml',  
 'nsp',  
 'numpang',  
 'nunggu',  
 'nuy',  
 'nya',  
 'nyalira',  
 'nyari',  
 'nyusul',  
 'obor',  
 'obral',  
 'oh',  
 'okcall',  
 'oke',  
 'onix',  
 'online',  
 'onyx',  
 'ooredohanya',  
 'ooredoo',  
 'operator',  
 'orang',  
 'otomatis',  
 'pa',  
 'pacar',  
 'pagi',  
 'pakai',  
 'paket',  
 'paketextra',  
 'pakettelkomselflash',  
 'pan',  
 'panjang',  
 'panpizzacoid',  
 'partner',  
 'pas',  
 'pasca',  
 'pasien',  
 'pasti',  
 'per',  
 'perbln',  
 'pergi',  
 'periksa',  
 'periode',  
 'perkedel',  
 'persero',  
 'pesan',  
 'pesawat',  
 'pesta',  
 'ph',  
 'pilih',  
 'pin',  
 'pinbfgs',  
 'pindhi',  
 'pinggir',  
 'pinggirmulai',  
 'pinjam',  
 'pinjf',  
 'pinmcxj',  
 'pinmhr',  
 'pinmk',  
 'pinnf',  
 'pizza',  
 'pizzaroti',  
 'pizzarotinya',  
 'pkl',  
 'pla',  
 'plasatelkommlg',  
 'platihan',  
 'play',  
 'playapple',  
 'plaza',  
 'plgn',  
 'pln',  
 'plus',  
 'plusplus',  
 'pmnang',  
 'pnwrn',  
 'poin',  
 'point',  
 'polisi',  
 'pool',  
 'pop',  
 'popmie',  
 'praktis',  
 'prasetya',  
 'prastia',  
 'premium',  
 'prodi',  
 'program',  
 'promo',  
 'proses',  
 'pt',  
 'ptcare',  
 'ptindofood',  
 'ptindosat',  
 'ptmkios',  
 'ptmtronik',  
 'pttelkomsel',  
 'pttri',  
 'pttricare',  
 'puas',  
 'pucuk',  
 'pulsa',  
 'pulsabayar',  
 'pulsacek',  
 'pulsarb',  
 'pusat',  
 'push',  
 'put',  
 'putus',  
 'quota',  
 'rabu',  
 'rahasia',  
 'raih',  
 'rakyat',  
 'ramadhan',  
 'range',  
 'rapih',  
 'rawit',  
 'rbbulan',  
 'rbminggu',  
 'rbnrml',  
 'rbu',  
 'rctiinfo',  
 'refgratis',  
 'regular',  
 'reguler',  
 'rejeki',  
 'rek',  
 'rekening',  
 'reloadpasti',  
 'remedial',  
 'rencang',  
 'repot',  
 'resmi',  
 'restart',  
 'resti',  
 'reza',  
 'ri',  
 'ribet',  
 'ribu',  
 'rica',  
 'rich',  
 'robbins',  
 'rpjt',  
 'rpjtdr',  
 'rpjuta',  
 'rpkunjungi',  
 'rpmggcs',  
 'rprb',  
 'rprbbln',  
 'rpribu',  
 'rs',  
 'rsuinfo',  
 'rtx',  
 'ruang',  
 'rumah',  
 'rupiah',  
 'sabtu',  
 'sadikin',  
 'sahabat',  
 'salah',  
 'sale',  
 'sambung',  
 'samsung',  
 'sayappaha',  
 'scoop',  
 'sctv',  
 'sd',  
 'sdnov',  
 'sebenernya',  
 'sebentar',  
 'sedia',  
 'sekrang',  
 'selamat',  
 'selamatanda',  
 'selasa',  
 'seleb',  
 'selesai',  
 'semanggi',  
 'sempat',  
 'seninkamis',  
 'sensasi',  
 'senyum',  
 'sep',  
 'septiana',  
 'seru',  
 'sesi',  
 'sesuai',  
 'setia',  
 'setiabudhi',  
 'sf',  
 'sgp',  
 'shio',  
 'si',  
 'siang',  
 'siaw',  
 'sibuk',  
 'sidang',  
 'sidomuncul',  
 'sih',  
 'silah',  
 'simcard',  
 'simpati',  
 'sin',  
 'sisa',  
 'situs',  
 'sjk',  
 'sk',  
 'skb',  
 'skripsi',  
 'sli',  
 'slm',  
 'slmtno',  
 'sma',  
 'smartfren',  
 'smartfrenmu',  
 'smartphone',  
 'sms',  
 'smsgift',  
 'smuanya',  
 'sob',  
 'solusi',  
 'sore',  
 'spasi',  
 'special',  
 'spesial',  
 'stasiunbandara',  
 'stickernya',  
 'stickershop',  
 'stiker',  
 'stop',  
 'store',  
 'streaming',  
 'suami',  
 'sudirman',  
 'suka',  
 'sukses',  
 'sumo',  
 'super',  
 'surat',  
 'survei',  
 'surveu',  
 'survey',  
 'suzuki',  
 'syarat',  
 'sync',  
 'table',  
 'tambah',  
 'tanah',  
 'tanp',  
 'tap',  
 'tapdptkan',  
 'tarif',  
 'tariff',  
 'tariflokasi',  
 'tau',  
 'tawar',  
 'taxi',  
 'tbk',  
 'tcash',  
 'tcashmu',  
 'tebal',  
 'teh',  
 'tekan',  
 'telat',  
 'telepon',  
 'telkomsel',  
 'telkomselpoin',  
 'telpon',  
 'temantemanmu',  
 'tembus',  
 'temen',  
 'tempo',  
 'terbang',  
 'terbataspromo',  
 'terbatassd',  
 'terbatasskb',  
 'terdekathubungi',  
 'terdekattelf',  
 'terima',  
 'terimakasih',  
 'teteh',  
 'th',  
 'thailand',  
 'the',  
 'themes',  
 'thr',  
 'tidak',  
 'tiga',  
 'timezone',  
 'timur',  
 'tinggal',  
 'tipon',  
 'tipu',  
 'tkn',  
 'tlb',  
 'tld',  
 'togel',  
 'toko',  
 'tolong',  
 'top',  
 'torch',  
 'tourch',  
 'toyota',  
 'tpilih',  
 'trans',  
 'transaksi',  
 'transfer',  
 'tri',  
 'tricare',  
 'tricoid',  
 'trima',  
 'trims',  
 'triple',  
 'trmksh',  
 'trnsfer',  
 'trpilih',  
 'truck',  
 'tselmeas',  
 'tselmefl',  
 'tselmeinternet',  
 'tselmemytsel',  
 'tselmetappromo',  
 'tst',  
 'tsu',  
 'tugas',  
 'tuh',  
 'tuju',  
 'tukar',  
 'tunai',  
 'tunggu',  
 'tunjuk',  
 'turbo',  
 'turun',  
 'twallet',  
 'twd',  
 'twitter',  
 'uang',  
 'ubah',  
 'uinfo',  
 'uji',  
 'ujiklyakankerjaberstifikat',  
 'ulang',  
 'ultima',  
 'umur',  
 'undang',  
 'undi',  
 'ungkapin',  
 'unit',  
 'unlimited',  
 'unreg',  
 'untuk',  
 'untung',  
 'update',  
 'upi',  
 'usa',  
 'ustad',  
 'value',  
 'vas',  
 'vchr',  
 'vd',  
 'verifikasi',  
 'video',  
 'view',  
 'visa',  
 'visit',  
 'voice',  
 'volume',  
 'voucher',  
 'vrb',  
 'wallppr',  
 'weekend',  
 'wendys',  
 'whatsapp',  
 'wid',  
 'wisuda',  
 'wkwk',  
 'wkwkwk',  
 'wonderland',  
 'wow',  
 'wwfindonesia',  
 'xl',  
 'xv',  
 'xxx',  
 'xxxx',  
 'xxxxx',  
 'xxxxxx',  
 'xxxxxxx',  
 'ya',  
 'yaa',  
 'yaaa',  
 'yacc',  
 'yah',  
 'yank',  
 'yaok',  
 'yaris',  
 'yellow',  
 'yesok',  
 'your',  
 'yrkz',  
 'ytg',  
 'ythmohon',  
 'ythsaya',  
 'yuk',  
 'yunit',  
 'zalora',  
 'zona']

In [36]:

*# Memilih fitur yang telah terseleksi dari vocabulary TF IDF*  
kbest\_feature **=** {}  
  
**for** (k,v) **in** tf\_idf**.**vocabulary\_**.**items():  
 **if** k **in** selected\_feature:  
 kbest\_feature[k] **=** v

In [37]:

len(kbest\_feature)

Out[37]:

1000

In [38]:

*# Menampilkan fitur-fitur yang sudah diseleksi*   
*# Beserta nilai vektornya pada keseluruhan data untuk dijalankan pada proses machine learning*  
  
*# Hanya k fitur yang terpilih sesuai parameter k yang ditentukan sebelumnya*  
  
data\_selected\_feature **=** pd**.**DataFrame(X\_kbest\_features, columns**=**selected\_feature)  
data\_selected\_feature

Out[38]:

|  | **aamiin** | **agam** | **agen** | **ags** | **agst** | **agsts** | **agt** | **agtskinfodlj** | **agun** | **agustuskunjungi** | **...** | **yesok** | **your** | **yrkz** | **ytg** | **ythmohon** | **ythsaya** | **yuk** | **yunit** | **zalora** | **zona** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **2** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **4** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1138** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1139** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1140** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1141** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1142** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

1143 rows × 1000 columns

In [39]:

**with** open('kbest\_feature.pickle', 'wb') **as** output:  
 pickle**.**dump(kbest\_feature, output)

Modelling (Machine Learning)[¶](#lnxbz9)

In [40]:

*# Memisihkan data training dan data testing dengan perbandingan 80:20*  
X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X\_kbest\_features, y, test\_size**=**0.2, random\_state**=**0)

In [41]:

*# Training Model*  
algorithm **=** MultinomialNB()  
model **=** algorithm**.**fit(X\_train, y\_train)

Prediction[¶](#35nkun2)

In [42]:

*# Prediksi untuk Data Latih*  
predicted\_train **=** model**.**predict(X\_train)  
  
*# hasil ini akan kita bandingkan dengan nilai y\_train (labeling dari manusia)*  
predicted\_train[1:20]

Out[42]:

array([2, 2, 0, 2, 0, 2, 1, 2, 0, 1, 0, 0, 0, 2, 0, 1, 0, 0, 1],  
 dtype=int64)

In [43]:

*# Prediksi untuk Data Uji*  
predicted\_test **=** model**.**predict(X\_test)  
  
*# hasil ini akan kita bandingkan dengan nilai y\_test (labeling dari manusia)*  
predicted\_test[1:20]

Out[43]:

array([2, 0, 0, 2, 0, 2, 0, 0, 0, 1, 0, 2, 1, 0, 0, 0, 1, 0, 0],  
 dtype=int64)

Model Evaluation[¶](#1ksv4uv)

In [44]:

*# Perhitungan Performa Model untuk Data Latih*  
CM **=** confusion\_matrix(y\_train, predicted\_train)  
  
print(CM,'\n')  
  
print("ACCURACY\t:", accuracy\_score(y\_train, predicted\_train)**\***100,"%\n")  
  
print(classification\_report(y\_train,predicted\_train))

[[433 3 7]  
 [ 12 249 8]  
 [ 7 5 190]]   
  
ACCURACY : 95.40481400437638 %  
  
 precision recall f1-score support  
  
 0 0.96 0.98 0.97 443  
 1 0.97 0.93 0.95 269  
 2 0.93 0.94 0.93 202  
  
 accuracy 0.95 914  
 macro avg 0.95 0.95 0.95 914  
weighted avg 0.95 0.95 0.95 914

In [45]:

*# Perhitungan Performa Model untuk Data Uji*  
CM **=** confusion\_matrix(y\_test, predicted\_test)  
  
print(CM,'\n')  
  
print("ACCURACY\t:", accuracy\_score(y\_test, predicted\_test)**\***100,"%\n")  
  
print(classification\_report(y\_test,predicted\_test))

[[120 4 2]  
 [ 7 58 1]  
 [ 3 3 31]]   
  
ACCURACY : 91.26637554585153 %  
  
 precision recall f1-score support  
  
 0 0.92 0.95 0.94 126  
 1 0.89 0.88 0.89 66  
 2 0.91 0.84 0.87 37  
  
 accuracy 0.91 229  
 macro avg 0.91 0.89 0.90 229  
weighted avg 0.91 0.91 0.91 229

Menyimpan Model[¶](#44sinio)

In [46]:

*# save the model to disk*  
dump(model, filename**=**"model\_spam\_tfidf\_nb.model")

Out[46]:

['model\_spam\_tfidf\_nb.model']