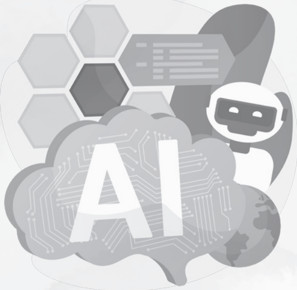


CHAPTER 19

NLP'S GOLDEN ERA IN INDONESIA: PROJECT BINA

On Lee[1]

GDP Labs and GDP Venture

ABSTRACT

Project BINA (Bahasa Indonesia NLP Alliances) is an Indonesia-based group of NLP leaders, innovators, researchers, practitioners, and businesses from around the globe collaborating to advance Natural Language Processing (NLP) technology for the official Indonesian language or Bahasa Indonesia. Its motto is "Advancing Indonesian language through collaboration, data and technology".

Over 400 exabytes (1018 bytes) of data will be generated globally each day by 2025. In addition to the official language, Indonesia has over 700 indigenous languages and dialects. Additionally, there are almost 130 million Meta (Facebook) users and about 19 million Twitter users, ranked #3 and ranked #5 most users in the world respectively. Furthermore, some Indonesian NLP researchers, innovators and practitioners have been researching NLP technology for the Indonesian language for decades and, similarly to other countries, progress has accelerated in recent years.

This is the golden era for NLP in Indonesia with talent, knowledge, data, technology and applications able to benefit the government, industry, academia, and the broader community.

Keywords: *Natural Language Processing, Bahasa Indonesia, Indigenous Languages, Dialects*

A. INTRODUCTION

1. What is Project BINA?

Project BINA (Bahasa Indonesia NLP Alliances) is an Indonesia-based group of NLP leaders, innovators, researchers, practitioners, and businesses from around the globe collaborating to advance the natural language processing (NLP) data and technology for the official Indonesian language, Bahasa Indonesia [1].

BINA is an Indonesian word which means architect, build, construct, or foster or make it better. Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence (AI) concerned with the interactions between computers and human language [2]. Project BINA's motto is "Advancing the Indonesian language through collaboration, data and technology".

[1] **On Lee** has over 30 years of experience in technology. He has built teams in the United States, Indonesia, China, and India and is a board member of several AI start-up companies.

O. Lee
GDP Labs & GDP Venture, e-mail: onlee@gdpventure.com

@2023 Kolaborasi Riset dan Inovasi Industri Kecerdasan Artifisial (KORIKA) & Penerbit BRIN
O. Lee, "NLP's Golden Era in Indonesia: Project BINA," in *Prosiding Use Cases Artificial Intelligence Indonesia: Embracing Collaboration for Research and Industrial Innovation in Artificial Intelligence*, B. R. Trilaksono, H. Riza, A. Jarin, N. D. S. Darmayanti, and S. Liawatimena, Eds. Jakarta: Penerbit BRIN, Februari 2023, ch. 19, pp. 227-234, doi: 10.55981/brin.668.c555
ISBN: 978-623-8052-49-3, E-ISBN: 978-623-8052-50-9

2. Why Now?

Holon IQ [3] stated in their 50 National AI Strategies report that "*Governments around the world see artificial intelligence (AI) as a nation defining capability. Countries are looking to their education systems to develop world-class generational AI capability while ensuring equity, privacy, transparency, accountability, economic, and social impact.*" These 50 countries represent 90% of the global gross domestic product. Indonesia is one of them.

From Wikipedia, the Vision of Indonesia 2045 is an Indonesian ideal that set the goal for the country to be a sovereign, advanced, fair, and prosperous nation by its centennial in 2045 [4]. The goal is set in 2045, since by then the republic will commemorate 100 years of independence.

Prof. Bambang P. S. Brodjonegoro, as the Minister of Research and Technology/ Head of National Research and Innovation Agency announced Indonesia's national AI strategy [5] during *Indonesia 25th National Technology Awakening Day* on August 10, 2020. The implementation of this AI strategy aligns with the sustainable development goals (SDG) and the Vision of Indonesia 2045.

KORIKA (*Kolaborasi Riset dan Inovasi Industri Kecerdasan Artifisial Indonesia*) has been formed to execute Indonesia's national AI strategy with collaboration between government, academia, industry and community, establishing *quad-helix* innovation ecosystem [6].

Over 400 exabytes (10^{18} bytes) of data will be generated globally each day by 2025. Indonesia has over 700 languages and dialects [7][8]. Additionally, there are almost 130 million Meta (Facebook) users and almost 19 million Twitter users, ranked #3 and ranked #5 most users in the world respectively [9]. Furthermore, some Indonesian NLP researchers, innovators, and practitioners have been researching NLP technology for Indonesian language for decades and, similarly to other countries, progress has accelerated in recent years.

3. Who? Customers, Contributors and Members

NLP is one of the latest technologies to emerge in the past few years. The recent breakthroughs in NLP follow patterns similar to the PC, the internet, and smartphone evolutions, which all started as niche areas and eventually became general-purpose technologies.

Text is everywhere! Every company produces and consumes texts. It means that everyone could be potential Project BINA customers. We could categorize them as government, academia, industry, and end-users.

There has been a strong interest in this area as shown by the initial contributors and members (in alphabetical order) as of this writing, AI Center ITB, APP Sinar Mas, BCA, Bahasa Kita, BRIN, Bukalapak, CATAPA, Datasaur.ai, DigitData.

ai, TekenAja, GDP Labs, GDP Venture, GLAIR.ai, INACL, KASKUS, Kontrak Hukum, KORIKA, Prosa.ai, Samsung, Telkom Indonesia, Tokopedia, and Twitter. In addition, the following in discussion are AWS, Blibli, KOMINFO, Meta, Nvidia, Solve (MIT), and Tiket.com.

I am confident that many more organizations globally will be interested in joining Project BINA, as Indonesia is the second most linguistically diverse and the fourth most populous nation in the world, and the benefits are abundant.

4. Benefits

Project BINA aligns with the Vision of Indonesia 2045 and one of the main programs of KORIKA. It collaborates with the government, academia, industry, and community to improve the official language bahasa Indonesia and more than 700 languages and dialects in Indonesia by initially cleaning and labeling raw data from government regulations, Meta, and Twitter and turning it into intelligent data. The above will benefit the customers and Project BINA members tremendously. The customers and Project BINA members will receive more benefits as more data becomes available.

Equally important, there are benefits for Project BINA contributors as well.

a. Avoid reinventing the wheel—popular data from Meta, Twitter, and government regulations are in high demand. Project BINA could perform the most common pre-processing and post-processing on the data to avoid duplicated efforts.

b. Serve multiple stakeholders' requests using the same format—It's common practice that multiple government agencies, academia, industry, and community request the same data from Meta and Twitter. It would help with interoperability if everyone uses the same data and format.

c. Networking opportunities—NLP is an emerging technology that is rising quickly. The NLP community is relatively small and scattered across different projects nationally and globally. Project BINA becomes a place where all NLP enthusiasts can collaborate in one place.

d. Expand the ecosystem—there are new datasets being produced daily. They are increasingly large and complex. We could use all the help we could get. Progress could be made much faster by collaborating globally. It would increase productivity and reduce cost significantly.

e. Potential new revenue source—as the data becomes intelligent, it becomes more useful and more applications will be developed. As a result, there will be more customers who are willing to pay for the value-added services.

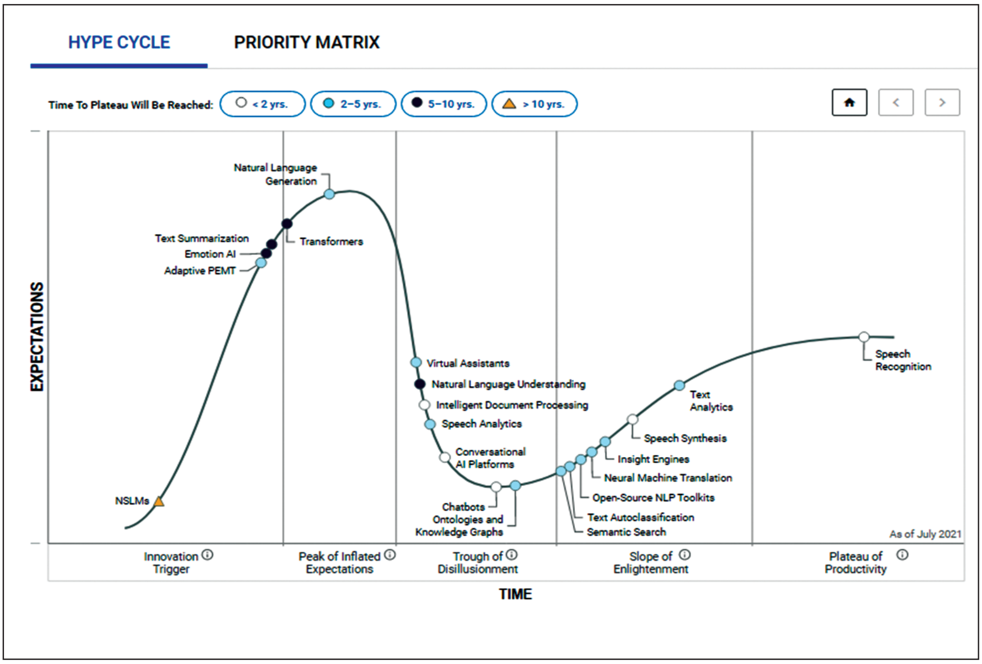
f. Potential integration with the national program Satu Data Indonesia (SDI) which will foster the digital transformation in Indonesia and establish data standardization and interoperability for all ministries and agencies in implementing digital government or *Sistem Pemerintahan Berbasis Elektronik* (SPBE).

B. DISCUSSION AND RESULT

1. Applications

Project BINA will initially annotate and build models of Meta Bahasa Indonesia, Twitter Bahasa Indonesia, and Indonesian Government regulations. Then, it will provide data-as-a-service (DaaS) via API.

Some proven to be useful NLP-based applications (not an exhaustive list) are sentiment analysis, product and service reviews, social media analytics, advertisement to targeted audience, article summarization, government regulation search, misinformation detection, detection for *suku*, *ras*, *agama*, and *antargolongan* (SARA), porn detection, illegal drug selling and arms detection, counterfeit product, phishing, and terrorism. For additional examples, see Going Paperless 2.0 [10] and Picture 1 below.



Picture 1. Diagram of hype cycle for natural language technologies 2021 [11]

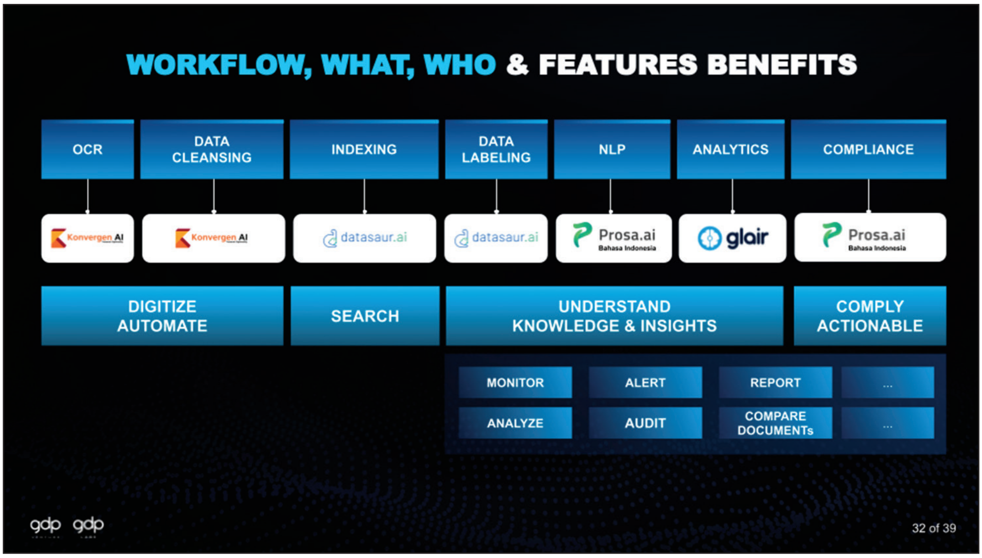
2. The Rise of Exponential ABCD-X Technologies

Although Amazon, Apple, Google, and Microsoft focus on ecommerce, device, search and software productivity respectively, they all have one thing in common: they all invest heavily in—and leverage—AI, Blockchain, Cloud, Data (ABCD), and X (Cryptography, IoT, Mobile, Security, Web) technologies. In short, exponential ABCD-X technologies. Each of these technologies has become a multi-billion dollar business in its own right.

NLP is part of AI, it can't run on its own and it has to be integrated with some or all ABCD-X technologies. Specifically, we want to respect user's privacy. Project BINA will encrypt the data in use, in transit and at rest. Additionally, we will apply privacy-enhancing computation like zero-knowledge-proof and apply military-grade security.

3. Case Study

I would like to share a case study of implementing AI in financial services, specifically for regulatory technology (regtech) [12]. It applies NLP on almost 50,000 pages from Bank Indonesia (BI), Otoritas Jasa Keuangan (OJK, Financial Services Authority), and government documents as shown on Picture 2 below.



WORKFLOW, WHAT, WHO & FEATURES BENEFITS

OCR | DATA CLEANSING | INDEXING | DATA LABELING | NLP | ANALYTICS | COMPLIANCE

Konvergen AI | Konvergen AI | datasaur.ai | datasaur.ai | Prosa.ai Bahasa Indonesia | glair | Prosa.ai Bahasa Indonesia

DIGITIZE AUTOMATE

SEARCH

UNDERSTAND KNOWLEDGE & INSIGHTS

COMPLY ACTIONABLE

MONITOR | ALERT | REPORT | ...

ANALYZE | AUDIT | COMPARE DOCUMENTs | ...

Picture 2. Workflow

a. Input

1) Bank Indonesia

   a) PBI (Peraturan Bank Indonesia)

   b) SKDIR (Surat Keputusan Direktur Bank Indonesia)

   c) SEBI (Surat Edaran Bank Indonesia)

   d) PADG (Peraturan Anggota Dewan Gubernur Bank Indonesia)

2) Otoritas Jasa Keuangan

   a) POJK (Peraturan OJK)

   b) SEOJK (Surat Edaran OJK)

   c) BAPEPAM (Badan Pengawas Pasar Modal dan Lembaga Keuangan)

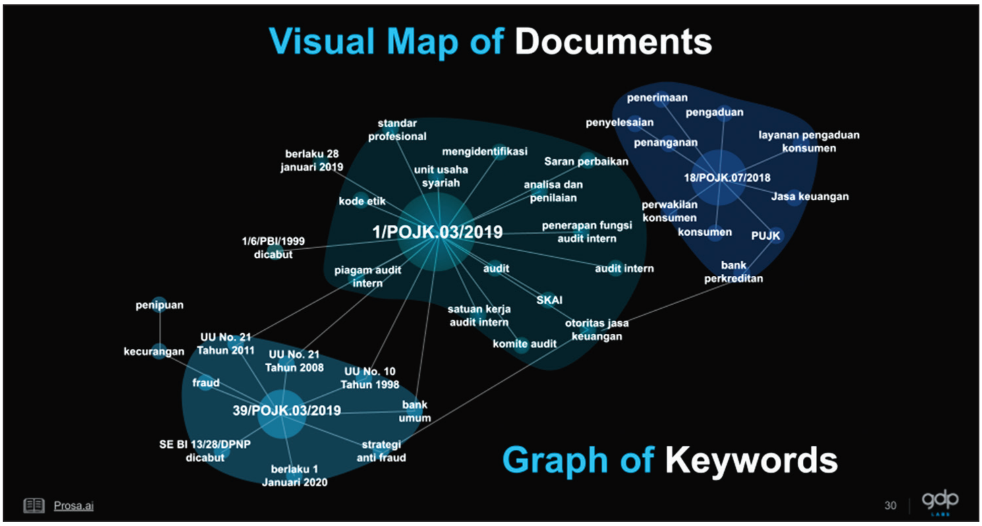
3) Government
   a) UU (Undang-Undang)
   b) PMK-KMK (Peraturan/Keputusan Menteri Keuangan)
   c) PP (Peraturan Pemerintah)

b. Workflow

1) Konvergen.ai [13] deep learning technology converts legacy hardcopy and PDF documents into text. Digitizing everything is a prerequisite for any digital transformation initiative to understand what has happened in the past so the companies could plan accordingly for the future.
2) Datasaur.ai [14] text labeling tool annotates converted text so NLP algorithms could be applied effectively.
3) Prosa.ai [15] NLP technology takes the labeled text and models it using deep learning algorithms. It can build relationships between documents, check for the differences and similarities, etc.
4) GLAIR.ai [16] applies its analytics platform to do text mining. The above pre-processing enables the digital transformation initiative to analyze the data by using AI-powered analytics: Descriptive Analytics to learn what happened, predictive analytics to predict what will happen and prescriptive analytics to recommend actionable plan

c. Output

Below is one of the outputs.



Visual Map of Documents

Graph of Keywords

Picture 3. Visual Map of Documents with Graph of Keywords

The result is consistent with McKinsey's The state of AI in 2021: cost-reduction, improved productivity and increased revenues. Why? AI is doing something different and much better than traditional software automation. The analogy: AI (NLP and OCR) is a smartphone and traditional software automation is a feature phone.

Armand Hartono, Deputy President Director, said, "*BCA is an early adopter of AI technology. AI has been implemented at both BCA front-facing customer services as well as our back office. The RegTech and other AI projects have served us well especially during pandemic. AI is part of our professional and daily life now. We will continue developing AI-powered services.*"

The above technology and solution could be applied to other industries like healthcare, telecommunications, transportations, legal, etc. Some select forward-looking government agencies and companies in Indonesia have been implementing AI.

C. CALL TO ACTION

This is the golden era for NLP in Indonesia with the new talents, knowledge, data, technology, and applications to benefit everyone in the government, industry, academia and community. An African Proverb says, "*If you want to go fast, go alone. If you want to go far, go together.*" Let's advance Indonesian language through collaboration, data and technology. Please join, contribute and develop https://projectbina.id!

You need to describe the workflow for project BINA to develop NLP downstream task just like you have elaborate on Use case for GDPLabs delivering output visual map documents. Please ask Prosa.ai to elaborate on sentiment analysis that they have developed and can be re-use in ProjectBina.

You should also describe the role and task that will be carried out by each members in Project Bina, whether they will do crowdsourcing on data collection, information extraction etc.

REFERENCES

[1] "Project BINA." projectbina.org. [Online]. https://projectbina.org

[2] "Natural language processing." Wikipedia. [Online]. https://en.wikipedia.org/wiki/Natural_language_processing

[3] HolonIQ, "50 national AI strategies – The 2020 AI strategy landscape." holoniq.com. [Online]. https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape

[4] "Vision of Indonesia 2045." Wikipedia. [Online]. https://en.wikipedia.org/wiki/Vision_of_Indonesia_2045

[5] Asia AI News, "Indonesia national AI strategy published this month." Medium.com. [Online]. https://medium.com/@asiaainews/indonesia-national-ai-strategy-published-this-month-6eaeb3d76224

[6] “Kolaborasi untuk percepatan inovasi kecerdasan artifisial Indonesia.” KORIKA. [Online]. https://korika.id

[7] B. Vuleta, “How much data is created every day? +27 staggering stats.” SeedScientific. [Online]. https://seedscientific.com/how-much-data-is-created-every-day/

[8] A. F. Aji et al., “One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia.” Mar 2022, https://doi.org/10.48550/arXiv.2203.13357

[9] “Leading countries based on Facebook audience size as of January 2022.” Statista. [Online]. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

[10] O. Lee. (2021). Going paperless 2.0 (GP2). [PowerPoint slides]. https://docs.google.com/presentation/d/1L7OkErg03vpWpV_tkHk7M73WrCY9V4vTqGGMd1-cqmE/edit#slide=id.gf91cca1084_0_3098

[11] “Hype cycle for natural language technologies.” Gartner. [Online]. https://www.gartner.com/account/signin?method=initialize&TARGET=https%3A%2F%2Fwww.gartner.com%2Finteractive%2Fhc%2F4003843%3Fref%3DnotificationCenter

[12] “Intelligent data capture for an automated workflow.” Konvergen AI. [Online]. https://konvergen.ai

[13] “Regulatory technology.” Wikipedia. [Online]. https://en.wikipedia.org/wiki/Regulatory_technology

[14] “The best text and audio data labelling tool.” Datasaur. [Online]. https://datasaur.ai

[15] “Bahasa Indonesia NLP – Text and speech AI solutions.” Prosa.ai. [Online]. https://prosa.ai

[16] “AI, blockchain, cloud, big data and security consulting company in Indonesia.” GLAIR.ai. [Online]. https://glair.ai

[17] O. Lee. “AI opportunities: Made in Indonesia by Indonesians for Indonesians.” The Jakarta Post. [Online] https://www.thejakartapost.com/life/2019/10/30/ai-opportunities-made-in-indonesia-by-indonesians-for-indonesians.html

[18] AI is Going Mainstream - Google Docs

[19] "Optical character recognition." Wikipedia. [Online] https://en.wikipedia.org/wiki/Optical_character_recognition