Data Analytics Term Project

Hoax News Detection by It's Headline

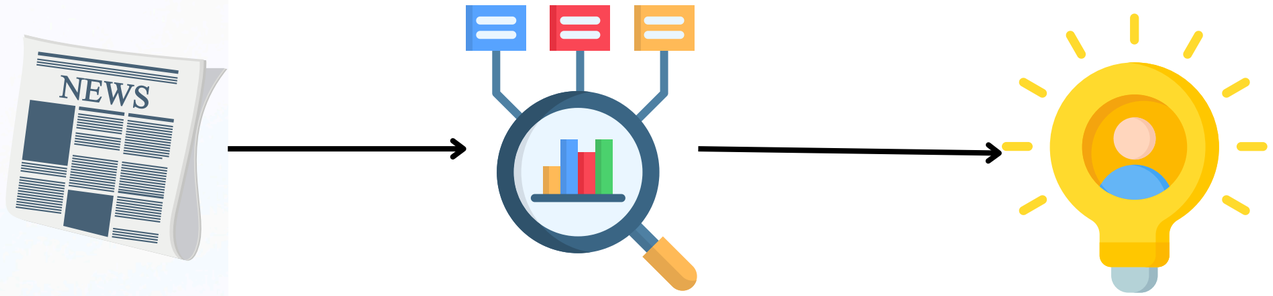
NATURAL LANGUAGE PROCESSING ON DETECTING A HOAX NEWS USING INDONESIAN LANGUAGES PROCESSING

Dhewa Radya
6701011256

What is the project about?

Hoax in Inondesia

In the context of Indonesia, hoax news is prevalent in various forms, including fabricated headlines, manipulated images, and conspiracy theories shared widely on social media platforms and messaging apps like WhatsApp.



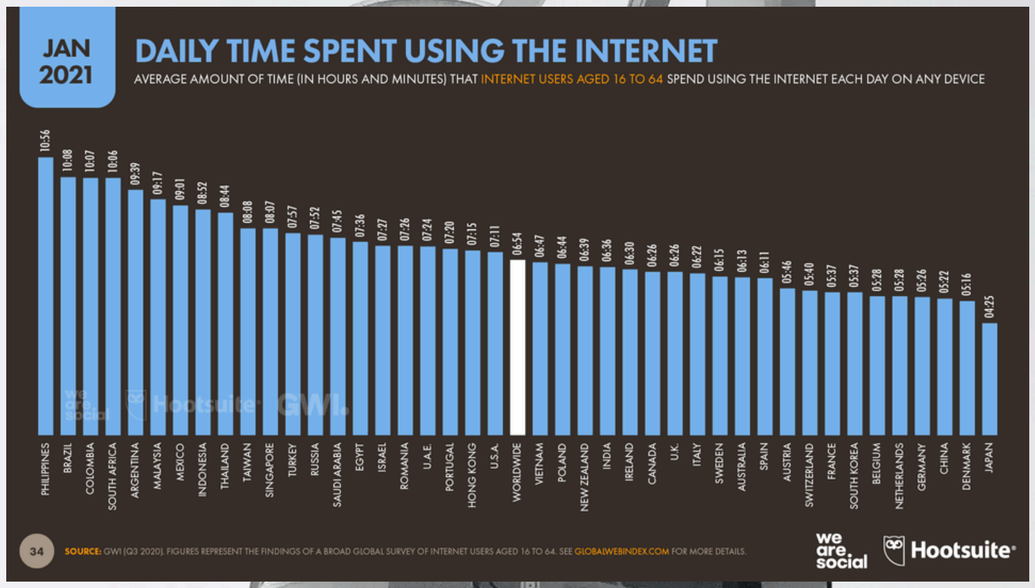
Why then?

Digital Literacy in Indonesia

3.54

Digital literacy index in Indonesia 2020-2022, by type. According to a survey in 2022, Indonesia's digital literacy index score increased from 3.46 in 2020 to **3.54 in 2022**. Overall, Indonesia's digital skills and digital culture had improved, while digital ethics and digital safety in the archipelago had weaken. Aug 9, 2567 BE

"Moreover, it's still not much papers and research on Indonesian languages in Data Processing"



05

Probabilistic Modeling

Detecting how words are classified as hoax

Supervised Machine Learning

Predicting the results to help people

Interactive Visualization

Interactivity on doing the predictions.

This project will focus on the usage of Data to do Predictions

Data Usage

Scrapping through Python

SATUSEHAT Application

```
result, continuation_token = reviews(
  'com.telkom.tracencare',
  lang='id',
  country='id',
  sort=Sort.NEWEST,
  count=2000,
  filter_score_with=None
)
```
[8]

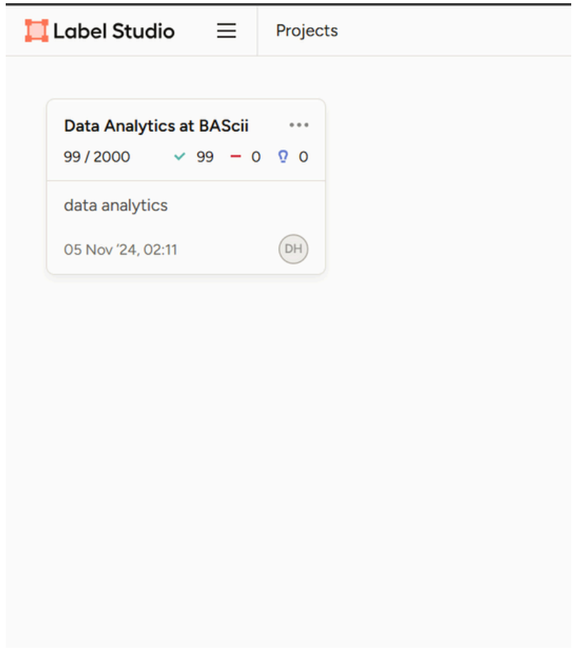
```
import pandas as pd
```
[1]

```
df = pd.read_csv('halodoc_review_dataset.csv')
```
[3]

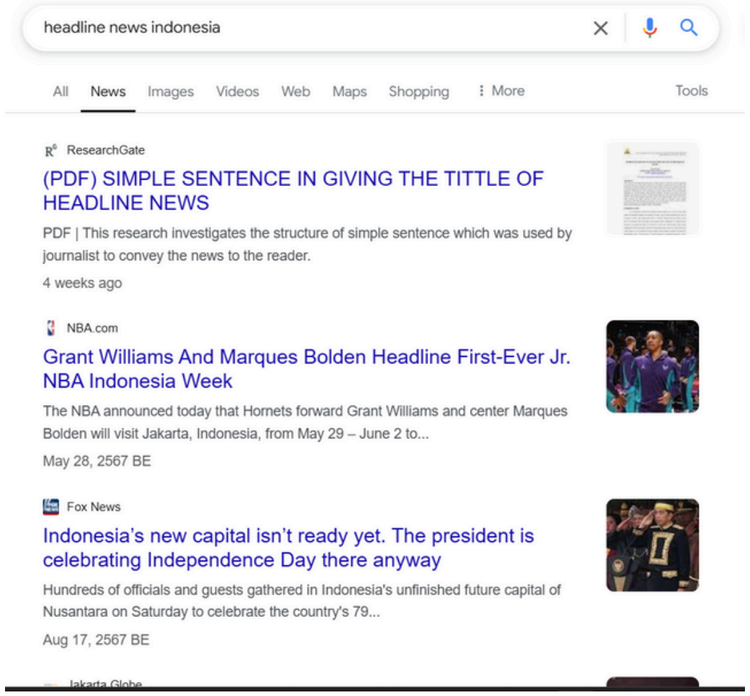
```
df.rename(columns={'content': 'text'}, inplace=True)
```
[4]

```
data = df.to_dict(orient='records')
```
[5]

Labeling through Data Studio



Preprocessing in Mathematica



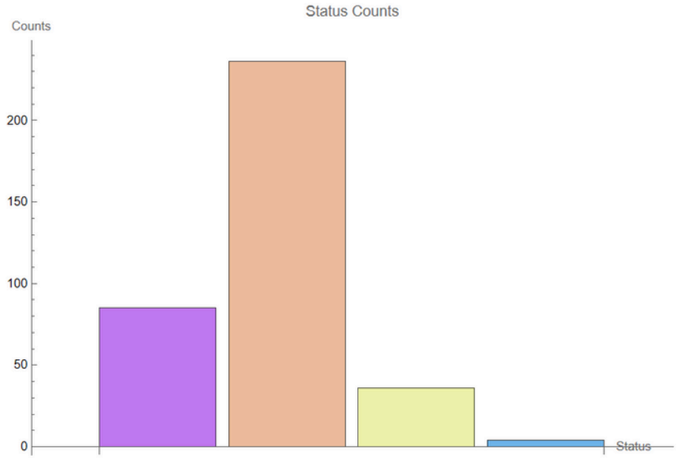
It contains 362 news headlines classified as HOAX and TRUE.

id	kode_provinsi	nama_provinsi	judul_berita	klasifikasi_utama	klasifikasi_menyinggung	status_berita	bulan	satuan	tahun
1	32	jawa_barat	RS HASAN SADIKIN BANDUNG....	KESEHATAN	0	BENAR	AGUSTUS	3	2022

```
In[3]:= dimData = Dimensions[dataset]

Out[3]= {362, 11}
```

Start with EDA

BENAR	85
DISINFORMASI (HOAKS)	236
MISINFORMASI (HOAKS)	36
0	4



Also checking the **duplicate** and **null data**

Checking Duplicates

```
duplicateCounts = Tally[
  Select[selectedData[All, 1], Count[selectedData[All, 1], #] > 1 &]
  ]
```

TUJUH PENUMPANG LION AIR RUTE TIONGKOK – MANADO TERINFEKSI VIRUS CORONA	2
PEMERINTAH BERI INTERNET GRATIS AKIBAT CORONA VIRUS	3
MUI PUSAT HIMBAU KIYAI DAN USTADZ AGAR TOLAK RAPID TEST, KARENA ITU MODUS OPERANDI PKI	3
KEMENTERIAN KESEHATAN ITALIA ; COVID19 BUKAN KARENA VIRUS MELAINKAN BAKTERI DAN DIPERKUAT RADIASI 5G MNYEBABKAN PERADANGAN DAN HIPOKSIA	2
COVID-19 BISA DISEMBUHKAN DENGAN MINYAK KAYU PUTIH	3
IMBAUAN DISDUKCAPIL BANDUNG TENTANG KASUS PENYALAHGUNAAN KTP	3
KUOTA BELAJAR TELKOMSEL 10GB HARGA RP10	2
BROADCAST PERAMPOKAN BERMODUS PEMBAGIAN MASKER GRATIS MENGANDUNG BIUS	2
BIMA ARYA GELAR RAZIA GDS, SISWA YANG TERJARING DIBERI SANKSI MENYEMPROT LINGKUNGAN DENGAN RADIUS 1000 METER	2
DAFTAR PRAKERJA MELALUI SITUS HTTPS://PRAKERJA.VIP	4
BROADCAST BSSN AKAN REKAM TELEPON DAN PANTAU WA, TWITTER, FACEBOOK	3
RAYAKAN ULANG TAHUN KFC TAWARKAN 3000 SNACK BUCKET UNTUK SEMUA ORANG	2
BIAYA TILANG TERBARU DI INDONESIA : KAPOLRI BARU MANTAP	2
BANTUAN SOSIAL FINANSIAL RP. 3.550.000 BAGI YANG KERJA TAHUN 2000 DAN 2021	2

Describing the Dataset

```
describeData = {
   "Count" → Length[selectedData],
   "Unique Titles" → Length[Union[selectedData[All, 1]]],
   "Unique Status" → Length[Union[selectedData[All, 2]]],
   "Most Frequent Title" → First@Commonest[selectedData[All, 1]],
   "Most Frequent Status" → First@Commonest[selectedData[All, 2]]
   }
```

```
{Count → 361, Unique Titles → 361, Unique Status → 361,
 Most Frequent Title → DAFTAR PRAKERJA MELALUI SITUS HTTPS://PRAKERJA.VIP, Most Frequent Status → DISINFORMASI (HOAKS)}
```

Data Cleaning

Drop Duplicates

Not taking the same value twice to learn, it will disturb the way the machine learns

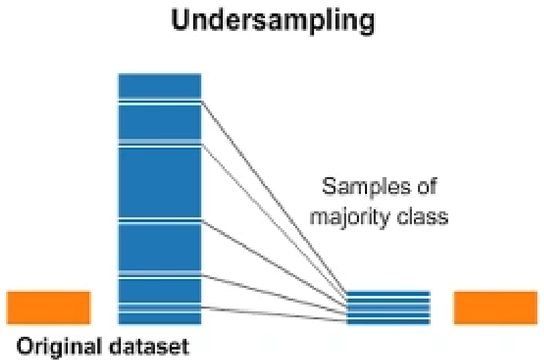
Fixing the Classification

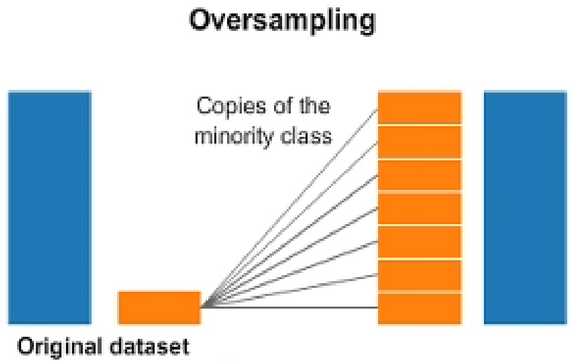
To just make the classifications as 2 different class

{BENAR}	82
{HOAX}	244
{0}	4

Class Balancing

So will not make confussion on some answers





Natural Language Processing

Wordopt, tokenize, stemming, stopwords

Word Opt

Natural Langueage Preprocessing

The Wordopt function is a custom text cleaning tool designed to prepare text data for analysis by removing various unwanted elements.

(I name it myself to sounds like word - optimization)

```
wordopt[text_String] := Module[{cleanedText = text},

  (* Convert text to lowercase *)
  cleanedText = ToLowerCase[cleanedText];

  (* Remove text within brackets *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["\\[.*?\\]"] → ""];

  (* Remove URLs *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["https?://[\\S]+|www\\.[\\S]+"] → ""];

  (* Remove HTML tags *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["<[^>]*>"] → ""];

  (* Remove HTML character codes (like &amp;) *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["&[A-Za-z0-9]+;"] → " "];

  (* Remove punctuation *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["[^\\w\\s]"] → ""];

  (* Remove extra whitespace *)
  cleanedText = StringReplace[cleanedText,
    RegularExpression["\\s+"] → " "];
```

- **Converting lo lowercase**
- **Removing brackets, URL, HTML, Tags, Character codes**
- **Removing punctuation**
- **Removing newline and extra whitespace**

Expected output

```
In[24]:= text = "Here's some [bracketed] text with a <b>HTML tag</b> and a URL https://example.com";
         wordopt[text]

Out[25]= heres some text with a html tag and a url
```

Tokenization

Natural Langueage Preprocessing

Breaking down text into individual words or tokens. To simplify processing, improve understanding, and enable further analysis

```
(* Basic word tokenization *)
tokenizeBasic[text_String] := DeleteStopwords[StringSplit[text]]

(* More advanced tokenization with options *)
tokenizeAdvanced[text_String, opts:OptionsPattern[]] := Module[
    {words, minLength = OptionValue[MinWordLength]},

    (* Split into words *)
    words = StringSplit[text];

    (* Remove stopwords if specified *)
    If[OptionValue[RemoveStopwords],
        words = DeleteStopwords[words]
    ];

    (* Filter by minimum length if specified *)
    If[minLength > 0,
        words = Select[words, StringLength[#] ≥ minLength &]
    ];

    words
]

(* Set default options *)
Options[tokenizeAdvanced] = {
    MinWordLength → 0,
    RemoveStopwords → True
};
```

Expected output

```
In[31]:= text = "this is a sample text";
         tokens = tokenizeBasic[text]

Out[32]= {this, is, a, sample, text}
```

StopWords

Natural Langueage Preprocessing

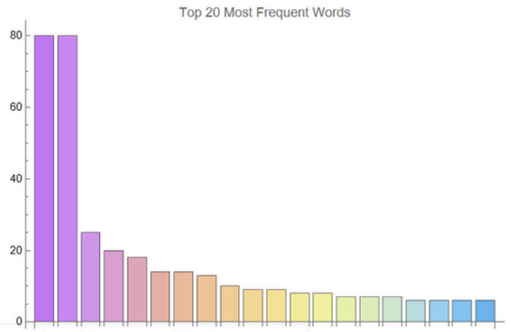
Removing common words that don't carry much meaning (e.g., "the," "and," "of").

```
(* Define Indonesian stopwords *)
indonesianStopwords = {
    "yang", "di", "ke", "dari", "pada", "dalam", "untuk", "dengan", "dan", "atau",
    "ini", "itu", "juga", "sudah", "saya", "aku", "kamu", "dia", "mereka", "kita",
    "akan", "bisa", "ada", "tidak", "saat", "oleh", "setelah", "para", "seperti",
    "saya", "anda", "dia", "mereka", "kita", "kami", "nya", "lah", "pun", "kan",
    "ku", "mu", "si", "sang", "kaum", "bagi", "sebuah", "seorang", "telah",
    "tetap", "buat", "masih", "hal", "ketika", "kepada", "sebagai", "sampai",
    "dahulu", "sangat", "sering", "sendiri", "sekarang", "tapi", "tentang",
    "selain", "tersebut", "apabila", "bagaimana", "menurut", "hampir", "dimana",
    "bagaimana", "siapa", "mengapa", "kapan", "yakni", "dimana", "kemana",
    "pula", "selama", "sekitar", "yaitu", "namun", "karena", "jika", "bila",
    "kalau", "oleh", "sejak", "ialah", "bahwa", "hanya", "lain", "sambil",
    "setelah", "sebab", "maka", "selagi", "sementara", "sebelum", "supaya",
    "semua", "setiap", "beberapa", "banyak", "sebagian", "lalu", "melalui",
    "dimana", "diantara", "keduanya", "semenjak", "sedangkan", "sebegitu",
    "seadanya", "sebetulnya", "sesungguhnya", "sepertinya"
};
```

Expected output

```
In[39]:= text = "Saya sedang belajar pemrograman untuk analisis data";
         tokens = tokenizeIndonesian[text]

Out[40]= {analisis, belajar, data, pemrograman, sedang}
```

Out[46]=



BENAR	HOAX	covid19	bandung
corona	indonesia	vaksin	virus
kota	china	pasien	positif
video	uang	tahun	kuota
ditutup	jabar	meninggal	juta

Stemming

Natural Langueage Preprocessing

Reducing words to their root form (e.g., "running" -> "run")

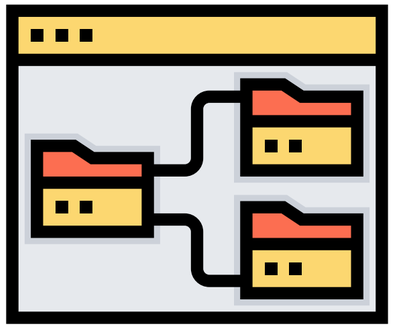
```
(* Define affixes *)
prefixes = {"be", "me", "pe", "te", "di", "ke", "se"};
complexPrefixes = {"ber", "bel", "pel", "per", "pem", "pen", "peng", "meng", "mem", "men", "ter"};
suffixes = {"i", "an", "kan"};
possessivePronounds = {"ku", "mu", "nya"};
particles = {"lah", "kah", "tah", "pun"};
```

```
(* Helper function to check if word ends with suffix *)
endsWithAny[word_, suffixList_] :=
    AnyTrue[suffixList, StringEndsQ[word, #] &]

(* Main stemming function *)
stemIndonesian[word_String] := Module[
    {result = word, step = 1},

    (* Step 1: Remove particles *)
    If[endsWithAny[result, particles],
        result = StringDrop[result, -3];
        step++;
    ];

    (* Step 2: Remove possessive pronouns *)
    If[endsWithAny[result, possessivePronounds],
        result = StringDrop[result, -3];
        If[StringLength[result] ≥ 3, step++];
    ];
```

but it hasnt work that well, so I can't use it yet

Modeling

(from left to right)



Train-test split

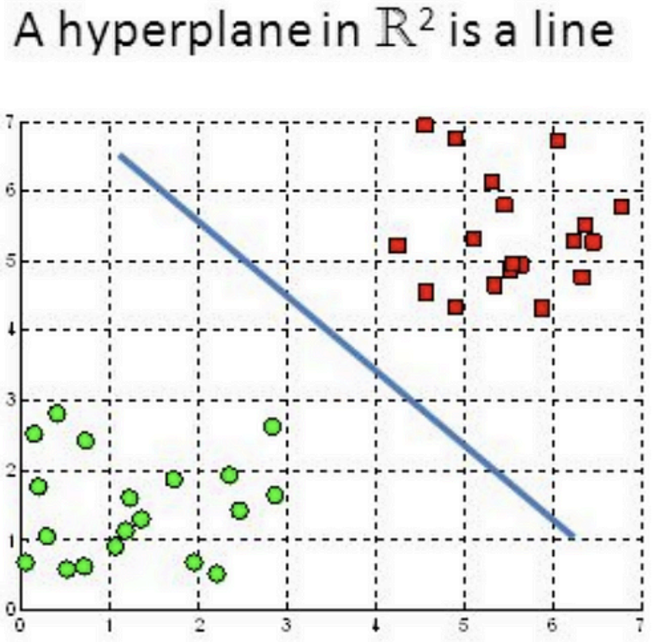
X -> Y

Association Thread

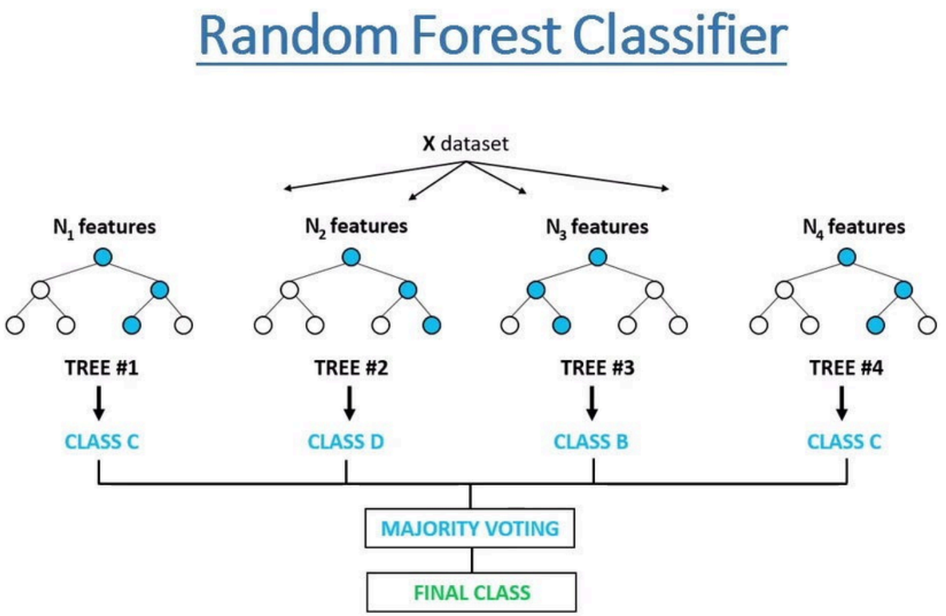
Machine Learning

Evaluation

Models

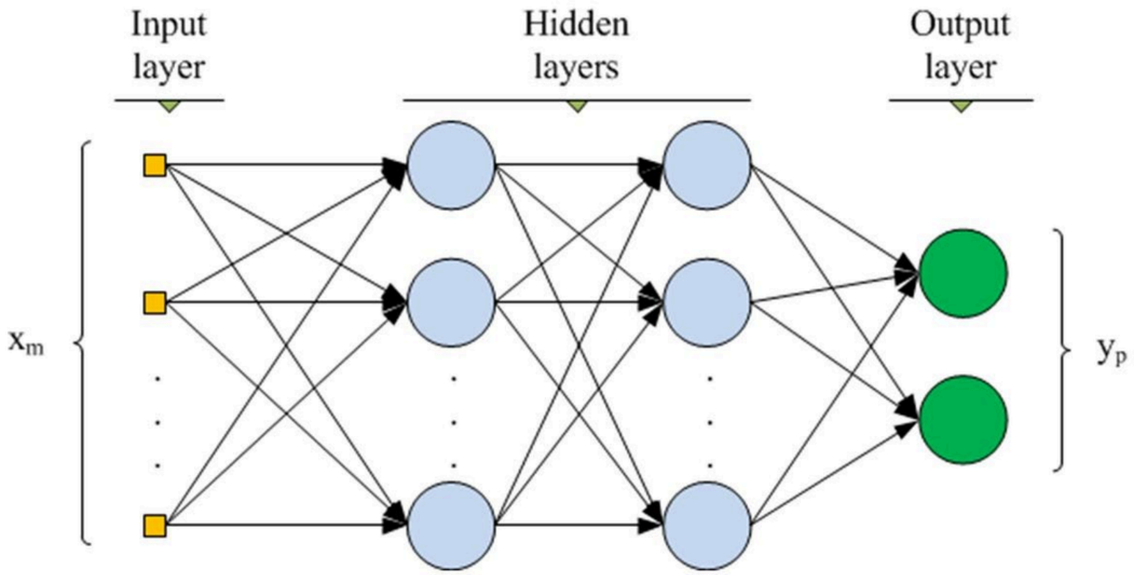


SVM

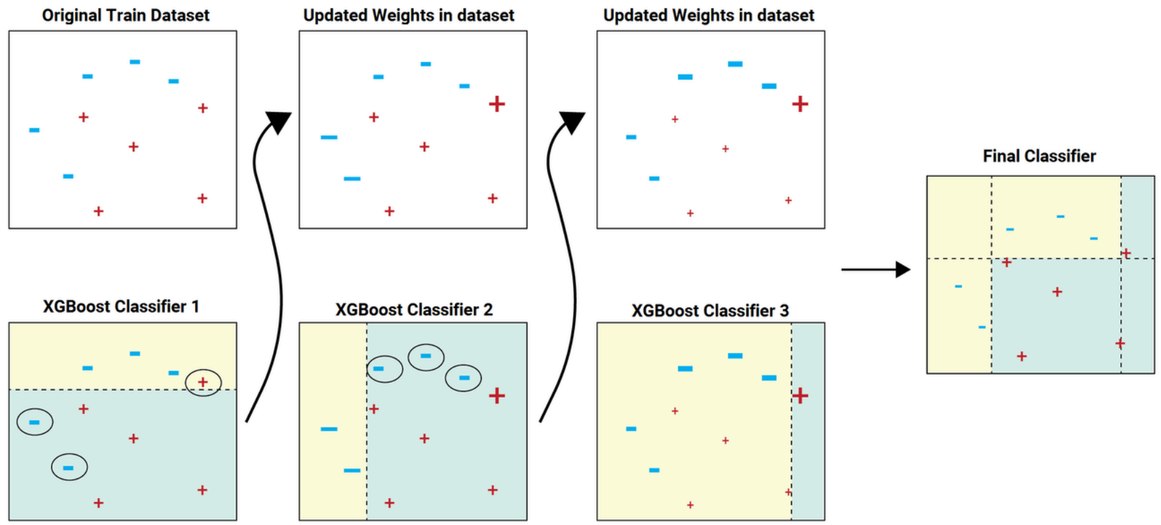


RFC

Models

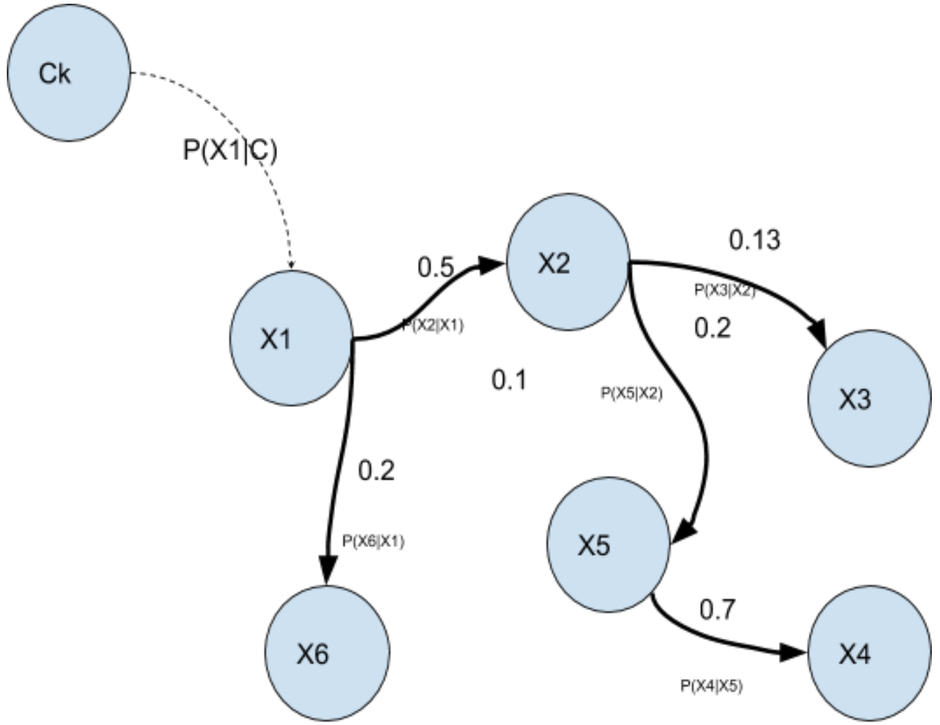


Neural Network



XGB

Models



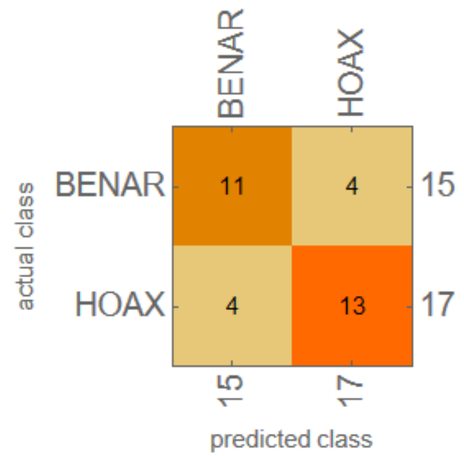
Example #1

Markov

The Model I Use

Classifier Measurements

Classifier method	Markov
Number of test examples	32
Accuracy	(75. ±8.)%
Accuracy baseline	(53. ±9.)%
Geometric mean of probabilities	0.382 ± 0.13
Mean cross entropy	0.962 ± 0.33
Single evaluation time	3.87 ms/example
Batch evaluation speed	724. examples/s



accuracy: **0.75 /1**

In the context of hoax news detection, recall is generally more important than precision.

False Negatives are More Harmful: means that a hoax news article is incorrectly classified as true. Lead misinformation spreading and potentially causing harm.

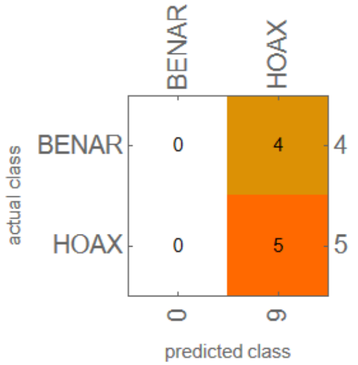
False Positives are Less Harmful: means a true news article is incorrectly classified as a hoax. While this can be annoying or misleading

Evaluation

Suport Vector Machine

Classifier Measurements

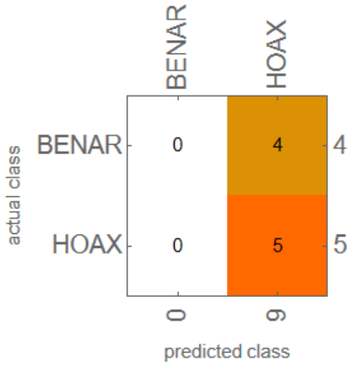
Classifier method	SupportVectorMachine
Number of test examples	9
Accuracy	(56. ±18.)%
Accuracy baseline	(56. ±18.)%
Geometric mean of probabilities	0.499 ± 0.040
Mean cross entropy	0.695 ± 0.081
Single evaluation time	6.97 ms/example
Batch evaluation speed	179. examples/s



Random Forest

Classifier Measurements

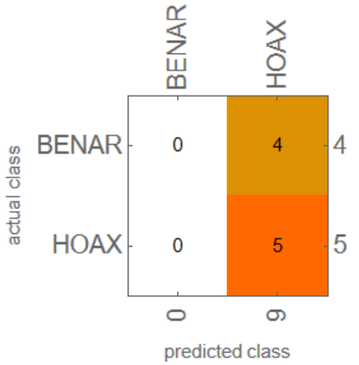
Classifier method	RandomForest
Number of test examples	9
Accuracy	(56. ±18.)%
Accuracy baseline	(56. ±18.)%
Geometric mean of probabilities	0.517 ± 0.014
Mean cross entropy	0.659 ± 0.027
Single evaluation time	31.9 ms/example
Batch evaluation speed	289. examples/s



Neural Network

Classifier Measurements

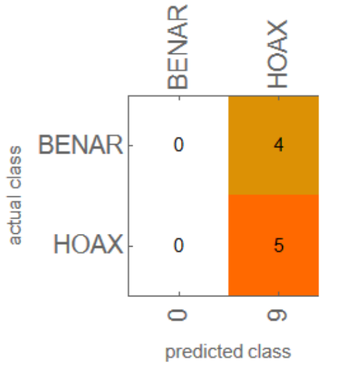
Classifier method	NeuralNetwork
Number of test examples	9
Accuracy	(56. ±18.)%
Accuracy baseline	(56. ±18.)%
Geometric mean of probabilities	0.498 ± 0.011
Mean cross entropy	0.697 ± 0.022
Single evaluation time	23.6 ms/example
Batch evaluation speed	248. examples/s



Gradient Boosting

Classifier Measurements

Classifier method	GradientB
Number of test examples	9
Accuracy	(56. ±18.)%
Accuracy baseline	(56. ±18.)%
Geometric mean of probabilities	0.503 ± 0.
Mean cross entropy	0.687 ± 0.
Single evaluation time	8.62 ms/e
Batch evaluation speed	294. exar



Precision: Important when false positives have a high cost (e.g., predicting fraud when it's not fraud).

Recall: Important when false negatives have a high cost (e.g., missing a cancer diagnosis).

F1-Score: Useful when there's a need to balance precision and recall, or when the class distribution is highly imbalanced.

Interactive Visualization

Indonesian Text Sentiment Classifier

Select Model:

Markov

Accuracy: % 75.00

Enter Text for Classification:

Aku cinta BAScii

Classify Text

Classification Result:
BENAR

Show Model Comparisons