# An Automatic Advisor to Detect Summarizable Chat Conversation in Online Instant Messaging

Fajri Koto

Dojo, 20 May 2015

# Outline

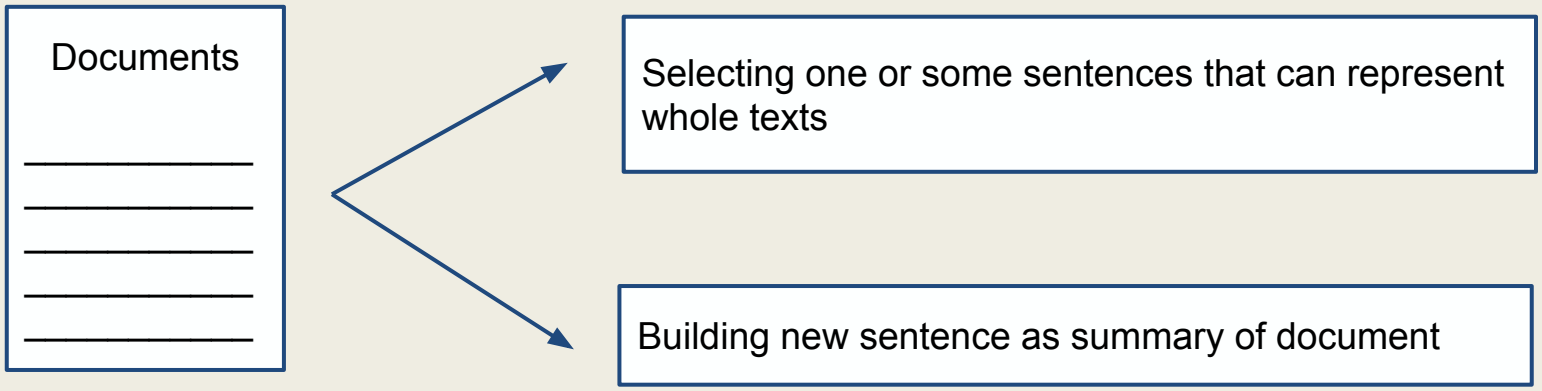
1. Introduction
2. Related Work
3. Machine Learning at glance
4. Data Construction
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work

# Outline

1. **Introduction** ✓
2. Related Work
3. Machine Learning at glance
4. Data Construction
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work

# What is text summarization?

## Definitions

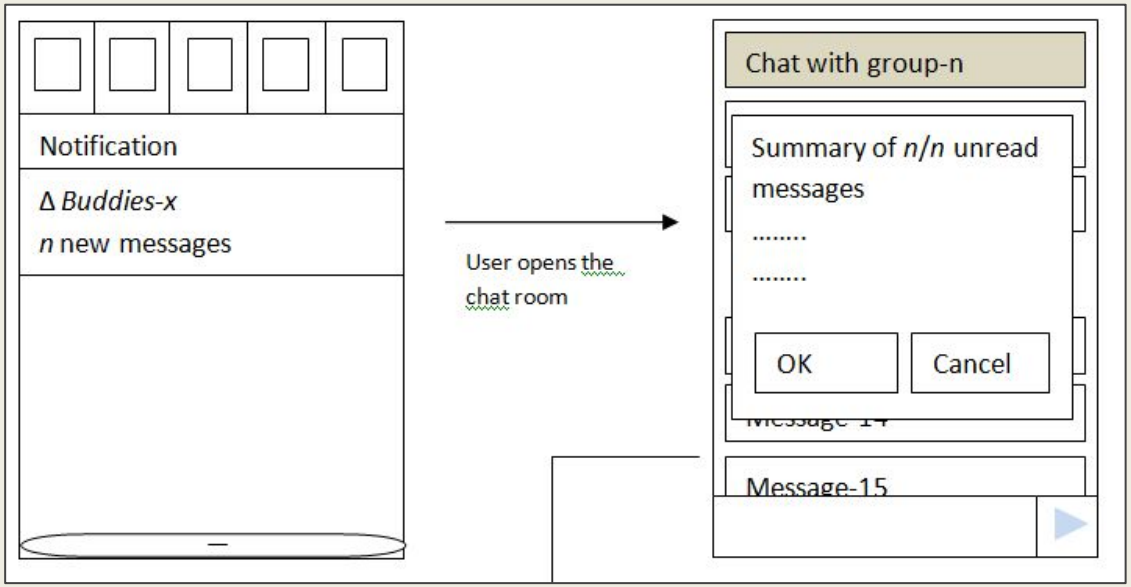



→ All summarization approaches work by directly applying bunch of messages without considering whether these messages have meaningful summary or not

→ The result is always provided

# Why chat summarization?

→ To ease user obtaining information quickly from various received messages.



# Summarizable chat definition

| Summarizable | Non-Summarizable |
|---|---|
| A: Hi guys, lets have holiday somewhere,, I need some fresh air<br>B: Where? I m free this weekend<br>A: Florida beach on Sunday?<br>C: Nice idea<br>A: Horray, lets do it<br>B: coooool idea | A: Hi dudes<br>B: Yo<br>A: How are you?<br>C: Tsup?<br>B: good<br>C: hm? what?<br>A: nothing |

Summarizable chat means that the document could produce a meaningful summary for human.

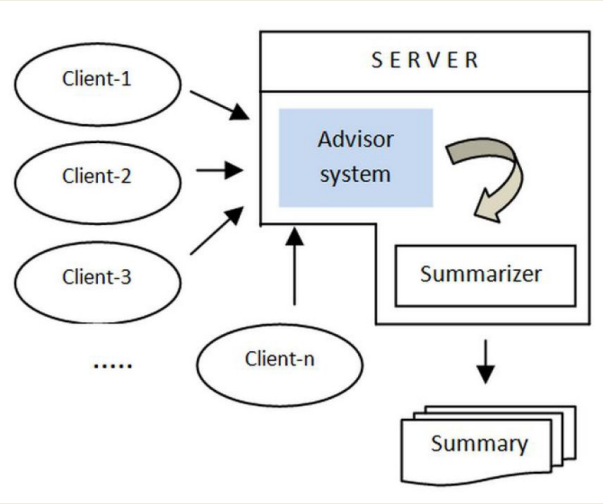
→ First conversation can be summarized as "*holiday to florida beach on Sunday*"

→ Second conversation is judged as non-summarizable conversation

# Why do we need summarizable chat detection?

*a. To optimize the summarization system*

→ Working on chat documents that contain many unstructured sentences in not trivial matter.
→ Summarization will take/seize big resources on its computing.




*b. To improve quality of summarization result*

Q: *How to detect this summarizable chat?*

# Outline

1. **Introduction**
2. **Related Work** ✓
3. Machine Learning at glance
4. Data Construction
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work

# Related works

→ There is only few numbers of works that have been published

- Uthus and Aha, 2011
  → It is caused by the difficulty in performing analysis of chat summarization: unstructured sentences, and the difficulty to obtain dataset

- Zhou and Hovy, 2005
  → worked on chat summarization by summarizing chat logs in order to create summaries comparable to the human made

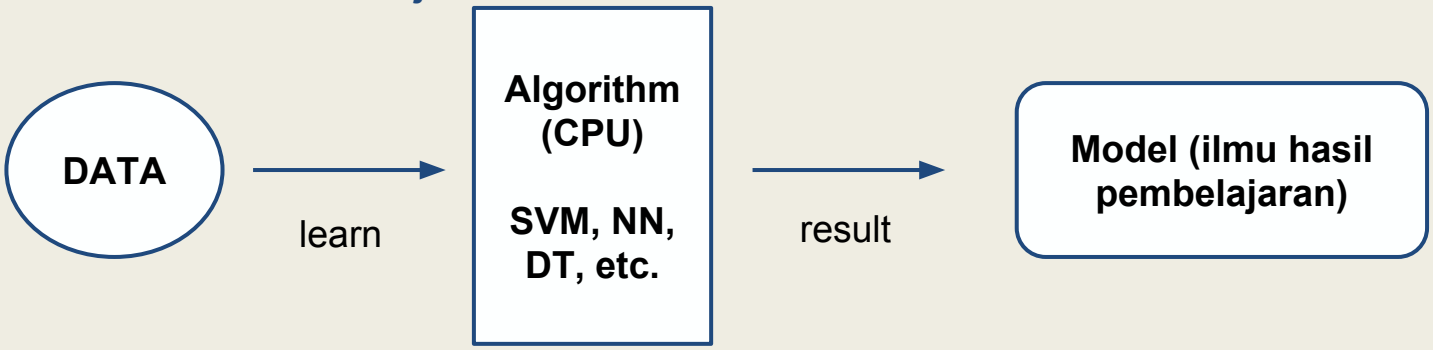
→ Our work is the first.
→ it is quite simple by applying machine learning.

# Outline

1. **Introduction**
2. **Related Work**
3. **Machine Learning at glance** ✓
4. Data Construction
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work

# Machine Learning at glance

→ Membuat mesin belajar.




**Contoh data:**

Atribut / Fitur

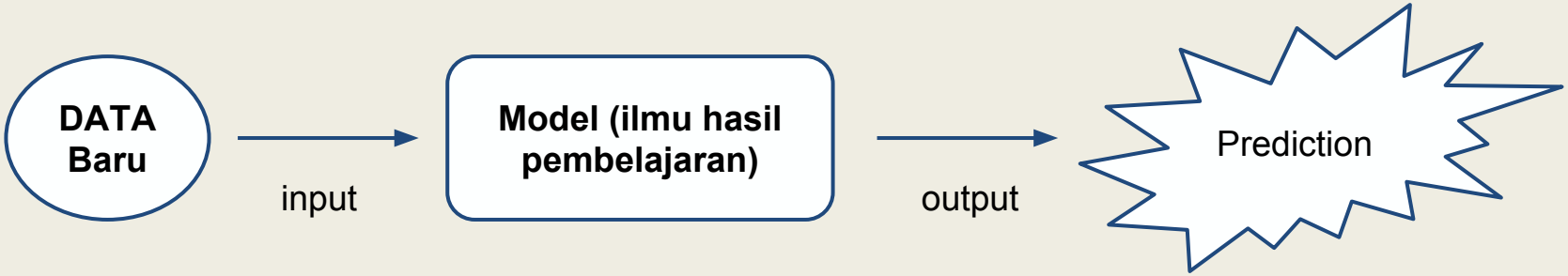
label / class

| Berat badan | Sel darah merah | Tekanan darah | Mutasi sel | ..... | Kanker otak |
|---|---|---|---|---|---|
| 50 | 12.000 | 150 | 55.000 | ..... | Yes |
| 45 | 15.000 | 170 | 26.000 | ..... | Yes |
| 65 | 18.000 | 135 | 55.000 | ..... | No |
| 72 | 20.000 | 152 | 30.000 | ..... | No |

We call it as Classification

# Machine Learning at glance

→ Menggunakan model (ilmu hasil pembelajaran).




**Contoh data baru:**

| Berat badan | Sel darah merah | Tekanan darah | Mutasi sel | ..... | Kanker otak |
|---|---|---|---|---|---|
| 70 | 12.000 | 150 | 55.000 | ..... | ?? |
| 25 | 15.000 | 170 | 26.000 | ..... | ?? |

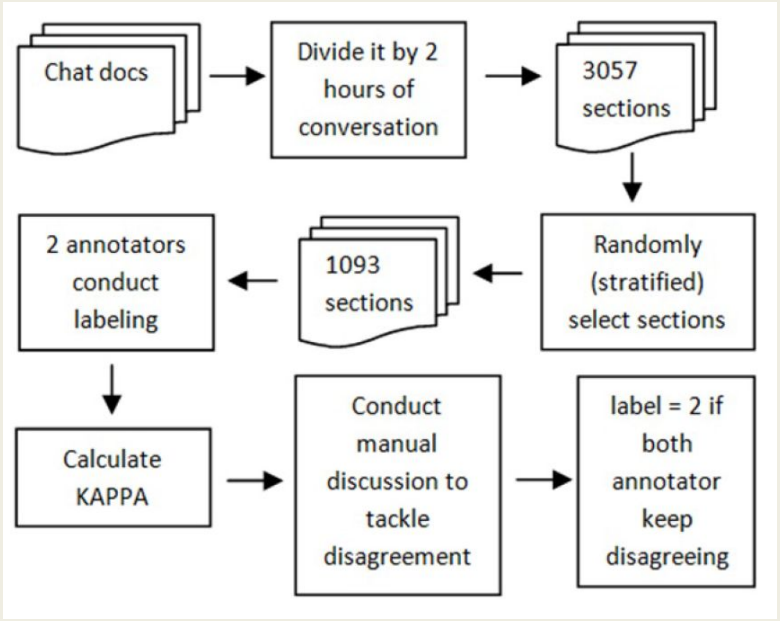
→ ini disebut sebagai **testing stage (tahap pengujian)**
→ Dari sini bisa diketahui akurasi sistem.

# Outline

1. **Introduction**
2. **Related Work**
3. **Machine Learning at glance**
4. **Data Construction** ✓
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work

# Data construction

The dataset was constructed by using seven WhatsApp groups chat in Bahasa Indonesia




Section 1
---
---
---
→ Summarizable

Section 2
---
---
---
→ Non-Summarizable

……

Section 1000
---
---
---
→ Summarizable

# Data construction

The dataset was constructed by using seven WhatsApp groups chat in Bahasa Indonesia

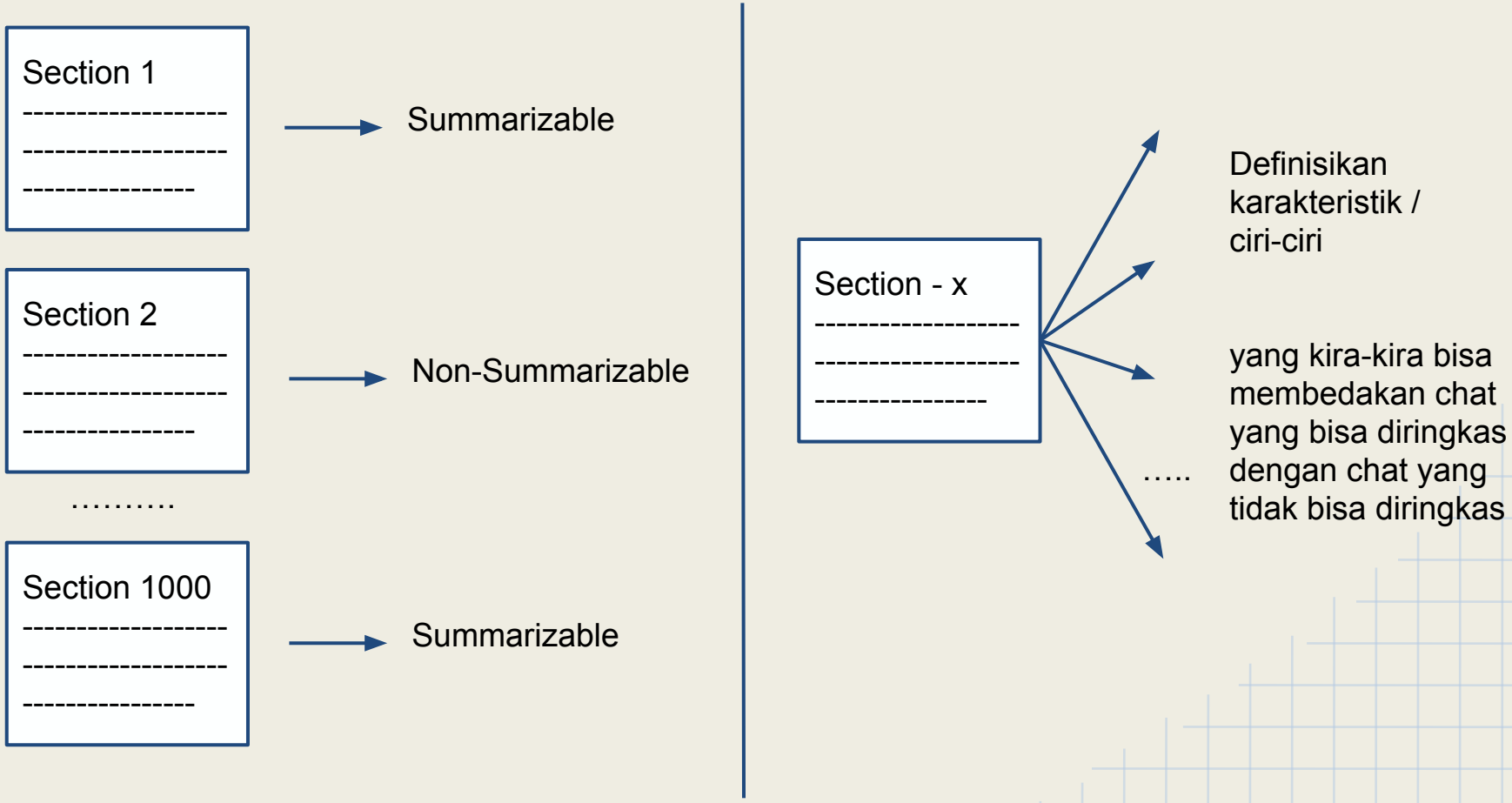
| Number of line | Total | #Labeling | Label = 1 | | Label = 2 | | Label = 3 |
|---|---|---|---|---|---|---|---|
| | | | Summarizable | | Non-Summarizable | | Disagreement |
| | | | Total | Sampling | Total | Sampling | |
| 1-10 | 1724 | 300 | 110 | 110 | 181 | 181 | 9 |
| 11-20 | 563 | 300 | 252 | 150 | 42 | 42 | 6 |
| 21-30 | 310 | 100 | 94 | 20 | 6 | 6 | 0 |
| 31-40 | 160 | 100 | 93 | 20 | 7 | 7 | 0 |
| 41-50 | 86 | 86 | 81 | 0 | 3 | 0 | 2 |
| >50 | 207 | 207 | 205 | 0 | 1 | 0 | 1 |
| Total | 3057 | 1093 | 853 | 300 | 240 | 236 | 18 |

# Outline

1. **Introduction**
2. **Related Work**
3. **Machine Learning at glance**
4. **Data Construction**
5. **Feature of Summarizable Chat Detection** ✓
6. Experiment Result
7. Conclusion and Future Work

# Feature for Classification

What we have at previous slide:



# Feature of Summarizable chat detection

→ In total we use 19 features, and grouped them into 3 sets of feature:

| Feature | Description |
|---|---|
| **Chat attribute** | |
| #user | Number of user participating in chat document |
| #chat | Number of chat section (line) in a document |
| ChatAvg | Average number of chat for each participant |
| **Lexical** | |
| #word | Number of words in a chat document |
| #unique | Number of unique words in a chat document |
| #non-stops | Number of non-stopword words in a chat document |
| #stops | Number of stopwords in a chat document |
| stopsRatio | Ratio between #stops and #word |
| nonStopsRatio | Ratio between #non-stops and #word |
| wordAvg | The average of non-stopword's frequency in a chat section |
| wordMax | Maximum value of non-stopword's frequency in a chat section |
| wordMin | Minimum value of non-stopword's frequency in a chat section |
| **RAKE** | |
| wordDegreeAvg | The average of words degree in the word co-occurrences graph |
| wordFreqAvg | The average of words frequency in the word co-occurrences graph |
| wordDegree | The ratio of wordDegreeAvg and wordFreqAvg |
| #keyword | number of extracted keywords generated by RAKE |
| keywordScoreAvg | The average of keyword score generated by RAKE |
| keywordScoreMax | Maximum value of keyword score generated by RAKE |
| keywordScoreMin | Minimum value of keyword score generated by RAKE |

*Hypothesis:*
Chat yang bisa diringkas cendrung memiliki **topik pembicaraan**

# Outline

1. **Introduction**
2. **Related Work**
3. **Machine Learning at glance**
4. **Data Construction**
5. **Feature of Summarizable Chat Detection**
6. **Experiment Result** ✔
7. Conclusion and Future Work

# Experiment result

| Classifier | All Feature | Feature selection | Selected feature |
|---|---|---|---|
| Linear Regression | 77.24% | 77.05% | #unique |
| Naive Bayes | 73.50% | 75.19% | stopsRatio, #keyword |
| Neural Network | 77.99% | 78.18% | #user, #keyword, keywordScoreMin |
| SVM | 76.88% | **78.36%** | #keyword, keywordScoreMin |

To perform classification, we use Rapid Miner tools, using 4 different classifier

→ There are two stages of experiment:

- Using all feature
- Performing feature selection → Mencari kombinasi fitur terbaik

# Outline

1. Introduction
2. Related Work
3. Machine Learning at glance
4. Data Construction
5. Feature of Summarizable Chat Detection
6. Experiment Result
7. Conclusion and Future Work ✓

# Conclusion

→ As the first study on summarizable chat detection, this study reveals that summarizability of chat document can be observed.

→ By employing three feature sets:

1) Chat attribute
2) Lexical
3) RAKE

We can distinguish summarizable chat by 78.36% as the highest accuracy performed by feature selection with SVM classifier