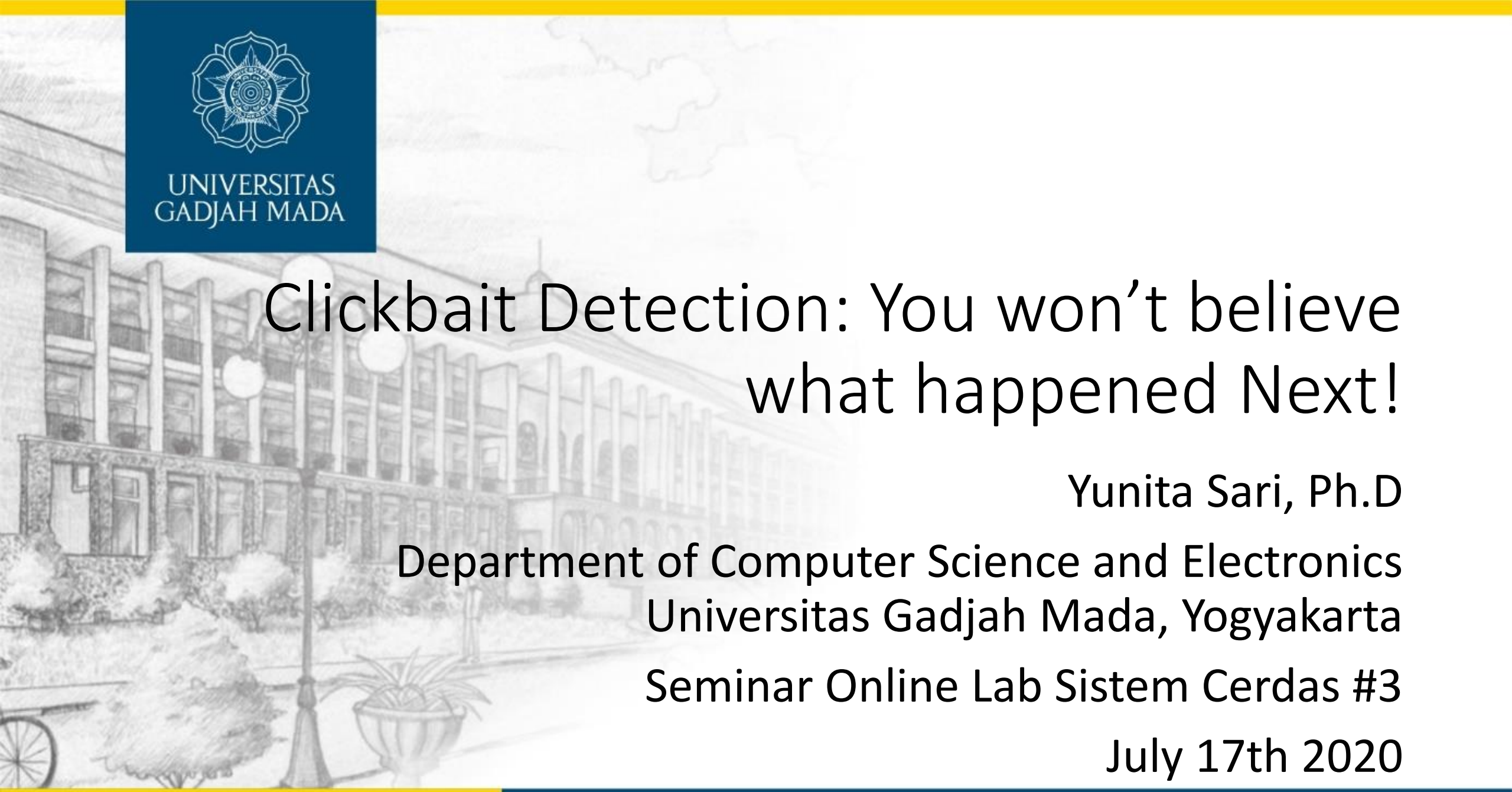


UNIVERSITAS
GADJAH MADA

Clickbait Detection: You won't believe what happened Next!

Yunita Sari, Ph.D

Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta

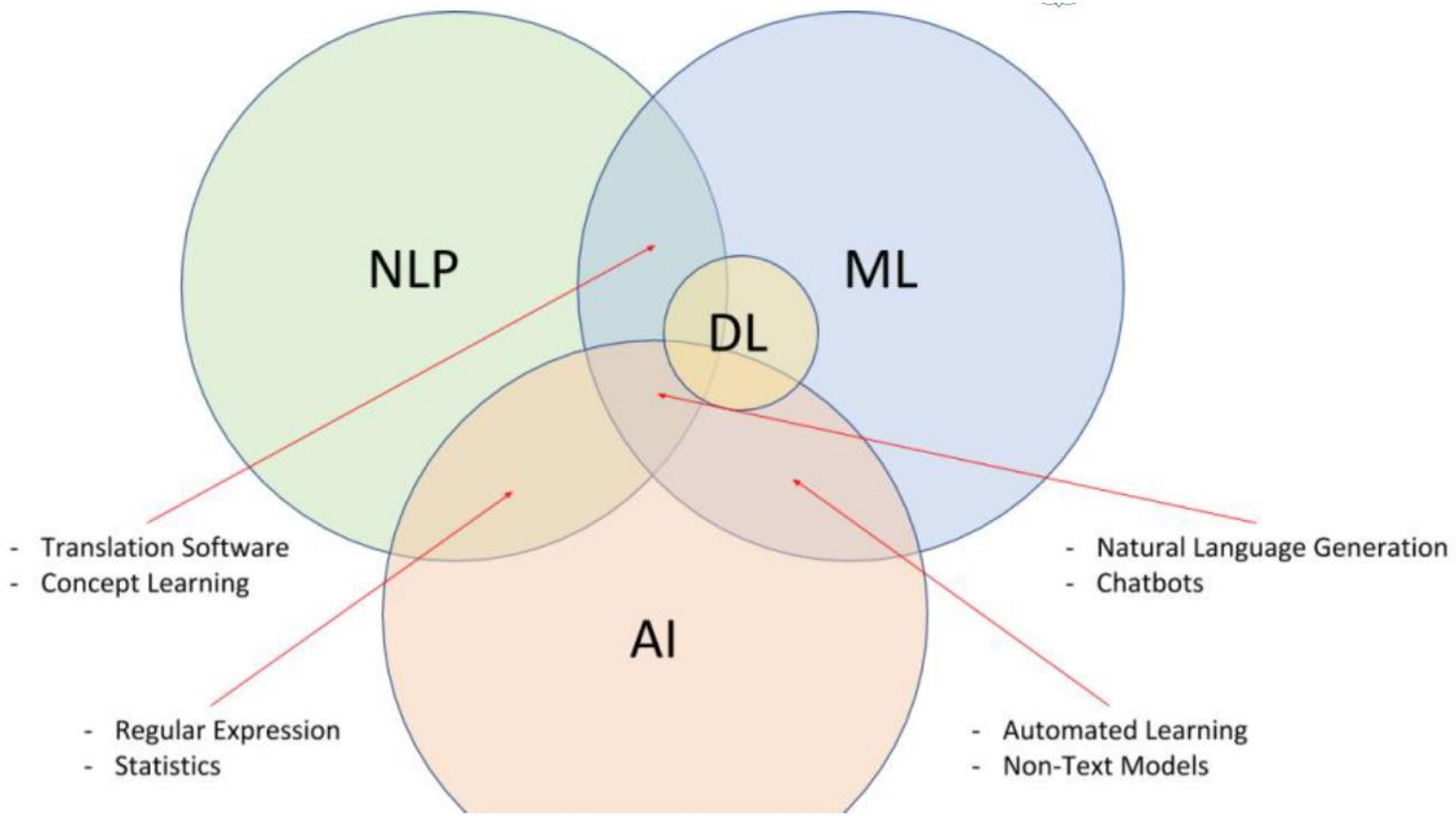
Seminar Online Lab Sistem Cerdas #3

July 17th 2020

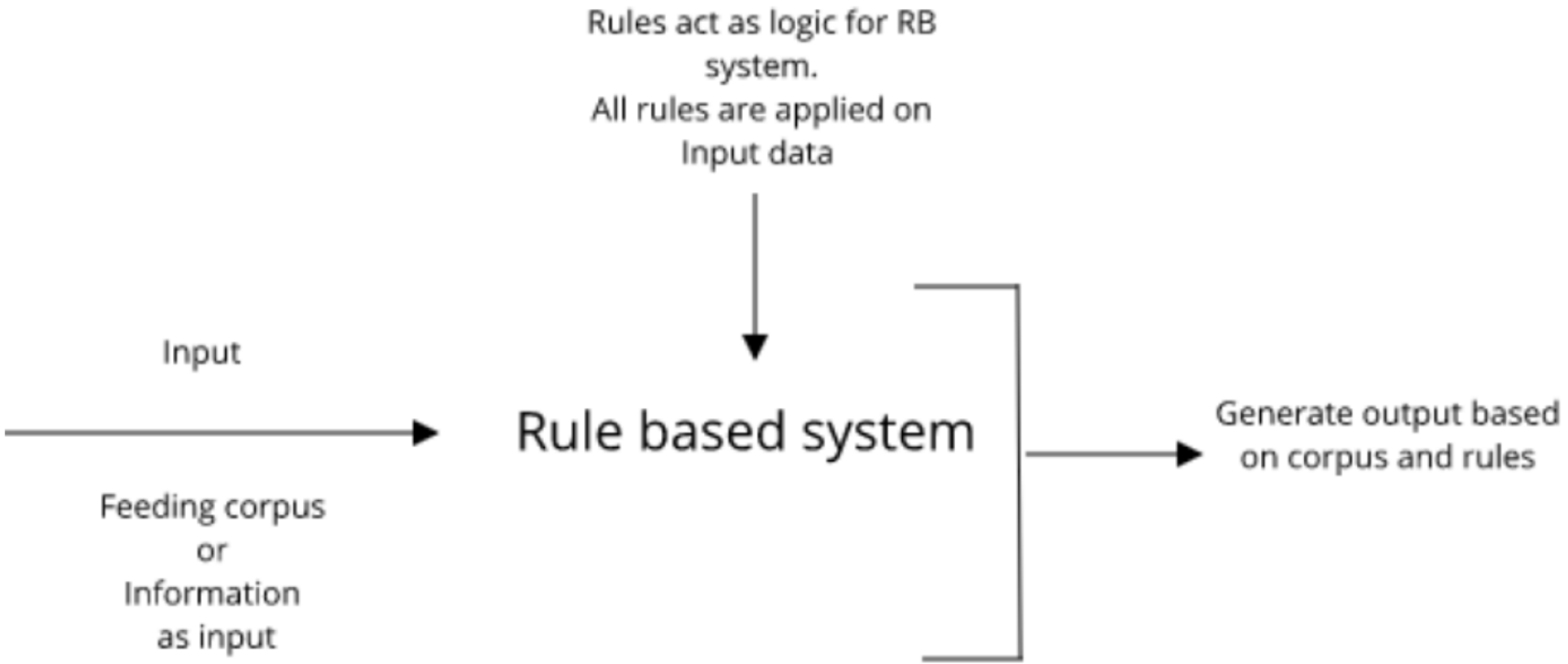
NLP is Challenging

" Human language is highly ambiguous … It is also ever changing and evolving. People are great at producing language and understanding language, and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings. At the same time, while we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language"
Page 1, Neural Network Methods in Natural Language Processing, 2017.

NLP-AI-ML-DL


NLP
ML
DL
AI
- Translation Software
- Concept Learning
- Natural Language Generation
- Chatbots
- Regular Expression
- Statistics
- Automated Learning
- Non-Text Models

Rule-based Approach

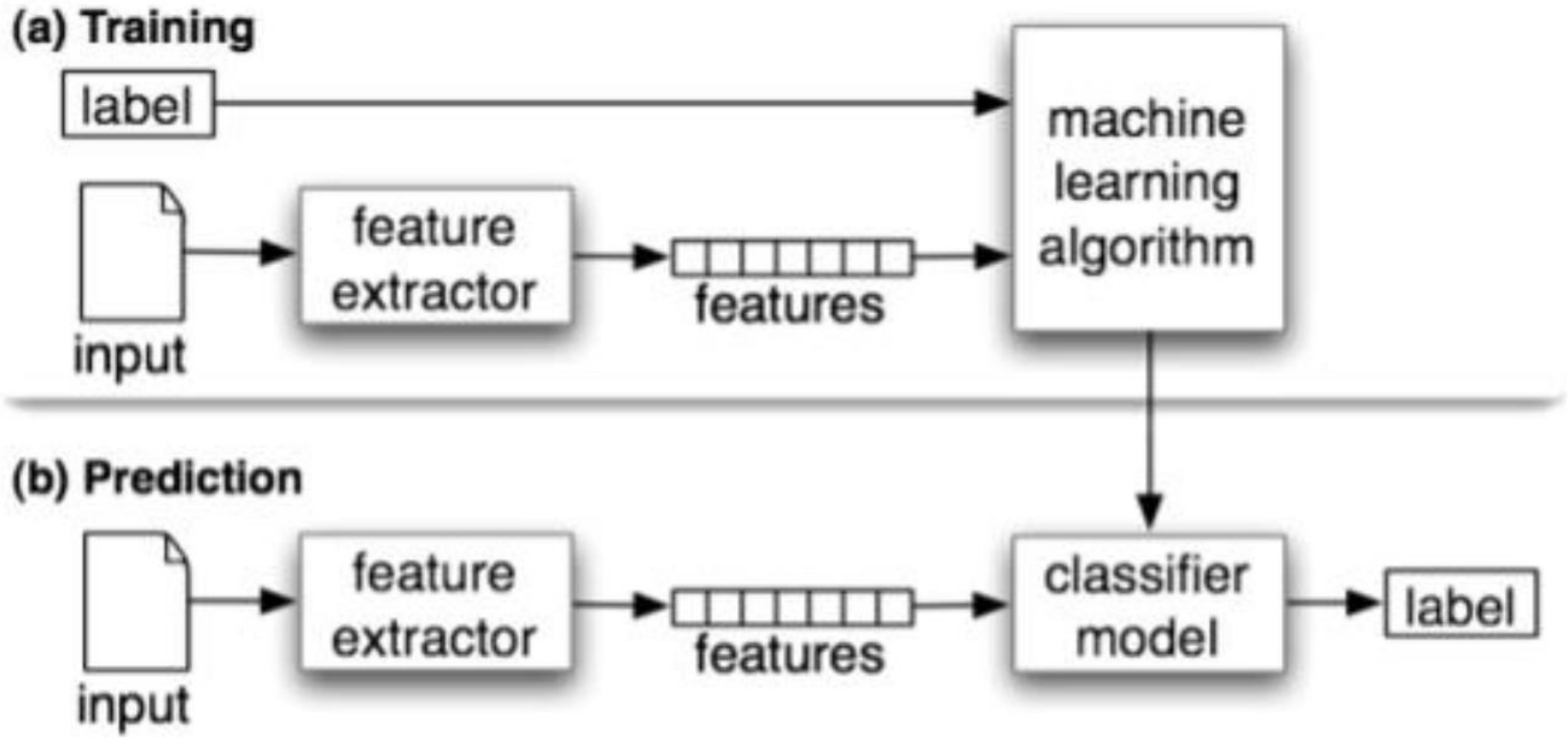


Example Rule-based

Rule: for every sentence containing the word “language”, remove the first word, and any remaining capitalized words are programming languages.

- The language above is **Python**.
- In 2013 the ten most popular programming languages are (in descending order by overall popularity): **C, Java, PHP, JavaScript, C++, Python, Shell, Ruby, Objective-C** and **C#**.
- Edsger **Dijkstra**, in a famous 1968 letter published in the **Communications** of the **ACM**, argued that **GOTO** statements should be eliminated from all “higher level” programming languages.
- Java came to be used for server-side programming.

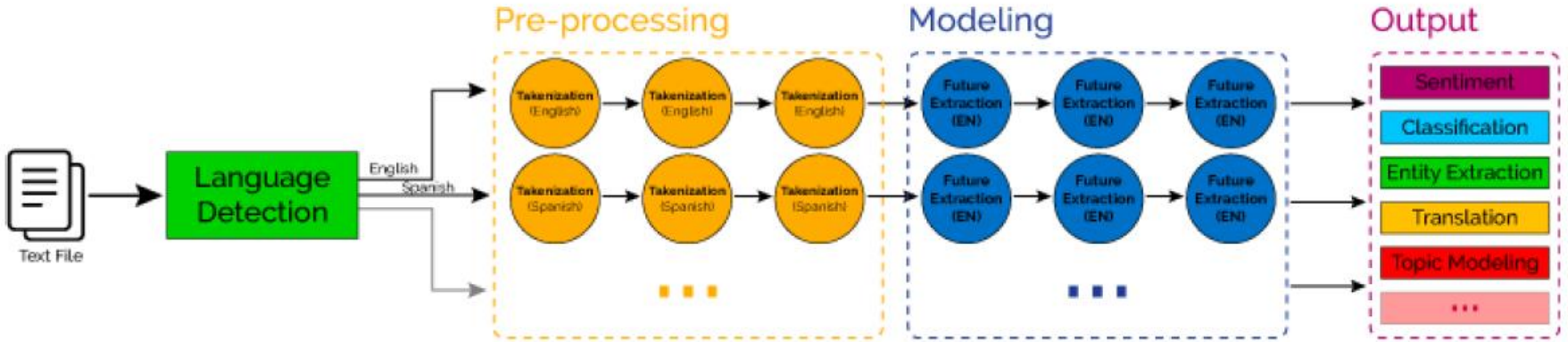
Supervised Machine Learning-based Approach



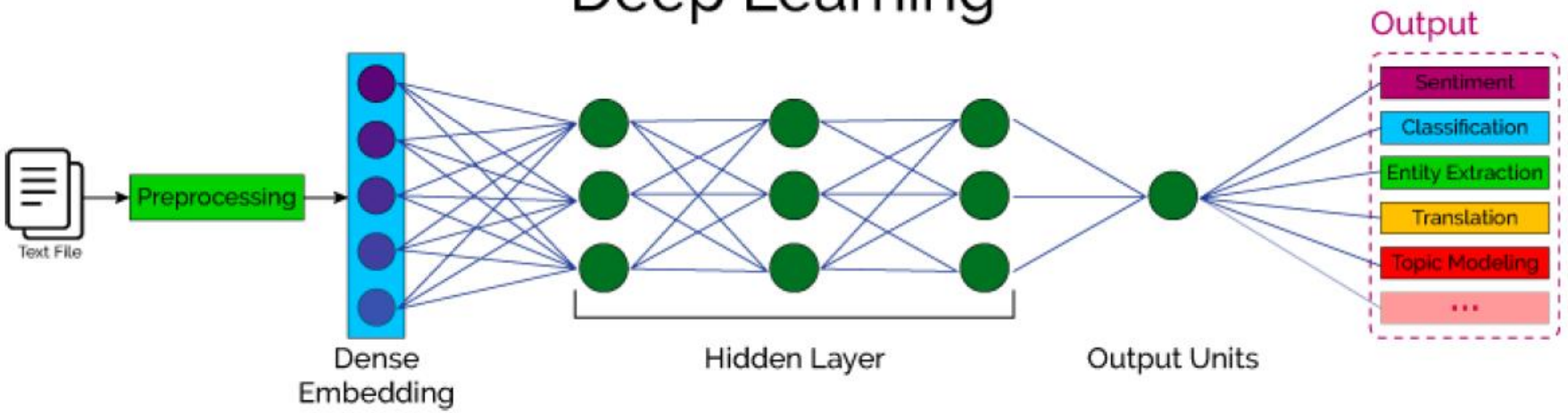
Rule-based vs ML approach

	Rule-based grammar	Machine Learning algorithm
+	★ Flexible ★ Easy to debug ★ Doesn't require a massive training corpus ★ Understanding of the language phenomenon ★ High precision	★ Easy to scale ★ "Learnability" without being explicitly programmed ★ Fast development (if datasets available) ★ High recall (coverage)
−	★ Requires skilled developers and linguists ★ Slow parser development ★ Moderate recall (coverage)	★ Requires training corpus with annotation ★ Difficult to debug ★ No understanding of the language phenomenon

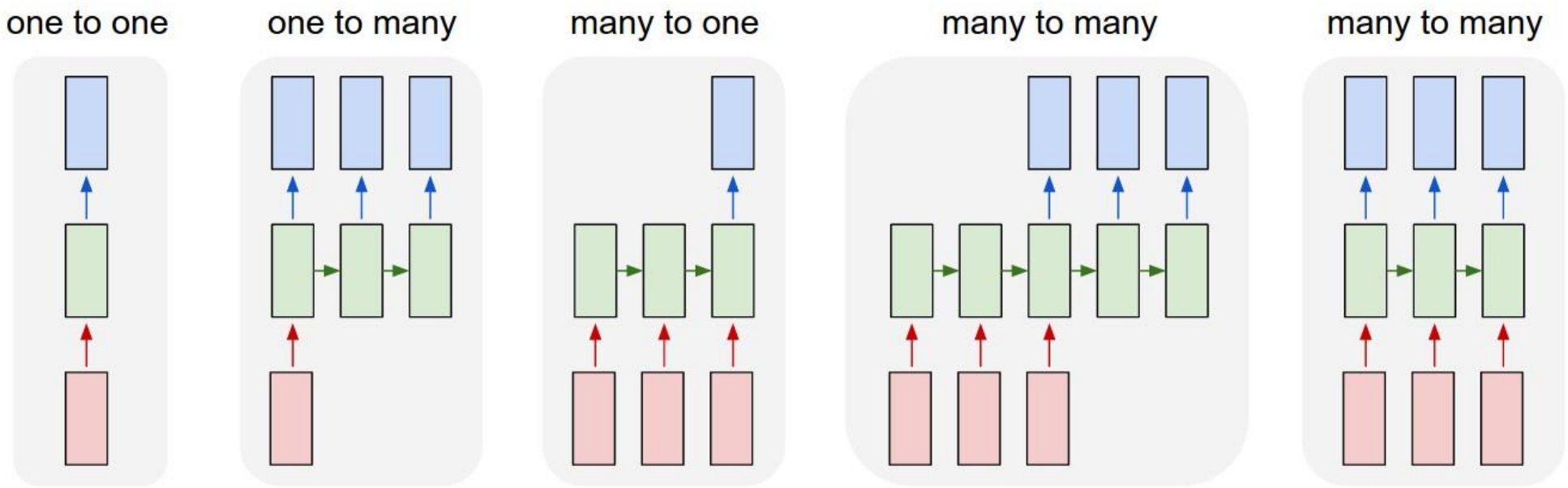
Classical NLP



Deep Learning



RNN Architectures



(1) Vanilla mode of processing without RNN, from fixed-sized input to fixed-sized output (e.g. image classification). (2) Sequence output (e.g. image captioning takes an image and outputs a sentence of words). (3) Sequence input (e.g. sentiment analysis where a given sentence is classified as expressing positive or negative sentiment). (4) Sequence input and sequence output (e.g. Machine Translation: an RNN reads a sentence in English and then outputs a sentence in French). (5) Synced sequence input and output (e.g. video classification where we wish to label each frame of the video).

Source: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

The Clickbait Team

Yunita

Andika

Gavin

Dennis

Hilda

Felix

Clickbait?

Viral! Driver Ojol di Bekasi Antar Pesanan Makanan Pakai Sepeda”

“Ada Motor Nyangkut di Atas Bambu di Sleman, Kok Bisa?”

“Seorang Pria Depresi Bakar Mobilnya Sendiri di Tempat Pembuangan Akhir”

Why Clickbait Detection?

There is a growing trend of publishers making use of clickbait

(Gianotto, 2014; Potthast *et al.* 2016)

The state of rampant clickbait usage is affecting the user experience as a whole.

(Tan and Ang, 2017)

(William, 2020)

Clickbait Corpus – Samples

No	Publisher	Articles	Sample
1	detik.com	5853	1000
2	fimela	789	700
3	kapanlagi	1007	1000
4	kompas.com	3253	1500
5	liputan6.com	4581	1500
6	okezone.com	4664	1500
7	postmetro medan	308	300
8	republika.co.id	5782	1500
9	sindonews.com	3572	1500
10	tempo.co	4026	1500
11	tribunnews.com	9662	1500
12	wowkeren.com	3020	1500
	Total	46517	15000

(William, 2020)

Clickbait Corpus - Results

Label	Headlines	Full Agreements (FA)	FA Percentage
Clickbait	6290	3316	52.7%
Non-Clickbait	8710	5297	60.8%
Total	15000	8613	57.4%

(William, 2020)

Clickbait Corpus – Clickbait Percentage

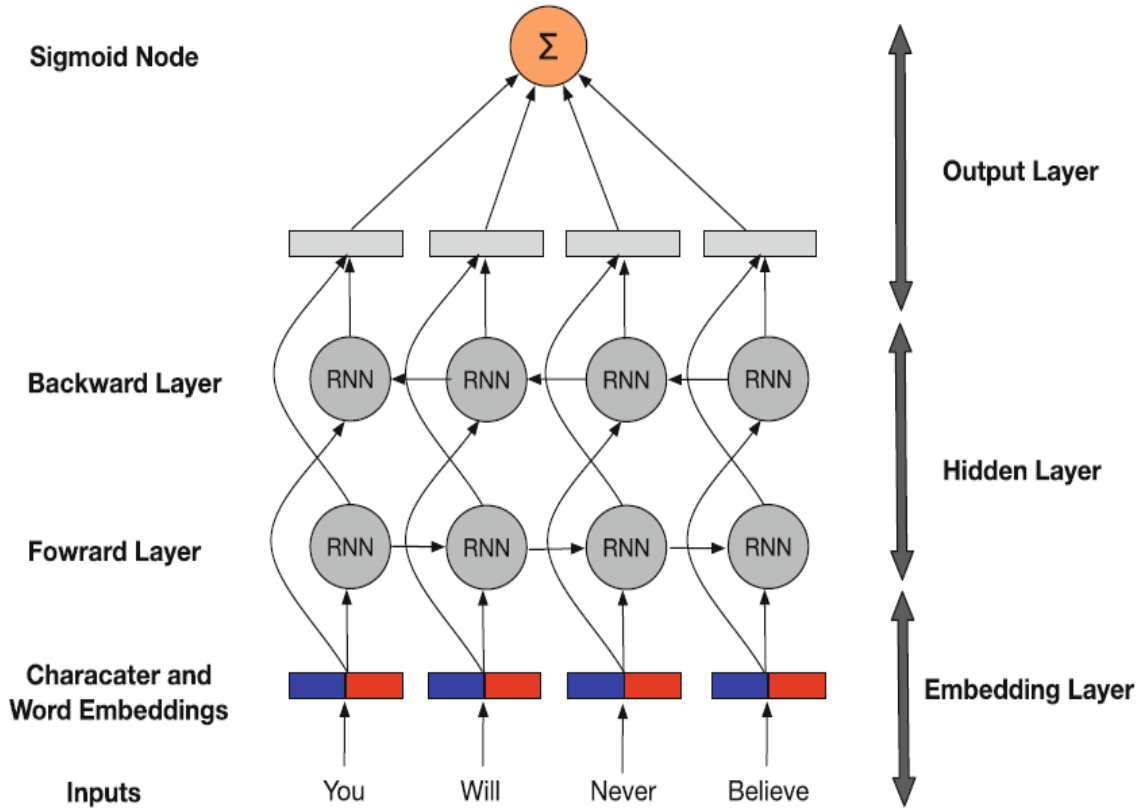
Publisher	Sample	Non-Clickbait	Clickbait	Clickbait Percentage
detikNews	1000	890	110	11.0%
fimela	700	306	394	56.3%
kapanlagi	1000	603	397	39.7%
kompas	1500	1157	343	22.9%
liputan6	1500	613	887	59.1%
okezone	1500	741	759	50.6%
Posmetro-medan	300	71	229	76.3%
republika	1500	1267	233	15.5%
sindonews	1500	1215	285	19.0%
tempo	1500	1118	382	25.5%
tribunnews	1500	451	1049	69.9%
wowkeren	1500	278	1222	81.5%
Total	15000	8710	6290	41.9%

(William, 2020)

Clickbait Corpus – Clickbait Percentage per category

Category Groups	Category Headlines	Clickbait	Clickbait percentage
News	6379	1918	30.1%
Celebrity & Entertainment	3009	1998	66.4%
Sports	1727	620	35.9%
Business & Economy	1463	380	26.0%
Lifestyle	1315	878	66.8%
Science & Technology	411	179	43.6%
Automotive	321	146	45.5%
Religion & Culture	194	91	46.9%
Others	181	80	44.2%
Total	15000	6290	41.9%

Research Methodology – Bi-LSTM



Bi-LSTM Architecture (Anand *et al.,* 2017)

(William, 2020)

Results – BILSTM Detection

	Bi-LSTM						
Dataset	Epochs	Duration*	Loss	Acc	Precision	Recall	F1
MainWithSymbol	7	45s	0.5082	0.7790	0.7851	0.6521	0.7052
MainNoSymbol	7	43s	0.4979	0.7847	0.7756	0.6726	0.7144
MainStemmedWords	7	49s	0.5556	0.7330	0.7095	0.6208	0.6527
FAWithSymbol	7	42s	0.3059	0.8503	0.8544	0.8531	0.8475
FANoSymbol	7	27s	0.3077	0.8845	0.8045	0.9106	0.8492
FAStemmedWords	7	22s	0.4225	0.8311	0.8079	0.7277	0.7598
Chakraborty *et al.* (2016)	10	136s	0.1111	0.9750	0.9784	0.9720	0.9743

(William, 2020)

Attempt for Extending the dataset

- To create larger dataset, thus the performance can be improved
- Providing reliable and free dataset that can be used for other NLP tasks
- Accelerating the progress of NLP tasks especially for Bahasa Indonesia.


Clickbait
Yunita
Media Australia: Ada Kemiripan dalam Penanganan COVID-19 Anies deng York
SHOW CONTENT
SKIP
TUTORIAL - STEP 1
Selamat datang di linguaksara!
Dimana anda dapat melabel artikel clickbait dan mengumpulkan poin.
Next
History
Jokowi: Terapi Plasma D
Polisi Tangkap 202 Mobil
Diterjang Banjir Bandang, 6 Rumah dan 1 Sekolah di Sukajaya Bogor Terancam Ambruk
Kasus Positif Covid-19 di Sulbar Kini Meniadi 73 Kasus.
Not Clickbait
Clickbait

linguaksara.site

thanks

/THaNGks/

noun

plural noun: **thanks**

an expression of gratitude.
"festivals were held to **give thanks for** the harvest"

Similar: gratitude | gratefulness | appreciation | acknowledgment | recognition

- another way of saying thank you.
  "**thanks for** being so helpful"