

# Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia

Yuliska<sup>1</sup>, Khairul Umam Syaliman<sup>2</sup>

Teknik Informatika, Politeknik Caltex Riau, Pekanbaru, Riau, Indonesia<sup>1,2</sup>

yuliska@pcr.ac.id<sup>1</sup>, khairul@pcr.ac.id<sup>2</sup>

---

## Article Info

### History :

Dikirim 01 Maret 2020

Direvisi 28 April 2020

Diterima 14 Juli 2020

---

### Kata Kunci :

Literatur Review

Peringkasan Teks

Text Summarization

Bahasa Indonesia

---

## Abstrak

Saat ini, kebutuhan akan mesin peringkasan dokumen teks menjadi semakin nyata karena semakin banyaknya informasi digital yang tersedia baik *online* maupun *offline*. Mesin peringkasan dokumen teks dibutuhkan agar pembacaan dan pencarian informasi menjadi lebih cepat. Review ini membahas metode, aplikasi, dataset dan Teknik evaluasi yang dapat diimplementasikan untuk riset di bidang peringkasan dokumen untuk teks berbahasa Indonesia. Review dilakukan terhadap berbagai teknik *text summarization* yang pernah digunakan pada penelitian-penelitian di bidang peringkasan teks berbahasa Indonesia, baik *unsupervised* maupun *supervised*, dataset yang dapat digunakan sebagai *baseline* dalam pengembangan sebuah metode dan *evaluation measure* yang tepat. Literatur Review ini juga akan menjelaskan sejauh apa perkembangan riset di bidang *text summarization* untuk dokumen berbahasa Indonesia.

© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

---

## Koresponden:

Yuliska

Program Studi Teknik Informatika, Jurusan Teknologi Informasi

Politeknik Caltex Riau (PCR)

Jalan Umban Sari No.1 Rumbai, Pekanbaru, 28265

Email : yuliska@pcr.ac.id

---

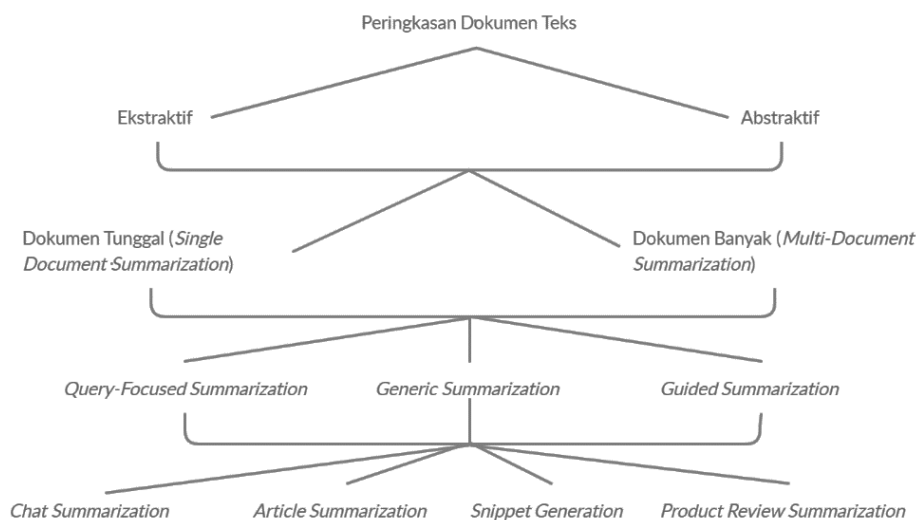
## 1. PENDAHULUAN

Peringkasan dokumen teks otomatis atau sering juga disebut sebagai *automatic text summarization* adalah sebuah cara untuk mengambil intisari informasi dari sebuah atau banyak dokumen teks. Peringkasan dokumen teks tersebut dapat dilakukan dengan 2 cara, yaitu peringkasan dokumen secara ekstraktif (*Extractive summarization*) dan peringkasan dokumen secara abstraktif (*Abstractive summarization*). Peringkasan secara ekstraktif dilakukan dengan cara mengambil beberapa kalimat penting dari dokumen teks asli yang mengandung informasi penting dari dokumen tersebut. Sedangkan peringkasan secara abstraktif dilakukan dengan cara membuat dan menyusun kalimat baru dimana kalimat-kalimat baru tersebut merupakan intisari informasi dari dokumen yang diringkaskan. Untuk jenis peringkasan dokumen teks, dapat dibagi menjadi peringkasan dokumen tunggal atau *single document summarization* [1, 2], peringkasan dokumen jamak atau *multi-document summarization* [3, 4], peringkasan dokumen berdasarkan *query* pengguna atau *query-based document summarization* [5, 6, 7], dan peringkasan dokumen teks yang berdasarkan pada fakta-fakta tertentu (5W+1H) atau *guided summarization* [8, 9].

Saat ini, kebutuhan akan mesin peringkasan dokumen teks menjadi semakin nyata karena semakin banyaknya informasi digital yang tersedia baik *online* maupun *offline*. Masih peringkasan dokumen teks dibutuhkan agar pembacaan dan pencarian informasi menjadi lebih cepat. Selain itu, peringkasan dokumen juga dapat diimplementasikan untuk mencegah adanya redundansi dan duplikasi di dalam sebuah dokumen teks. Implementasi mesin peringkasan dapat ditemukan di berbagai bidang, seperti peringkasan artikel berita [5, 10], generasi *snippet* atau teks pendek untuk hasil mesin pencarian [6, 11], generasi sinopsis buku [12], peringkasan email (*email summarization*) [13], peringkasan artikel ilmiah [14], generasi headline berita [15] dan generasi deskripsi singkat untuk promosi sebuah produk [16]. Di bidang *biomedical*, peringkasan dokumen juga diterapkan agar pencarian dan pembacaan dokumen medis menjadi lebih cepat dan efisien [17].

Peringkasan dokumen teks merupakan salah satu cabang ilmu dalam pemrosesan Bahasa alami manusia (*Natural Language Processing*). Berbagai Teknik atau metode telah dikembangkan untuk menghasilkan sebuah ringkasan yang padat dan informatif, mulai dari metode *unsupervised* [4, 6, 12, 18, 25] hingga metode *supervised* seperti *machine learning* dan *deep learning* [19-24]. Berbagai dataset atau corpus juga dibangun untuk kepentingan riset di bidang *text summarization*, seperti dataset *Document Understanding Conference (DUC)* dan *WikiHow* [26].

Sayangnya, berbagai metode, teknik dan dataset tersebut dikembangkan hanya untuk dokumen berbahasa Inggris. Pemrosesan Bahasa Indonesia dan Bahasa Inggris berbeda, terutama pada tahap *preprocessing*. Sedangkan untuk Teknik atau metode peringkasan dokumen teks berbahasa Indonesia terpaku pada metode-metode perangkangan tradisional, seperti *Maximal Marginal Relevance* [27] dan *Text Rank* [28]. Untuk dataset, sebuah dataset khusus untuk riset di bidang *text summarization* berbahasa Indonesia juga telah dikembangkan [29], namun dataset ini juga belum banyak digunakan. Dan untuk *evaluation measure* atau pengujian, riset *text summarization* berbahasa Indonesia lebih banyak menggunakan *recall* dan *precision* [4, 6, 12], dimana *evaluation measure* yang lebih tepat untuk *text summarization* adalah *ROUGE* [30].



Gambar 1. Peringkasan Dokumen Teks: Jenis dan Aplikasinya

Sebuah literatur review yang menjelaskan secara komprehensif tentang metode *text summarization* untuk dokumen teks berbahasa Inggris telah banyak dilakukan [31-32], namun review yang membahas secara khusus metode peringkasan dokumen teks berbahasa Indonesia belum pernah dilakukan. Untuk itu, review ini membahas metode, aplikasi, dataset dan Teknik evaluasi yang dapat diimplementasikan untuk riset di bidang peringkasan dokumen untuk teks berbahasa Indonesia. Review ini akan menjelaskan berbagai Teknik *text summarization*, baik *unsupervised* maupun *supervised*, dataset yang dapat digunakan sebagai *baseline* dalam pengembangan sebuah metode dan *evaluation measure* yang tepat. Review ini juga akan

menjelaskan sejauh apa perkembangan riset di bidang *text summarization* untuk dokumen berbahasa Indonesia, sehingga diharapkan dapat menjadi referensi yang baik untuk riset *text summarization* berbahasa Indonesia selanjutnya.

## 2. PERINGKASAN DOKUMEN TEKS OTOMATIS

Pada bab ini dijelaskan tahapan yang dilalui dalam meringkas dokumen secara otomatis. Mulai dari tahap *preprocessing*, pemilihan fitur, hingga ekstraksi kalimat atau kata dengan menggunakan beberapa metode.

### 2.1. Tahap *Preprocessing*

*Preprocessing* adalah salah satu tahapan penting dalam *text summarization*, di mana pada tahap ini teks diubah menjadi bentuk yang lebih mudah dicerna oleh komputer.

Tabel 1. Tahap *Preprocessing*

Tahap	Keterangan
<i>Tokenization</i>	Memecah kalimat menjadi kumpulan kata
<i>Stopword</i>	Penghapusan kata-kata yang tidak bermakna bagi dokumen, seperti “dan”, “yang”
<i>Stemming</i>	Penghapusan imbuhan, contoh “peringkasan” menjadi “ringkas”
<i>Case Folding</i>	Mengubah semua huruf kapital menjadi huruf kecil
<i>Sentence Splitting</i>	Memecah paragraf menjadi kumpulan kalimat

### 2.2. Ekstraksi Fitur (*Feature Extraction*)

Tahapan ini dilakukan setelah teks selesai di-*preprocessing*, yaitu tahap pemilihan atau ekstraksi fitur. Dalam peringkasan dokumen teks, fitur adalah kalimat atau kata yang dianggap penting dan memiliki probabilitas tinggi untuk dipilih sebagai ringkasan akhir.

#### 2.2.1. Fitur Tradisional

Fitur ini sering juga disebut sebagai *hand-engineered features*, karena ditentukan secara manual. Fitur ini digunakan pada metode-metode *unsupervised* dan metode-metode *machine learning*.

Tabel 2. Fitur Tradisional

Fitur	Keterangan
<i>Sentence Position</i>	Posisi kalimat dalam paragraph
<i>Sentence Length</i>	Jumlah kata dalam kalimat
<i>Keyword</i>	Jumlah <i>keyword</i> dalam kalimat
<i>Sentence Similarity to Title</i>	Tingkat kesamaan kalimat dengan judul dokumen
<i>Sentence Centrality</i>	Kesamaan suatu kalimat dengan kalimat lainnya
<i>Numerical Data</i>	Angka, tanggal, umur, alamat, <i>currency</i> pada kalimat
<i>Entity Name</i>	Sebuah nama benda, orang atau tempat khusus pada kalimat
<i>Double Quotes</i>	Tanda petik, biasanya merupakan tanda percakapan
<i>Cue Word</i>	Kalimat yang mengandung kata-kata seperti: “artikel ini menjelaskan”, “dapat disimpulkan”, atau “dengan demikian”
<i>Tf/Idf</i>	<i>Term Frequency/Inverse Document Frequency</i>

#### 2.2.2. *Bag of Words (BoW)*

*BoW* merupakan salah satu cara untuk merepresentasikan sebuah kalimat yang mendeskripsikan kemunculan kata tertentu dalam kalimat tersebut. Teknik *feature extraction* ini sering digunakan dalam riset *natural language processing* maupun *text mining* ketika metode yang digunakan adalah *machine learning*. Dalam pembentukan fitur menggunakan teknik *BoW*, terdapat 2 hal penting, yaitu kamus kata yang mengandung seluruh kata unik pada dokumen dan perhitungan frekuensi kemunculan kata tertentu [33].

### 2.2.3. One Hot Encoding

*One hot encoding* juga merupakan salah satu cara merepresentasikan sebuah kata atau kalimat di dalam sebuah dokumen teks. Pembentukan *One hot encoding* yaitu dengan cara memberikan sebuah integer unik pada setiap kata dan kemudian mengubah integer ke- $i$  menjadi sebuah *binary vector* (bernilai 0 atau 1) dengan jumlah dimensi sebesar  $K$  (jumlah kata pada kamus kata). Semua vektor yang terbentuk akan bernilai 0, kecuali kata dengan indeks  $i$ , yaitu 1 [34].

### 2.2.4. Fitur Word Embedding

Fitur ini digunakan untuk metode-metode *Deep Learning* [22, 35, 36]. Dibanding fitur tradisional yang memiliki dimensi terlalu kecil dan BoW serta *One Hot Encoding* yang dapat memiliki dimensi yang terlalu besar, *Word Embedding* mengungguli kedua jenis fitur ini. *Word Embedding* merepresentasikan satu kata dalam 50-300 *fixed* dimensi, dimana masing-masing vektor merupakan sebuah *dense vector*, sehingga kumpulan *vector* tersebut dapat merepresentasikan dengan baik hubungan antar kata, baik secara semantik maupun sintaksis. Saat ini, telah tersedia berbagai *word embedding* yang telah siap digunakan (*pre-trained word embedding*) dalam berbagai Bahasa, termasuk Bahasa Indonesia.

Tabel 3. Variasi *Pre-trained Word Embedding*

Nama	Keterangan
<i>Word2Vec</i> [37]	Pre-trained word embedding dengan dimensi 300. Dapat didownload pada link <a href="https://code.google.com/archive/p/word2vec/">https://code.google.com/archive/p/word2vec/</a>
<i>Glove</i> [38]	Pretrained word embedding dengan pilihan dimensi 50, 100, 200 dan 300. Dapat didownload pada link <a href="https://nlp.stanford.edu/projects/glove/">https://nlp.stanford.edu/projects/glove/</a>
<i>FastText</i> [39]	FastText hanya menyediakan word embedding dengan jumlah dimensi 300. Dapat didownload pada link <a href="https://fasttext.cc/docs/en/english-vectors.html">https://fasttext.cc/docs/en/english-vectors.html</a>
<i>BERT</i> [40]	BERT memiliki dimensi yang jauh lebih besar dibanding word embedding yang telah dijelaskan di atas, dapat didownload pada link: <a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a>

## 2.3. METODE PERINGKASAN DOKUMEN TEKS

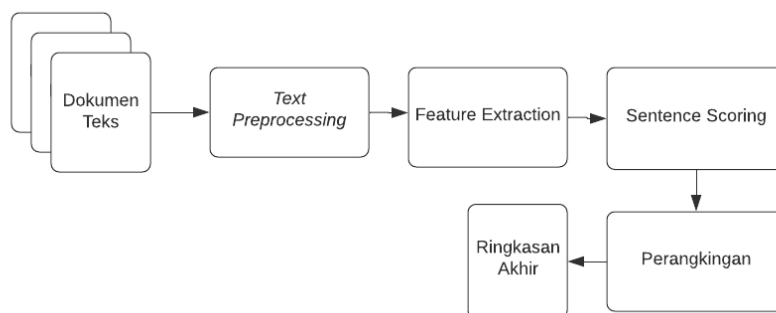
Sejak *Text summarization* dikenalkan oleh *NIST* pada tahun 2001, berbagai metode telah dikembangkan untuk berbagai aplikasi. Di sub bab ini, kami menjelaskan tentang metode-metode yang pernah digunakan untuk membuat sebuah mesin peringkasan dokumen teks berbahasa Indonesia. Selain itu, kami juga menjelaskan beberapa metode yang potensial yang belum pernah digunakan dalam meringkas dokumen berbahasa Indonesia.

### 2.3.1. Metode *Unsupervised*

Hampir semua metode yang bersifat *unsupervised*, dikembangkan untuk peringkasan teks yang bersifat ekstraktif, yaitu dengan cara memberikan skor kepada kalimat dan memilih kalimat dengan skor terbaik sebagai ringkasan akhir.

- **Maximal Marginal Relevance (MMR)** [27]: *MMR* pada awalnya dikembangkan untuk *query-focused summarization* [5, 7], yaitu dengan cara menghitung *similarity* antara kalimat dan kalimat, lalu kalimat dan kueri user. Namun, pada implementasinya *MMR* juga diterapkan pada peringkasan teks tanpa kueri [25]. *MMR* juga sering diterapkan dalam perangkaian untuk menentukan kalimat untuk ringkasan akhir. Jika kesamaan antara kalimat satu dan yang lain tinggi, maka dapat dipastikan terjadi redundansi.
- **Text Rank** [28]: *Text Rank* adalah teknik perangkaian kalimat berbasis graph. Pada *Text Rank*, setiap kalimat dianggap sebagai sebuah *vertex*. Konsepnya adalah semakin tinggi skor sebuah *vertex*, maka semakin penting *vertex* tersebut. Untuk dokumen berbahasa Indonesia, Eris, et.al [18] telah menerapkan *Text Rank* untuk meringkas dokumen berita.

- **Clustering Method** [4, 16, 41]: Despande, et.al [4] dan Cai, et.al [41] pada dasarnya menggunakan konsep yang sama dalam meringkas dokumen teks berbahasa Indonesia. Kedua studi ini menggolongkan kalimat-kalimat yang memiliki *similarity* yang tinggi dalam satu kelompok dengan menggunakan algoritma seperti *K-Means*. Lalu, kalimat-kalimat dengan skor tertinggi di dalam masing-masing kelompok diekstrak sebagai ringkasan akhir.



Gambar 2. Pendekatan Umum Peringkasan Teks dengan Metode *Unsupervised*

### 2.3.2. Metode Statistik

Metode statistik merupakan salah satu metode yang juga populer digunakan dalam peringkasan dokumen teks dan bersifat *unsupervised*, termasuk teks berbahasa Indonesia. Berikut beberapa metode statistik yang dapat diterapkan untuk peringkasan teks otomatis:

- **Non-Negative Matrix Factorization (NMF)** [42]: *NMF* digunakan dalam *extractive summarization* seperti yang dilakukan oleh Ridok [43], yaitu dengan cara memberikan bobot lebih pada kalimat-kalimat penting. Pemberian bobot dilakukan dengan cara perhitungan frekuensi term dalam kalimat, di mana term tersebut direpresentasikan sebagai matrix non-negative  $A$  (tidak negatif) berukuran  $s \times t$ , dengan  $s$  adalah jumlah term dalam kalimat dan  $t$  adalah jumlah kalimat dalam dokumen. Selanjutnya, perhitungan relevansi kalimat dilakukan untuk pemilihan kalimat hasil ringkasan akhir.
- **Log Likelihood Ratio (LLR)**: Metode statistik yang digunakan untuk mencari topik dokumen. Lalu, topik yang didapatkan tersebut digunakan untuk memberikan bobot kepada masing-masing kalimat. Akhmad, et.al [25], mengkombinasikan *LLR* dan *MMR* dalam studinya untuk peringkasan artikel berbahasa Indonesia.
- **Latent Dirichlet Allocation (LDA)**: Sama seperti *LLR*, *LDA* juga digunakan sebagai *topic modeling* atau ekstraksi topik tersembunyi dari sebuah dokumen. Hidayat, et. Al [45], menggunakan *LDA* untuk mengekstrak topik dari dokumen-dokumen yang ada di dalam datasetnya, lalu dengan topik tersebut dilakukan pengelompokan dokumen dengan *K-Means*, hingga akhirnya dilakukan pemilihan kalimat sebagai hasil ringkasan akhir. Lalu Silvia, et. al [44], mengkombinasikan *LDA* sebagai *topic modeling* dan algoritma genetika untuk memberikan bobot pada kalimat.

### 2.3.3. Metode Supervised

Metode *supervised* yang dibahas pada literatur review ini adalah metode-metode *Machine Learning* dan *Deep Learning*, karena kedua pendekatan ini yang paling sering digunakan oleh banyak studi di bidang *text summarization* saat ini. Untuk peringkasan teks secara ekstraktif, *Machine learning* dan *Deep Learning* menganggap peringkasan teks sebagai sebuah *classification problem* [22, 24], yaitu dengan memprediksi apakah sebuah kalimat “layak” untuk dijadikan sebagai ringkasan akhir atau tidak. Selain sebagai *classification problem*, peringkasan teks juga dapat diselesaikan secara regresi (*regression problem*) [46], namun untuk teks berbahasa Indonesia, pendekatan ini belum dikembangkan. Hal selanjutnya yang harus diperhatikan ketika melakukan peringkasan teks menggunakan metode *supervised* adalah proses pelabelan setiap kalimat untuk data latih (*training data*), apakah dilakukan secara manual (*manual/human labelling*) atau secara otomatis (*automatic labelling*) [22, 36, 48]. *Automatic labeling* dapat menjadi salah satu solusi,

walau solusi ini belum dapat dikatakan sebagai solusi terbaik [31]. Berikut beberapa metode *machine learning* yang dapat diterapkan untuk peringkasan teks otomatis:

- **Machine Learning**

Beberapa metode *machine learning* terbukti dapat menghasilkan ringkasan yang baik [24, 49]. Metode *machine learning* memiliki performa yang baik walau dengan data latih yang terbatas. Berikut beberapa metode yang dapat diterapkan untuk peringkasan dokumen berbahasa Indonesia:

- **Support Vector Machine (SVM):** *SVM* mengklasifikasi kalimat dengan mengoptimalkan sebuah garis pemisah (*hyperplane*). Sebagai contoh, kumpulan data latih di mana masing-masing data latih memiliki fitur  $x$  dan label  $y \in \{1, -1\}$ ,  $-1$  adalah kelas negatif (kalimat tidak penting) dan  $1$  adalah kelas positif (kalimat penting). Selama *training*, tujuan *SVM* adalah mengoptimalkan jarak antara garis-garis pemisah yang paralel dan membagi data latih ke dalam dua kelas ( $-1$  atau  $1$ ). Untuk dokumen berbahasa Indonesia, Somantri, et.al [24] telah menerapkan *SVM* untuk meringkas artikel berita.
- **K-Nearest Neighbour (KNN):** Ide dasar *KNN* adalah mengklasifikasi kalimat (data latih) ke dalam dua kategori, kalimat penting dan kalimat tidak penting. Hal ini dilakukan dengan cara menghitung jarak atau *similarity* antara data latih dan data uji menggunakan sebuah *similarity measure*, lalu mengurutkan nilai jarak mulai dari yang terkecil, kemudian memilih  $k$  data latih dengan jumlah  $k$  telah ditentukan sebelumnya. Terakhir, menentukan kategori data latih berdasarkan mayoritas  $k$  tetangga terdekat. Indrianto [50] telah menerapkan *KNN* untuk meringkas artikel berita berbahasa Indonesia dengan topik kesehatan.
- **Naïve Bayes:** Sebuah naïve bayes mengekstrak kalimat penting dengan cara mengklasifikasi kalimat ke dalam dua kategori, *keep* dan *reject* [49]. *Keep* artinya kalimat tersebut memiliki probabilitas tinggi untuk dijadikan sebagai ringkasan akhir, dan *reject* berarti kalimat tersebut tidak akan dipilih sebagai ringkasan akhir. Selama *training*, *Naïve Bayes* menghitung probabilitas kalimat-kalimat pada data latih, semakin tinggi probabilitas kalimat ( $0-1$ ), maka semakin tinggi kalimat tersebut untuk diberi label “*keep*” dan dipilih sebagai ringkasan akhir. Najibullah [49] membuat sebuah korpus yang terdiri dari 100 buah artikel berbahasa Indonesia dan menerapkan *Naïve Bayes* untuk melakukan peringkasan terhadap artikel-artikel tersebut.

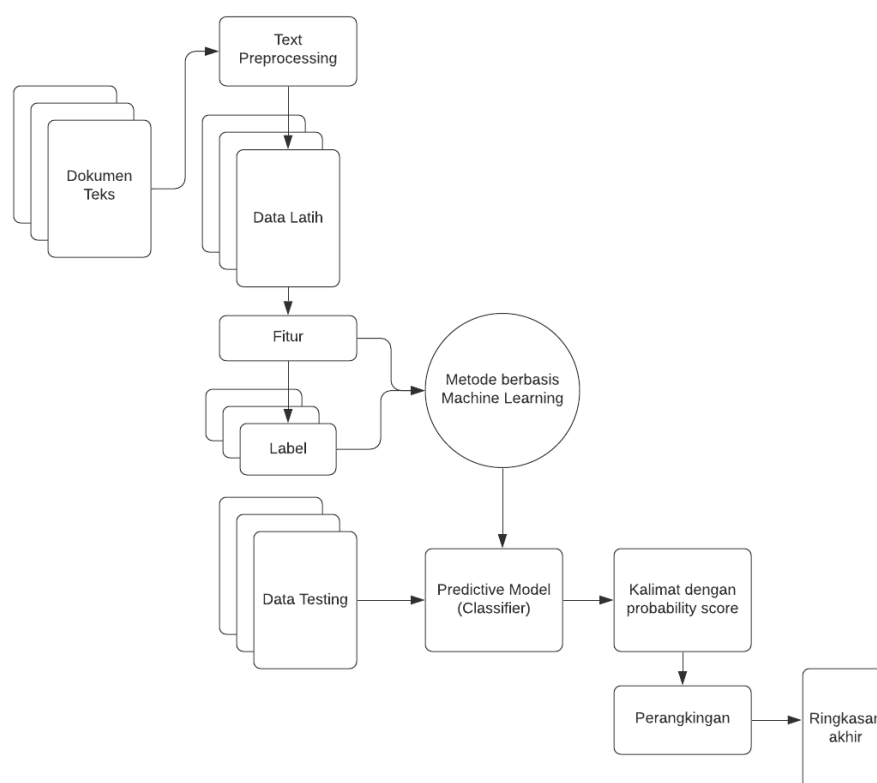
- **Deep Learning**

Studi peringkasan dokumen teks berbahasa Indonesia berbasis *deep learning* masih sangat terbatas. Sejauh ini, kami hanya menemukan dua studi, yaitu peringkasan teks menggunakan *Recurrent Neural Network (RNN)* [19] dan *Graph CNN* [20]. Perbedaan meringkas dokumen teks menggunakan *machine* dan *deep learning* terletak pada penggunaan fitur. *Machine learning* masih menggunakan *hand-engineered features*, sedangkan kebanyakan metode berbasis *deep learning* menggunakan fitur yang di-generate oleh mesin secara otomatis, seperti *word embedding* atau *one hot encoding*. Berikut beberapa *deep learning network* yang dapat diterapkan untuk peringkasan teks otomatis:

- **Convolutional Neural Network (CNN):** *CNN* pada awalnya diimplementasikan pada *Computer Vision* [34], namun Kim [51] membuktikan bahwa *CNN* juga dapat diterapkan pada *natural language processing*. Konsep *CNN* untuk dokumen teks adalah melakukan *convolution operation* untuk mengekstrak fitur baru dari kalimat atau kata pada dokumen. *Convolution operation* melibatkan pembentukan beberapa *window*, dimana sebuah *window* dapat memiliki  $h$  kata. Sebagai contoh, dari kalimat “peringkasan dokumen teks berbahasa indonesia” dapat dibentuk beberapa *window*, yaitu “peringkasan dokumen”, “dokumen teks”, “teks bahasa” dan “bahasa indonesia”, dengan  $h=2$ . Kemudian dilakukan *concatenation operation* terhadap *window-window* tersebut, sehingga didapatkan sebuah fitur baru yang lebih *representative*

untuk diteruskan ke *classifier layer*. Kim [51] menjelaskan proses ini secara detail pada studinya.

- **Recurrent Neural Network (RNN):** RNN memproses data sekuensial, seperti kalimat yang terdiri dari dua atau lebih kata. RNN memproses data sekuensial tersebut dengan melakukan iterasi ke setiap kata pada kalimat. Iterasi ke setiap kata disebut *state*, dimana *state* berisi informasi setiap kata yang sudah dilewati. RNN bersifat *forward*, dimana *internal loop* hanya dilakukan satu kali. Salah satu jenis RNN yang sering digunakan pada *text summarization* adalah *Long Short-Term Memory Network (LSTM)* [22].



Gambar 3. Peringkasan Teks dengan Metode *Supervised* (contoh: *Machine Learning*)

### 3. DATASET

Dataset adalah sekumpulan data yang dapat digunakan sebagai bahan percobaan riset. Beberapa studi mengumpulkan data mereka sendiri [49-50] sebagai bahan percobaan. Namun, pada bidang peringkasan dokumen teks otomatis berbahasa Indonesia, sejauh ini tersedia 2 dataset yang dapat digunakan, yaitu dataset yang berisi kumpulan *chat* dan IndoSum.

#### 3.1. Dataset Peringkasan Chat

Dataset ini dikembangkan oleh Koto [52], yang terdiri dari 300 sesi pembicaraan berbahasa Indonesia yang bersumber dari aplikasi *online chatting*, *WhatsApp*. Dataset ini juga memiliki 6 hasil ringkasan manusia yang dilakukan secara manual (*gold standard summaries*) untuk masing-masing sesi *chat*, 3 *gold standard summaries* bersifat ekstraktif, dan 3 *gold standard summaries* yang lainnya bersifat abstraktif. Masing-masing *gold standard summaries* merupakan ringkasan yang dibuat oleh manusia secara manual.

#### 3.2. Dataset Indosum

IndoSum bukan merupakan dataset pertama sebagai corpus untuk peringkasan dokumen teks otomatis berbahasa Indonesia, namun merupakan yang terbaru dan terbesar. Dataset yang dikembangkan oleh Kurniawan dan Louvan [29] ini terdiri dari 20K artikel berita. Artikel berita ini diambil dari portal berita berbahasa Indonesia, seperti CNN Indonesia dan Kumparan. Setiap

artikel memiliki judul, kategori dan 2 buah *gold standard summaries* yang dibuat secara manual. Di dalam dataset ini, terdapat 6 buah kategori, yaitu Hiburan/Entertainment, Inspirasi/Inspiration, Olahraga/Sport, Gosip/Showbiz, Berita Utama/Headline dan Teknologi/Tech.

#### 4. EVALUASI HASIL PERINGKASAN DOKUMEN

Evaluasi hasil peringkasan teks oleh mesin dilakukan untuk mengetahui kualitas hasil peringkasan. Pada bab ini, kami hanya membahas parameter *evaluation* bernama *ROUGE*, karena parameter inilah yang sangat banyak digunakan di bidang peringkasan dokumen teks.

##### 4.1. ROUGE

*Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) [30] adalah *evaluation metric* atau parameter evaluasi yang paling populer untuk evaluasi hasil peringkasan dokumen teks secara otomatis. *ROUGE* mengevaluasi hasil peringkasan dengan cara membandingkan hasil ringkasan oleh mesin dan hasil ringkasan manusia (*gold standard summary*). Ada beberapa macam jenis *ROUGE*, namun pada review ini, kami hanya akan membahas *ROUGE-N*, *ROUGE-L* dan *ROUGE-SU* karena *evaluation metrics* ini yang paling sering digunakan.

##### 4.1.1. ROUGE-N

*ROUGE-N* pada dasarnya adalah perhitungan *recall* yang berdasarkan pada perbandingan *n-gram* antara *gold standard summary* dan teks hasil peringkasan mesin. Jumlah *n-gram* yang digunakan beragam, yaitu  $n=1-4$ . Namun, yang paling sering digunakan adalah *n-gram* dengan jumlah  $n=1$  (*ROUGE-1*) dan  $n=2$  (*ROUGE-2*). Misal  $p$  adalah jumlah *n-gram* yang sama antara *gold standard summary* dan teks hasil peringkasan mesin, dan  $q$  adalah jumlah *n-gram* pada *gold standard summary*. Maka *ROUGE-N* dapat dihitung dengan rumus sebagai berikut:

$$ROUGE-N = \frac{p}{q} \quad (1)$$

##### 4.1.2. ROUGE-L

*ROUGE-L* mengevaluasi ringkasaan teks dengan cara membandingkan *longest common subsequence* (*LCS*) atau rangkaian kata terpanjang yang sama antara hasil ringkasan teks mesin dan *gold standard summary*. Misal  $m$  adalah jumlah kata pada *gold standard summary, maka *ROUGE-L* dapat dihitung dengan rumus sebagai berikut:*

$$ROUGE-L = \frac{LCS}{m} \quad (2)$$

##### 4.1.3. ROUGE-SU

*SU* pada *ROUGE-SU* merupakan *SKIP-Bigrams* dengan tambahan jumlah perhitungan *Unigram*. Sedangkan *SKIP-Bigrams* adalah pasangan kata (*bigrams*) dalam sebuah kalimat, dimana dalam perhitungannya dilewatkan sebanyak  $s$  kata, dimana  $s$  adalah jumlah kata yang dapat dilewati (*SKIP*) dalam pembentukan *SKIP-Bigram*. Untuk memahami pembentukan *SKIP-Bigrams*, silahkan pahami contoh sebagai berikut:

Kalimat: “Peringkasan dokumen teks otomatis”

*SKIP-Bigrams* dengan  $s=4$  yang terbentuk dari kalimat di atas adalah “peringkasan dokumen”, “peringkasan teks”, “peringkasan otomatis”, “dokumen teks”, “dokumen otomatis” dan “teks otomatis”.

Jadi, jika  $m$ -*SKIP* adalah jumlah *SKIP-Bigrams* yang sama antara hasil peringkasan teks mesin dan *gold standard summary*, dan  $U$  adalah jumlah *Unigram* yang sama antara hasil peringkasan teks mesin dan *gold standard summary*, maka *ROUGE-SU* dapat dihitung dengan rumus sebagai berikut:

$$ROUGE-SU = \frac{m - SKIP + U}{m} \quad (3)$$



## 5. APLIKASI PERINGKASAN DOKUMEN TEKS

Pada bab ini, kami hanya menjelaskan sebagian kecil dari aplikasi peringkasan dokumen teks yang kami anggap memiliki potensi besar untuk dikembangkan.

### 5.1. *Snippet Generation*

*Snippet* adalah deskripsi singkat mengenai sebuah halaman *website* pada halaman hasil mesin pencari. *Snippet* dianggap penting, karena semakin singkat dan informatif deskripsi sebuah *snippet*, maka semakin cepat pengguna menemukan informasi yang diinginkan. *Snippet generation* untuk mesin pencarian dokumen berbahasa Indonesia telah dilakukan oleh Saraswati, et.al [6] dengan menggunakan metode *MMR*, dimana untuk *men-generate snippet*, mereka menghitung kesamaan antara kalimat dan kalimat dan kesamaan antara kalimat dan kueri user.

### 5.2. *News Article Summarization*

Peringkasan artikel berita merupakan riset yang di bidang *text summarization* yang paling banyak dikembangkan untuk artikel berita berbahasa Indonesia [5, 8, 20]. Peringkasan artikel dapat berdasarkan pada fakta-fakta tertentu (*guided summarization*) [8], berdasarkan pada kueri user [5], atau hanya berdasarkan pemilihan informasi-informasi yang dianggap penting [25].

### 5.3. *Product Review Summarization*

*Product review summarization* adalah salah satu bentuk *multi-document summarization*, yaitu peringkasan review pengguna sebuah *online shop* terhadap produk tertentu [16]. Peringkasan review produk digunakan untuk mendapatkan informasi pendapat pengguna mengenai produk tertentu secara cepat dan tepat. Sejauh ini, belum ada studi pun yang membahas peringkasan review produk untuk *online shop* berbahasa Indonesia.

### 5.4. *Scientific Articles Summarization*

Peringkasan artikel ilmiah hampir sama dengan peringkasan dokumen teks biasa, yaitu proses transformasi artikel ilmiah yang menjadi satu dokumen teks pendek yang padat dan informatif. Hal ini telah dilakukan oleh Anggraini dan Wulandari [14] dengan memberikan bobot lebih pada kalimat yang dianggap penting dan informatif. Peringkasan artikel ilmiah juga dapat berupa *citation-based summarization*, yaitu dengan cara menemukan artikel-artikel lain yang merujuk artikel target, sehingga dapat ditemukan kalimat-kalimat penting untuk dapat diekstrak sebagai ringkasan artikel target. Hal ini telah dilakukan oleh beberapa peneliti [53-54], namun Peringkasan artikel ilmiah dengan metode ini belum dikembangkan untuk artikel ilmiah berbahasa Indonesia.

### 5.5. *Chat Summarization*

*Chat summarization* adalah peringkasan pembicaraan secara online yang menggunakan teks (*chat*). Namun, bahkan untuk teks berbahasa Inggris, penelitian ini masih sangat terbatas. *Chat summarization* dianggap lebih sulit karena bahasa yang digunakan ketika melakukan pembicaraan online berbeda dengan bentuk teks lainnya. *Online Chatting* cenderung lebih pendek, tidak terstruktur, mengandung singkatan, mengandung banyak kesalahan ejaan (*typo*) dan tidak terstruktur, sehingga sulit untuk menghasilkan ringkasan yang baik jika hanya menggunakan metode NLP biasa.

## 6. KESIMPULAN

Dengan berkembangnya teknologi, semakin banyak pula dokumen yang tersedia baik *online* maupun *offline*, sehingga kebutuhan akan mesin peringkasan teks otomatis menjadi sangat nyata agar proses pembacaan dan pencarian informasi menjadi lebih cepat. Pada literatur review ini, kami menekankan penjelasan pada peringkasan dokumen teks yang bersifat ekstraktif. Kami membahas berbagai metode yang sampai saat ini digunakan untuk peringkasan teks berbahasa Indonesia sampai dengan metode-metode yang belum pernah diimplementasikan, sehingga membuka lebar peluang riset. Sejauh ini, kami menemukan bahwa peringkasan dokumen teks secara otomatis didominasi oleh teknik yang bersifat ekstraktif. Peringkasan dokumen teks berbahasa Indonesia juga didominasi oleh metode-metode *unsupervised*, sementara metode-metode *supervised* seperti *machine learning* dan *deep learning* masih sangat jarang ditemukan. Review juga dilakukan

terhadap kesalahan yang umum terjadi dalam mengevaluasi hasil ringkasan dan memberikan parameter evaluasi yang lebih tepat untuk diimplementasikan, yaitu *ROUGE*. Namun, tidak semua metode dapat kami jelaskan melalui review ini, namun studi ini dapat memberikan pemahaman yang baik bagi perkembangan *text summarization* berbahasa Indonesia dan peluang risetnya.



## DAFTAR PUSTAKA

- [1] W. Yulita, S. Priyanta, and Azhari, "Automatic Text Summarization Based on Semantic Network and Corpus Statistics," *Indonesian Journal of Computing and Cybernetics Systems*, 2019.
- [2] P.P. Tardan, A. Erwin, K.I. Eng and W. Muliady, "Automatic Text Summarization Based on Semantic Analysis Approach for Documents in Indonesian Language," in *International Conference on Information Technology and Electrical Engineering*, 2013.
- [3] G. Yapinus, A. Erwin, M. Galinium and W. Muliady, "Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive- Extractive Summarization Technique," in *International Conference on Information Technology and Electrical Engineering*, 2014.
- [4] A. R. Deshpande and Lobo L. M. R. J., "Text Summarization using Clustering Technique," *International Journal of Engineering Trends and Technology*, 2013.
- [5] D. Annisa and M.L. Khodra, "Query-based Summarization for Indonesian News Article," in *International Conference on Advanced Informatics, Concepts, Theory, and Applications*, 2017.
- [6] N. F. Saraswati, I. Indriati and R.S. Perdana, "Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2018.
- [7] P. Bhaskar and S. Bandyopadhyay, "A Query Focused Multi Document Automatic Summarization," in *Pacific Asia Conference on Language, Information and Computation*, 2010, p.p 545-554.
- [8] D. T. Massandy and M.L. Khodra, "Guided Summarization for Indonesian News Articles," in *International Conference on Advanced Informatics, Concepts, Theory, and Applications*, 2014.
- [9] P. E. Gennest and G. Lapalme, "Fully Abstractive Approach to Guided Summarization," in *Annual Meeting of the Association for Computational Linguistics*, 2012, p.p. 354-358.
- [10] A. Ardianto, J. Praghantha and V. Christiani .M, "Perancangan Peringkasan Berita Otomatis Dengan Memperhatikan Sinonim Menggunakan Metode Weight of Feature," *Jurnal Ilmu Komputer dan Sistem Informasi*, 2013.
- [11] W.F. Chen, S. Syed, B. Stein, M. Hagen and M. Hattest, "Abstractive Snippet Generation," in *International World Wide Web Conference Committee*, 2020.
- [12] A. Indriani, "Maximum Marginal Relevance Untuk Peringkasan Teks Otomatis Sinopsis Buku Berbahasa Indonesia," in *Seminar Nasional Teknologi Informasi dan Multimedia*, 2014, p.p 29-34.
- [13] M. Alam and M. Kakkar, "Email Summarization-Extracting Content from the Email," *International Journal of Innovative Research in Computer and Communication Engineering*, 2015.
- [14] D. Anggraini and L. Wulandari, "Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia Menggunakan Teknik Ekstraktif dan Fitur Kalimat Untuk Dokumen Tunggal," in *Seminar Nasional Rekayasa Komputer dan Aplikasinya*, 2015, p.p. 126-130.
- [15] Ayana, Y.K. Lin and Z.Y. Liu, "Recent Advances on Neural Headline Generation," *Journal of Computer Science and Technology*, 2015.
- [16] Y. E. Ariska, W. Maharani and M.S. Mubarok, "Peringkasan Review Produk Berbasis Fitur

- Menggunakan Semantic Similarity Scoring dan Sentence Clustering,” in *e-Proceeding of Engineering*, 2016, p.p. 5323-5331.
- [17] D. K. Gaikwad and C.N. Mahender, “A Review Paper on Text Summarization,” *International Journal of Advanced Research in Computer and Communication Engineering*, 2016.
- [18] E. Eris, V. Christiani .M and C. Pragantha, “Penerapan Algoritma Textrank Untuk Automatic Summarization Pada Dokumen Berbahasa Indonesia,” *Jurnal Ilmu Teknik dan Komputer*, 2017.
- [19] R. Adelia, S. Suyanto and U.N. Wisesty, “Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit,” *International Conference on Computer Science and Computational Intelligence*, 2019.
- [20] G. Garmastewira and M.L. Khodra, “Summarizing Indonesian News Articles Using Graph Convolutional Network,” *Journal of ICT*, 2019.
- [21] C. Khatri, G. Singh and N. Parikh, “Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks,” in *International Conference on Knowledge Discovery & Data Mining*, 2018.
- [22] J. Cheng and M. Lapata, “Neural Summarization by Extracting Sentences and Words,” in *Annual Meeting of the Association for Computational Linguistics*, 2016, p.p. 484-494.
- [23] P.M. Sabuna and D.B. Setyohadi, “Summarizing Indonesian Text Automatically by Using Sentence Scoring and Decision Tree,” in *International Conferences on Information Technology, Information Systems and Electrical Engineering*, 2017.
- [24] P.G. Somantri, A. Komarudin and R. Ilyas, “Peringkasan Teks Otomatis Berita Berdasarkan Klasifikasi Kalimat Menggunakan Support Vector Machine,” in *Seminar Nasional Teknologi dan Informatika*, 2018.
- [25] I.N. Akhmad, A. S. Nugroho and B. Harjito, “Peringkasan Multidokumen Otomatis dengan Menggunakan Log-Likelihood Ratio (LLR) dan Maximal Marginal Relevance (MMR) untuk Artikel Bahasa Indonesia,” *Jurnal Linguistik Komputasional*, 2018.
- [26] M. Koupae dan W.Y. Wang, “WikiHow: A Large-Scale Text Summarization Dataset,” *arXiv*, 2018.
- [27] J. Carbonell dan J. Goldstein, “The Use of MMR Diversity-Based Reranking For Reordering Documents and Producing Summaries,” in *Special Interest Group on Information Retrieval*, 1998, p.p. 335-336.
- [28] R. Mihalcea dan P. Tarau, “TextRank: Bringing Order into Texts,” in *Conference on Empirical Methods in Natural Language Processing*, 2004.
- [29] K. Kurniawan dan S. Louvan, “INDOSUM: A New Benchmark Dataset for Indonesian Text Summarization,” in *International Conference on Asian Language Processing*, 2018.
- [30] C.Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Annual Meeting of the Association for Computational Linguistics*, 2004.
- [31] A. Malahyari, S. Pouriye, M. Assefi, S. Safaiei, E.D. Trippe, J.P. Gutierrez and K. Kochut, “Text Summarization Techniques: A Brief Survey,” *International Journal of Advanced Computer Science and Applications*, 2017.
- [32] D.K. Gaikwad and C.M. Mahender, “A Review Paper on Text Summarization,” *International Journal of Advanced Research in Computer and Communication Engineering*, 2016.
- [33] C.D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval,” *Cambridge University Press*, 2008.
- [34] F. Chollet, “Deep Learning with Python,” *Manning Publications*, 2017.
- [35] P. Kouris, G. Alexandridis and A. Stafylopatis, “Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization,” in *Annual Meeting of the Association for Computational Linguistics*, 2019.
- [36] Y. Yuliska and T. Sakai, “A Comparative Study of Deep Learning Approaches for Query-Focused Extractive Multi-Document Summarization,” in *International Conference on Information and Computer Technologies*, 2019, p.p. 153-157.

- [37] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems*, 2013.
- [38] J. Pennington, R. Shocker and C.D. Manning, "GloVe: Global Vectors for Word Representation," in *Conference on Empirical Methods in Natural Language Processing*, 2016, p.p. 1532-1543.
- [39] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in *Conference of the European Chapter of the Association for Computational Linguistics*, 2017.
- [40] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019.
- [41] Z. Cai, N. Lin, C. Ma and S. Jiang, "Indonesian Automatic Text Summarization Based on A New Clustering Method in Sentence Level," in *International Conference on Big Data Engineering*, 2019.
- [42] J.H. Lee, S. Park, C. M. Ahn and D. Kim, "Automatic generic document summarization based on non-negative matrix factorization," *Information Processing and Management: an International Journal*, 2009.
- [43] A. Ridok, "Peringkasan Dokumen Bahasa Indonesia Berbasis Non-Negative Matrix Factorization (NMF)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, 2014.
- [44] S. Silvia, P. Rukmana, V. R. Aprilia, D. Suhartono, R. Wongso and M. Meiliana, "Summarizing Text for Indonesian Language by Using Latent Dirichlet Allocation and Genetic Algorithm," in *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics*, 2014, p.p. 148-153.
- [45] E.Y. Hidayat, F. Firdausillah, K. Hastuti, I. N. Dewi and A. Azhari, "Automatic Text Summarization Using Latent Dirichlet Allocation (LDA) for Document Clustering," *International Journal of Advances in Intelligent Informatics*, 2015.
- [46] M. Zoph, E. L. Mencia and J. Fürnkranz, "Which Scores to Predict in Sentence Regression for Text Summarization?," in *the Proceedings of NAACL-HLT*, 2018, p.p. 1782-1791.
- [47] M. Zoph, E. L. Mencia and J. Fürnkranz, "Which Scores to Predict in Sentence Regression for Text Summarization?," in *the Proceedings of NAACL-HLT*, 2018, p.p. 1782-1791.
- [48] Y. Yuliska and T. Sakai, "Query-Focused Extractive Summarization based on Deep Learning: Comparison of Similarity Measures for Pseudo Ground Truth Generation", in *the Data engineering and Information Management Forum*, 2019.
- [49] A. Najibullah, "Indonesian Text Summarization based on Naïve Bayes Method", in *the Proceeding of the International Seminar and Conference*, 2015, p.p. 67-78.
- [50] R. Indrianto, M. A. Fauzi and L. Muflikhah, "Peringkasan Teks Otomatis Pada Artikel Berita Kesehatan Menggunakan K-Nearest Neighbor Berbasis Fitur Statistik", *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2017.
- [51] Y. Kim, "Convolutional Neural Networks for Sentence Classification", in *Conference on Empirical Methods in Natural Language Processing*, 2014, p.p. 1746-1751.
- [52] F. Koto, "A Publicly Available Indonesian Corpora for Automatic Abstractive and Extractive Chat Summarization", in *International Conference on Language Resources and Evaluation*, 2016, p.p. 801-805.
- [53] Q. Mei and C. X. Zai, "Generating Impact-Based Summaries for Scientific Literature", in *Annual Meeting of the Association for Computational Linguistics*, 2008, p.p. 816-824.
- [54] V. Qazvinian and D. R. Radev, "Scientific paper summarization using citation summary networks", in *International Conference on Computational Linguistics*, 2008, p.p. 689-696.

**BIOGRAFI PENULIS**

	<p><b>Yuliska</b>, lahir di Kotabaru, 08 Juli 1991. Menyelesaikan Pendidikan Sarjana (S1) di jurusan Teknik Informatika, UIN SUSKA Riau dan Pendidikan Magister (S2) di Jurusan Computer Science and Communications Engineering, Waseda University-Jepang. Saat ini mengajar di Jurusan Teknik Informatika, Politeknik Caltex Riau dan melanjutkan melakukan penelitian di bidang <i>Natural Language Processing</i>, <i>Text Mining</i>, <i>Machine Learning</i>, <i>Deep Learning</i> dan <i>Human Computer Interaction</i>.</p>
	<p><b>Khairul Umam Syaliman</b>, lahir di Perawang, 21 Juni 1992. Menyelesaikan Pendidikan sarjana (S1) di jurusan Teknik Informatika, Universitas Islam Riau (UIR) dan Pendidikan master (S2) di jurusan yang sama, Universitas Sumatra Utara (USU). Saat ini mengajar di Jurusan Teknik Informatika, Politeknik Caltex Riau dan melanjutkan melakukan penelitian di bidang <i>Data Mining</i> dan <i>Machine Learning</i>.</p>