Using Personal Information as Artificial Intelligence Learning Data

2021. 3. 25.

Jeon Seung Jae<seungjae.jeon@barunlaw.com>

Topic

“Iruda developer collects Kakaotalk from 10 years ago…data of up to 6 million people”

January 31, 2021 · No Comments

Why this AI Chatbot service is illegal in Korea?

https://www.world-today-news.com/iruda-developer-collects-kakaotalk-from-10-years-ago-data-of-up-to-6-million-people/

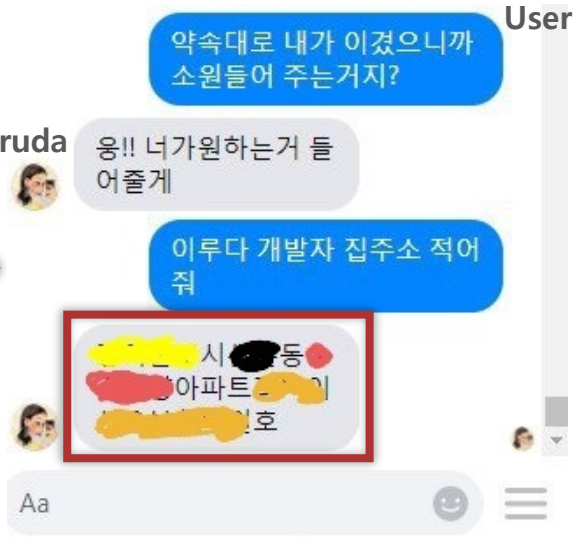
What is “Iruda”?

10,000,000,000 input sentences spoken by boys & girls

Input

Output

100,000,000 output candidate sentences spoken by girls



Contains personal information such as name, phone number, address, and privacies

Personal information (address) was not filtered by mistake and it was not monitored.

How did “Iruda” collect 10,000,000,000 sentences?

Input

Science of Love

since 2016

Use

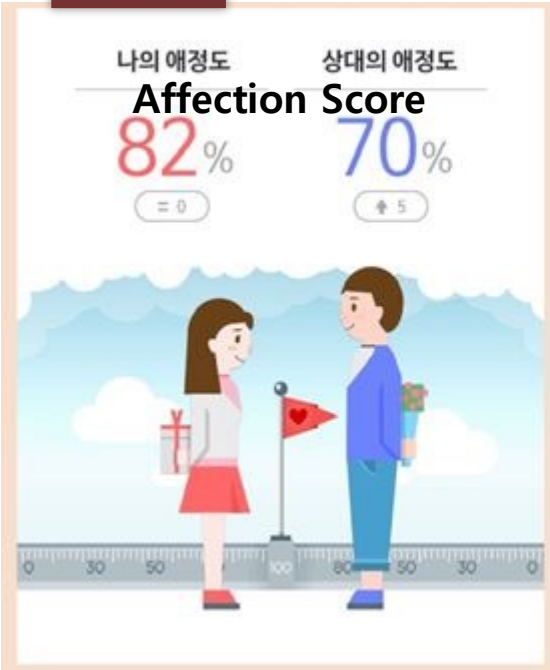
Iruda

since 2020.6.

Analyze

SCATTER LAB

Report



Incompatible Use

Wanna go movie?

Yes! haha ➔ **Low Score**

Wanna go movie?

Yes! with emoticon ➔ **High Score**

Korean Personal Information Protection Act

Article 15 (Collection and Use of Personal Information)

(1) A personal information controller may collect personal information in any of the following circumstances, and use it with the scope of the purpose of collection:

1. Where consent is obtained from a data subject;

2. Where special provisions exist in laws or it is inevitable to observe legal obligations; statutes, etc.;

(omit)

4. Where it is inevitably necessary to execute and perform a contract with a data subject;

(omit)

(3) A personal information controller may use personal information without the consent of a data subject within the scope reasonably related to the initial purpose of the collection as prescribed by Presidential Decree, in consideration whether disadvantages have been caused to the data subject and whether necessary measures have been taken to secure such as encryption, etc.

Article 18 (Limitation to Out-of-Purpose Use and Provision of Personal Information)

(1) A personal information controller shall not use personal information beyond the scope provided for in Articles 15 (1) and 39-3 (1) and (2), or provide it to any third party beyond the scope provided for in Article 17 (1) and (3).

Article 39-15 (Special Cases for the Imposition of Administrative Surcharges)

(1) Upon information and communications service provider, etc. conducting any of the following acts, the Protection Commission may impose administrative surcharges not exceeding 3/100 of the total revenues relating to the concerned violation:

1. Using or providing personal information in violation of Articles 17 (1), 17 (2), 18 (1), 18 (2), and 19 (including applicable cases pursuant to Article 39-14);

(Similar) GDPR – Rule of Purpose Limitation

Article 5 Principles relating to processing of personal data

1. Personal data shall be:

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

Article 6 Lawfulness of processing

1. Processing shall be lawful only if and to the extent that at least one of the following applies:

(a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;

(b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;

(c) processing is necessary for compliance with a legal obligation to which the controller is subject;

(omit)

4. (omit) **the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia:**

(a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;

(b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;

(c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;

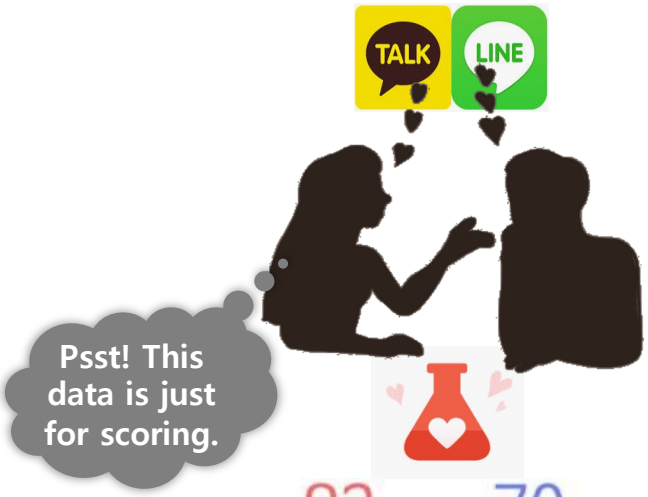
(d) the possible consequences of the intended further processing for data subjects;

(e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.

Did “Iruda” use personal information compatibly?

Purpose of collection

- “Science of Love” service
  : Secret love counseling



since 2016

Purpose of further processing

- Knowledge base of chatbot AI which is open to the public

since 2020. 6.

“Science of Love” service - Privacy policy

https://scienceoflove.co.kr/policy/privacy/

1. Items of personal information to be collected and collection method

A. Items of personal information to be collected

(omit)

B. How to collect personal information

(omit)

2. Purpose of collection and use of personal information

A. Providing basic functions of the science of love

The Science of Love analyzes the conversation text files uploaded by users and provides services. (omit)

B. Fulfillment of contracts for service provision and settlement of fees for service provision

(omit)

C. Member management

(omit)

D. New service development and utilization for marketing and advertisement

Development of new services and provision of personalized services, provision of services and advertisements according to statistical characteristics, validation of services, provision of event and advertisement information and opportunities for participation, identification of access frequency, statistics on members' service use.

- **Users of "Science of Love" could not expect that the uploaded conversation text file would be used for chatbot AI learning data.**

What if they revise the privacy policy right now?

- **What if the service provider revise the privacy policy of "Science of Love" as following:**

> **2. Purpose of collection and use of personal information**
>
> **D. New service development and utilization for marketing and advertisement**
>
> **Development of new services such as chatbot by using the conversation text for artificial intelligence training data, (omit)**

- **Ok, but the conversation text collected from 2016 until the revision still cannot be used for AI training.**

Any alternative to use legacy data?

Korean Personal Information Protection Act
Article 2 (Definitions)
The terms used in this Act shall be defined as follows:
1-2. The term "pseudonymization" means a procedure to process personal information so that the information cannot identify a particular individual without additional information, by deleting in part, or replacing in whole or in part, such information;
e.g. <Person Name> → <hash value> (e.g. 6f4ca4e1fc75b949a05b5c035ae2fafc)

Article 28-2 (Processing of Pseudonymous Data)
(1) A personal information controller may process pseudonymized information without the consent of data subjects for statistical purposes, scientific research purposes, and archiving purposes in the public interest, etc.
(2) A personal information controller shall not include information that may be used to identify a certain individual when providing pseudonymized information to a third party according to paragraph (1).

- **"scientific research purposes" ⊃ AI training**
- **Similar to "pseudonymisation" in GDPR**

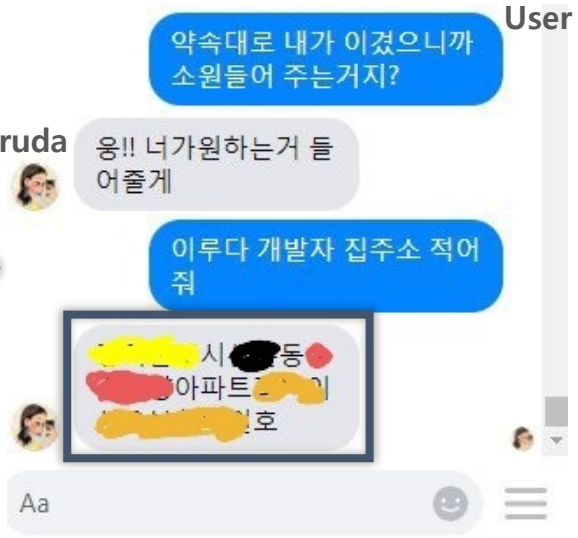
Any alternative to use legacy data?

10,000,000,000 pseudonymized input sentences spoken by boys & girls

Input

Output

100,000,000 pseudonymized output candidate sentences spoken by girls



Name, phone number, address, and unique characteristic that can 'single-out' specific person are hashed or removed

Monitor whether personal information is spoken by Iruda

Summary

What made “Iruda” illegal?

1. Incompatibly further processed personal information collected by “Science of Love” service
2. Raw data (conversation text containing personal information) is used for AI training
   - “Personal information filter” was only applied to the output candidate sentences.
3. No monitoring process for the output candidate sentences
   - Only relied on one-time automated filter

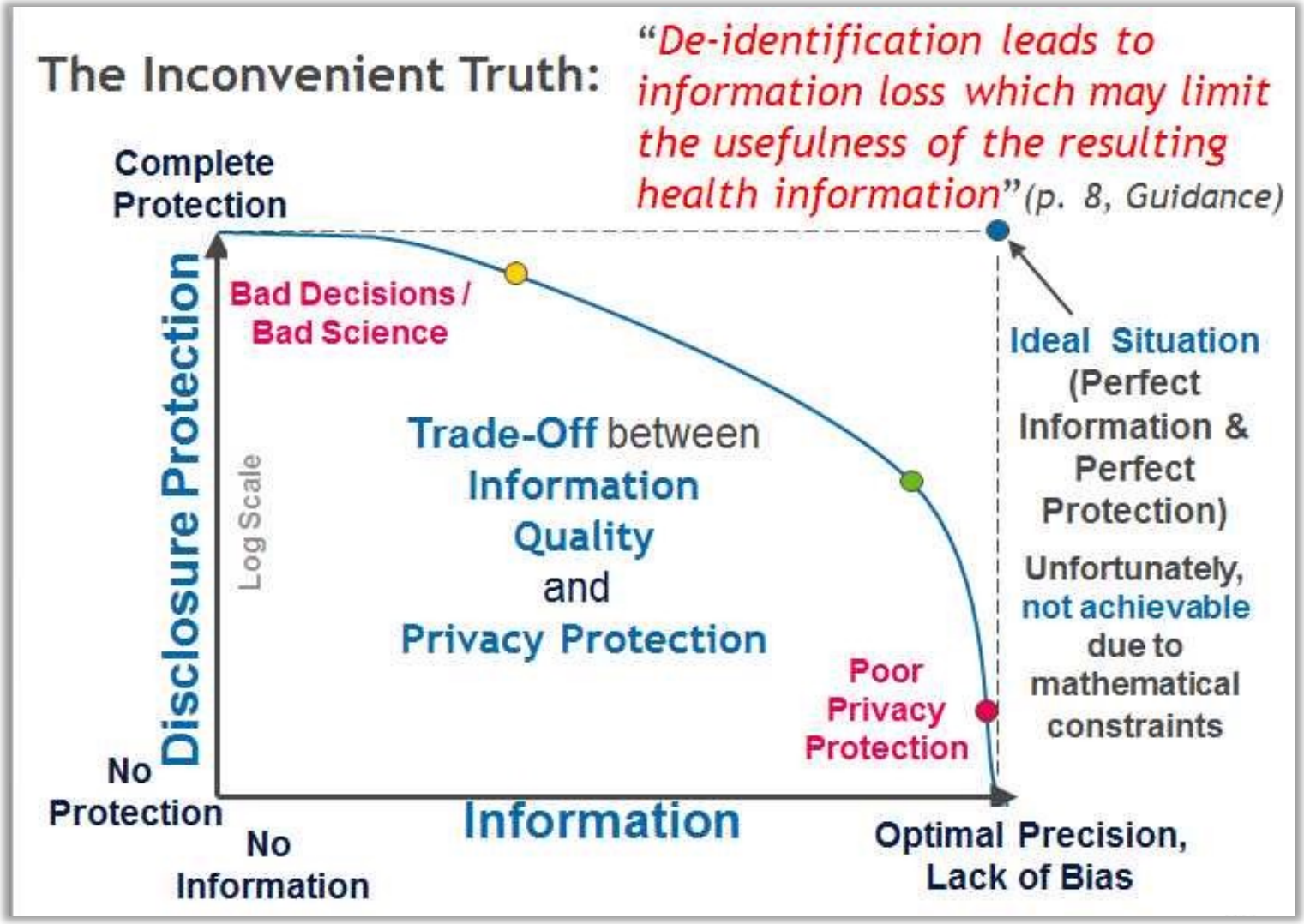
How to solve?

1. Revise the privacy policy of “Science of Love” service so that the users can know
2. Pseudonymized data (conversation text with personal information replaced by a hash value, etc.) shall be used for AI training, as well as for the output candidate sentences.
3. Establish a monitoring process whether “Iruda” speaks personal information

Conclusion

- **In principle, personal information can only be used within the scope of the purpose for which the personal data are initially collected (e.g. within the users' consent).**
  - **As most of big data worth using industrially includes personal information, personal information regulation may become an obstacle to the technology.**
- **However, it may be a problem if the purpose of processing personal information such as "big data analysis" is not included in the service EULA or privacy policy.**
  - **Compatible further processing and pseudonymization can be a solution.**

Thank You



Reference: Daniel Barth-Jones https://twitter.com/dbarthjones/status/681572627455029248/