# COMPARISON OF LEXRANK ALGORITHM AND MAXIMUM MARGINAL RELEVANCE IN SUMMARY OF INDONESIAN NEWS TEXTS IN ONLINE NEWS PORTALS

**Siti Tuhpatussania [1] ; Ema Utami [2] ; Anggit Dwi Hartanto [3]**

Informatics Engineering Study Program
Universitas AMIKOM Yogyakarta
Yogyakarta, Indonesia
www.amikom.ac.id
[1] tuhpatussania@students.amikom.ac.id; [2] ema.u@amikom.ac.id ; [3] anggit@amikom.ac.id

(*) Corresponding Author

***Abstract*** *— The presence of online media has shifted print media for news readers to get information that is fast, accurate and easy to access. However, the problem arises because the length of the news text makes the reader bored to search for the news as a whole so that the news that is obtained will be less accurate. For this reason, it is necessary to have an automatic text summary that was raised in this study as well as to compare the Maximum Marginal Relevance (MMR) algorithm and the LexRank algorithm to the summary of Indonesian news texts on the online news portal graphanews.com. the results of the comparison test of text summarization using fmeasure, precision and recall show the performance of text summarization with the MMR algorithm is better where fmeasure is 91.65%, precision is 91.08% and recall is 92.23%.*

***Keywords:*** *autotext summarization, lexrank , mmr , tf-idf .*

**Abstrak**—Kehadiran media online menggeser media cetak bagi pembaca berita dalam mendapatkan informasi yang cepat, akurat dan mudah di akses. Akan tetapi masalah muncul karena panjangnya teks berita membuat pembaca jenuh untuk mencari berita secara utuh sehingga berita yang di dapatkan akan menjadi kurang akurat. Untuk itu perlunya ada peringkasan teks otomatis yang diangkat pada penelitian ini sekaligus membandingkan algoritma *Maximum Marginal Relevance*(MMR) dan algoritma LexRank pada peringkasan teks berita berbahasa Indonesia di portal berita online grafikanews.com. hasil dari pengujian perbandingan peringkasan teks menggunakan *fmeasure, precision* dan *recall* menunjukkan kinerja peringkasan teks dengan algoritma MMR lebih baik dimana *fmeasure* 91.65%, *precision* 91.08% dan *recall* 92,23%.

**Keywords :** peringkasan teks otomatis, lexrank, mmr, tf-idf.

## INTRODUCTION

The development of the internet today this the more uphill rapidly which has an impact on the development technology communication for publish articles in online media. in line with development of the internet, letters news already switch upload articles news through online media and online news portals where Thing that make it easy community in get news with fast and efficient. Likewise with you interest Indonesian people in Internet usage in 2020 amounted to about 200 million or equivalent with ± 65% of the total population , and the number of the internet user will keep going increase (Shiddiqi et al., 2020)

Already becomes obligation for online news portals to present contents digital news with arrangement of words and delivery that is easily understood by the reader online news but in reality contents news often no organized with long sentence cause reader difficult understand meaning from news that and easy fed up for complete news that has been read. Summary news is needed so that content old news long could presented by short with the goal for readers understand the essence of something thinking main from news the (Ayu Syahfitri et al., 2022).

Study related summary text done by(Fauzi, 2022) about use Text Mining algorithms and algorithms LexRank for carry out the summary process text, the method used in the research the is method based graph that is algorithm LexRank that can be proven use research that has been tested on news data in Indonesian from liputan6.com. Amount extracted sentences only 25% -50% of the total the sentence listed in documents, results obtained from summary algorithm LexRank is order from highest weight to low. On research summary text use algorithm LexRank this capable produce summary text automatic without remove meaning actually will but application algorithm the still counted enough weak because there is score duplicate so that needed combination with other

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i2.3190

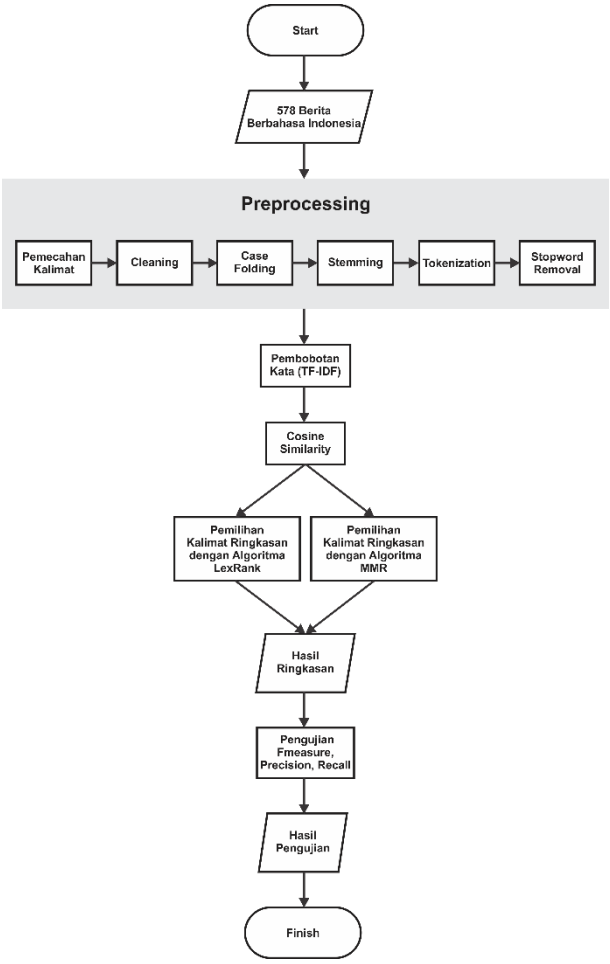
algorithms, for example, in stages preprocessing to get results more ending high. Study before related warning text automatic too done by(Dimas et al., 2022) them to do study for summary text auto on news portal sport use Maximum Marginal Relevance method and influence Maximum Marginal Relevance to results accuracy summary system that is testing taken of 5 samples news online news using lamda 0.7 ago results summary the used for compared with summary system and summary by expert after that searching for Correct then searched wrong and missed. Test results got from find the average of precision, recall, f-measure, so that influence lamda 0.7 produces an average accuracy of 57.7% Precision, 48.5% Recall and 50.3% F-Measure.

Based on studies literature that researchers do so studies case in research this is on the grafikanews.com news portal because on the grafikanews.com news portal the news that is presented dominant Indonesian with various type category news as well as different grammatical arrangements in the same meaning by each the journalist of course could influence results summary text automatic. And based on study before related summary text automatic that has been described previously produce level performance algorithm LexRank and MMR are enough low in summary text automatic motivate researcher for use second algorithm the in produce score more accuracy tall for summary text using the dataset in the study case study this. Final result study is see level more accuracy tall from algorithm LexRank and MMR algorithm on the same dataset with to do comparison from results summary text automatic.

## MATERIALS AND METHODS

In study this use methodology quantitative where research quantitative is research that describes or explain something problem that results could generalized . Method quantitative here cover where data collection in research this data is text in the next CSV format file generalized in form numbers for make it easy data processing and presentation return in form text .

processing on summary text automatic this through a number of stages among them stages dataset input, stages preprocessing, stage last word weighting cosine similarity stage, next stages election sentence summary use algorithm LexRank and Algorithm Maximum Marginal Relevance (MMR) of each result summary text from second algorithm the will be tested using fmeasure , precision and recall to find out results comparison the most accurate algorithm . Flow chart stages study as in Figure 1.




Source (Tuhpatussania, 2022)

Figure 1. Stages Study

### A. Dataset

Stages first on research this that is input the retrieved dataset from grafikanews.com. grafikanews.com is an online news portal that has variety content start from news politics, events, economy, sports, technology and more where news presented dominant Indonesian.

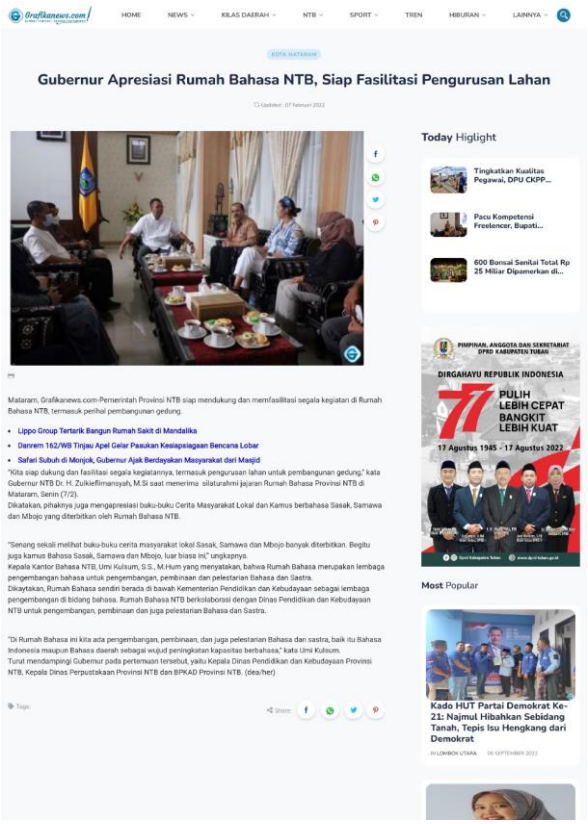
Grafikanews.com is also one of the online news portals that has get away Press Council Verification so that news presented can be sure its accurate.

Data used in research this taken for 5 months in the period January 2022 to May 2022 with news in Indonesian with a total of 578 news items. *Sample* dataset title can seen in figure 2 and details news could seen in figure 3 and figure 4, like following:

| Judul Berita | Tgl Posting | Status |
|---|---|---|
| Ditelantarkan Orang Tak Dikenal Di Pulau Merah, Polresta Banyuwangi Bantu Pulangkan Tiga Bocah | 23 Mei 2022 | Published |
| Semarakkan World Surf League, Banyuwangi Gelar Pelatihan Selancar Anak Muda | 21 Mei 2022 | Published |
| Jelang Idul Adha, Banyuwangi Pastikan Hewan Ternak Aman | 21 Mei 2022 | Published |
| Momen Harkitnas, Bupati Ipuk Ajak Masyarakat Bangkit dari Pandemi | 21 Mei 2022 | Published |
| Meriahkan Salatiga Expo Hybrid 2022, Danrem 073/MKT, Ikut Menjadi Peserta Yos Sudarso Fun Bike | 21 Mei 2022 | Published |
| Kirab Budaya Hari Jadi Ke 56 Kabupaten Batang Di Gelar | 21 Mei 2022 | Published |
| Menteri Sosial Kunjungi Tarno Warga Patean Kendal, Penderita Tumor Otak | 21 Mei 2022 | Published |
| Banyuwangi Tutup Akses Masuk Ternak dari Luar Daerah untuk Antisipasi PMK | 21 Mei 2022 | Published |
| Wakil ketua I DPRD Banyuwangi Minta SKPD Segera Selesaikan Persoalan Di Kabupaten Banyuwangi | 21 Mei 2022 | Published |
| Terima Kunjungan Menteri Keuangan Singapura, Bupati Kendal Dorong Perluasan Kerjasama Sektor Pariwisata | 20 Mei 2022 | Published |

Source (Tuhpatussania, 2022)

Figure 3. News Dataset Indonesian




Source (Tuhpatussania, 2022)

Figure 4. Details News Indonesian

## B. Preprocessing

Stages early on processing summary text automatic this that is stages preprocessing. Preprocessing is stages where application To do selection of data to be processed on every document (Hermawan et al., 2020).

Preprocessing process this cover a number of stages that is as following :

a. Solution sentence that is the first process in stages preprocessing this where is the splitting process document or all the data entered is broken be per sentence use dot delimiter comma and sign ask ( . , ?).
b. Next is the cleaning process working for delete noise in the text that has been through Step solving sentence example noise like number, sign open brackets and more .
c. After through stages last cleaning to the case folding process, in this process will replace letter big (uppercase) becomes letter small all (lowercase) so that the letters in the dataset are equal,
d. Furthermore, the stemming process is the process of changing words that have affixes such as " passed away " replaced becomes " home " or the root word .
e. Tokenization process in preprocessing this almost the same function with the first process that is solving sentence, the difference with tokenization that is break structured sentences on basic words and letters small all be per word using space delimiter.
f. Then the last process is stopword removal that is deletion of words that are considered no important such as “which”, “and”, “ is ” to reduce word count on processing summary text next (Sari Yunita & Fatonah Nenden, 2021).

## C. Word Weight

In stages this will use method The weighting of the Term-Frequency Inverse Document Frequency (TF-IDF) word is calculation or word weighting and frequency the emergence of the word in given document show the importance of that word in a document(Rofiqi et al., 2019). Formula TF-IDF calculation looks as following :

$$W_{td} = Tf_{td\_} * IDF_t = Tf_{td} x \log \frac{N}{dft} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Description :
*d* = document to - *d*
*t* = word to – *t* from *term*
W = weight term *t* on document *d*
*tf* = sum appearance *term* at *t* on document *d*
*N* = total documents
*df* = sum documents that have *term t.*

## D. Cosine Similarity

Stages next that is calculation *Cosine Similarity* aim count *similarity* or similarity of the words generated in the previous process. Formula for *Cosine Similarity* are :

$$\cos a = \frac{A \,.\, B}{|A| \,.\, |B|} = \frac{\sum_{i=1}^{n} A_i \, X \, B}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \; x \; \sqrt{\sum_{i=1}^{n} (B_i)^2}} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Description :
A = vector A, which will compared similarity
B = Vector B, which will compared similarity

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i2.3190

AB = dot product between vector A and vector B
| $A$ | = length of vector A
| $B$ | = length of vector B
| $A$ || $B$ | = cross product between | $A$ | and | $B$ |

**E. Election Sentence Summary**

Stages election sentence summary text automatic on research this use different algorithm that is algorithm LexRank and the MMR algorithm where Algorithm *Maximum Marginal Relevance* (MMR) is technique summary text with destination take accurate information without contain redundancy. MMR technique in summarizing document that is count similarity Among part text with destination get score sentence based on *similarity* and with the given query could reduce redundancy in results summary (Arisandi & Sutrisno, 2022). Following formula MMR :

$$\text{MMR}\,(S_i) = \lambda . Sim_1\,(S_i, Q) - (1 - \lambda)\,.\,max\,Sim_2(S_i, S_j) \quad \text{.....( 3)}$$

Description :
sentence vector candidate answer
Q = sentence vector question
Sim ( S,Q ) = cosine similarity between Si and Q . sentences
Sim (Si, Sj ) = cosine similarity between Si and Sj kalimat sentences

Algorithm in stages election sentence summary next is algorithm LexRank in the process of summarizing use approach centroid based. Algorithm technique LexRank is with combine score *probability stationary* with feature other like use combination linear and position sentence. LexRank use processing post heuristic that is produce summary with add sentence in order rating (Fauzi, 2022).

**F. Test**

After getting results summary text of each algorithm the so will conducted testing use *precision, recall* and *fmeasure. Precision* that is comparison Among *True Positive* (TP) with predictable amount of data positive whereas *recall* that is comparison Among *True Positive* (TP) with the actual amount of data positive and testing *fmeasure* is *harmonic mean* from *precision* and *recall*.

Test using these 3 parameters could measure performance from algorithm used in study this in make predict and not only give information about errors made by the model but also the type mistakes made.

## RESULTS AND DISCUSSION

Explanation related results testing the proposed model and discussion the process will be reviewed in section this. Summary text through stages *Preprocessing*, Weighting, *Text Summarization* second the proposed method then testing. The whole implementation process algorithm use language python programming and results from testing then will evaluated with *f-measure* or *f1score, precision* , and *recall.*

**A. Preprocessing**

In stages this filled document on arrangement sentence will through stages normalization for find the root word and delete duplicate words that aim to be at the stage next that is word weighting get more results accurate.

**B. TF-IDF and Cosine Similarity**

Results in the next TF-IDF stage saved in form *arrays,* such as calculation TF-IDF value for each word in the dataset used for study this. Example results TF-IDF calculation on one of the news on grafikanews.com is described in the following table :

Table 1. Calculation of IDF

| Term | Frekuensi Kata di setiap kalimat | | | | | | | IDF= log(D/DF) |
|---|---|---|---|---|---|---|---|---|
| | N1 | N2 | N3 | N4 | N5 | N6–N11 | N12 | |
| Rumah | 0 | 1 | 1 | 0 | 0 | 0 | 0 | Log(2/12) =0,778 |
| Bahasa | 0 | 1 | 1 | 0 | 0 | 0 | 0 | Log(2/12) =0,778 |
| Lahan | 1 | 0 | 0 | 2 | 0 | 0 | 1 | Log(4/12) =0,477 |
| Fasilitas | 0 | 0 | 0 | 1 | 1 | 0 | 0 | Log(2/12) 0,778 |
| Gubernur | 2 | 0 | 0 | 0 | 2 | 0 | 0 | Log(4/12) = 0,477 |
| Tingkat | 1 | 0 | 1 | 0 | 0 | 0 | 1 | Log(3/12) =0,602 |
| Buku | 1 | 1 | 0 | 0 | 0 | 0 | 2 | Log(4/12) 0,477 |

Source (Tuhpatussania, 2022)

Table 2. Calculation of TF-IDF

| Term | N1 | N2 | N3 | N4 | N5 | ... | N12 |
|---|---|---|---|---|---|---|---|
| Rumah | 0 | 0,778 | 0,778 | 0 | 0 | | 0 |
| Bahasa | 0 | 0,778 | 0,778 | 0 | 0 | | 0 |
| Lahan | 0,477 | 0 | 0 | 0,954 | 0 | | 0,477 |
| Fasilitas | 0 | 0 | 0 | 0,778 | 0,778 | | 0 |
| Gubernur | 0,954 | 0 | 0 | 0 | 0,954 | | 0 |
| Tingkat | 0,602 | 0 | 0,602 | 0 | 0 | | 0,602 |
| Buku | 0,477 | 0,477 | 0 | 0 | 0 | | 0,954 |

Source (Tuhpatussania, 2022)

In Table 2 it can be seen score results TF-IDF calculation for every *term* or word. Next the results of the TF- IDF will used for calculation score vector length of each sentence. Count score vector length with method the value of the raised IDF then, value *term* in one sentence rooted after adding up. long value *vector* as in Table 3 and calculations

*Cosine Similarity* use formula *Cosine Similarity* with results calculation as in Table 4:

Table 3. Vector Length Value Every Sentence

| Term | N1 | N2 | N3 | N4 | N5 | ... | N12 |
|---|---|---|---|---|---|---|---|
| Rumah | 0 | 0,605 | 0,605 | 0 | 0 | | 0 |
| Bahasa | 0 | 0,605 | 0,605 | 0 | 0 | | 0 |
| Lahan | 0,227 | 0 | 0 | 0,910 | 0 | | 0,227 |
| Fasilitas | 0 | 0 | 0 | 0,605 | 0,605 | | 0 |
| Gubernur | 0,910 | 0 | 0 | 0 | 0,910 | | 0 |
| Tingkat | 0,362 | 0 | 0,362 | 0 | 0 | | 0,362 |
| Buku | 0,227 | 0,227 | 0 | 0 | 0 | | 0,910 |
| Nilai Vektor | 1,313 | 1,198 | 1,253 | 1,230 | 1,230 | | 1,224 |

Source (Tuhpatussania, 2022)

Table 4. Calculation *Cosine Similarity*

| Term | N1 | N2 | N3 | N4 | N5 | ... | N12 |
|---|---|---|---|---|---|---|---|
| Rumah | 0 | 0,736 | 0,703 | 0 | 0 | | 0 |
| Bahasa | 0 | 0,736 | 0,703 | 0 | 0 | | 0 |
| Lahan | 0,562 | 0 | 0 | 0,794 | 0 | | 0,564 |
| Fasilitas | 0 | 0 | 0 | 0,717 | 0,717 | | 0 |
| Gubernur | 0,743 | 0 | 0 | 0 | 0,794 | | 0 |
| Tingkat | 0,590 | 0 | 0,619 | 0 | 0 | | 0,633 |
| Buku | 0,562 | 0,576 | 0 | 0 | 0 | | 0,797 |

Source (Tuhpatussania, 2022)

### C. LexRank and MMR

Algorithm LexRank and MMR are used in stages summarizing for get weight end inside all existing *vertices* in graph. After results weighting each *vertex* is obtained, weight it is sorted based on score highest. On weighting end use MMR algorithm generates score between 0.2327-52.668 and weighting use LexRank is 0.000-0.5789.

### D. Comparison Results

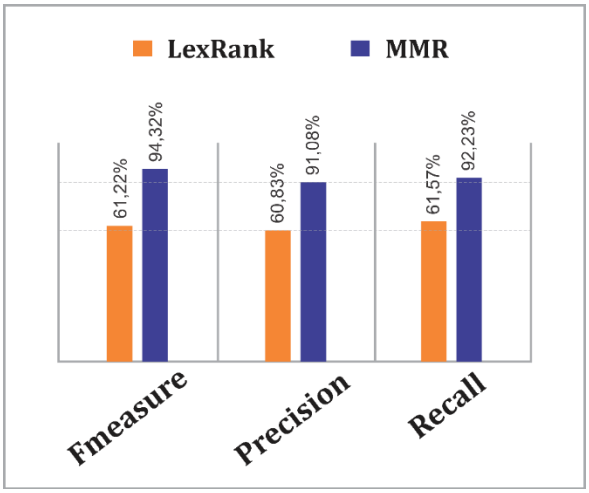
Test end on research this is compare results from summary text use algorithm LexRank and the MMR algorithm. Test results summary text automatic this use method *fmeasure*, *precision* and *recall* as in the following table :

Table 5. Comparison of Calculation Results *fmeasure* , *precision* , *recall*

| Evaluation | LexRank | MMR |
|---|---|---|
| *Fmeasure* | 0.6119 | 0.9165 |
| *Precision* | 0.6083 | 0.9108 |
| *Recall* | 0.6157 | 0.9223 |

Source (Tuhpatussania, 2022)

the table show results testing that MMR method more superior in comparison method LexRank, with superiority as big as *fmeasure* 0.3046, *precision* 0.3025, *recall* 0.3066. Figure 5 following description percentage comparison results testing algorithm summary text used:




Source (Tuhpatussania, 2022)

Figure 5. Percentage Comparison of Test Results Algorithm LexRank and MMR

## CONCLUSION

From result testing summary text auto on news dataset in Indonesian at grafikanews.com you can taken conclusion for summary text automatic researcher do with use two algorithm that algorithm *Maximum Marginal Relevance* (MMR) shows results more performance good where score *fmeasure* 0.9165 or 91.65%, *precision* 0.9108 or 91.08% and *recall* 0.9223 or 92.23% compared summary text use algorithm LexRank with the average value *fmeasure, precision* and *recall* of 61.19%.

## REFERENCE

Andriani, D., & Tanzil Furqon, M. (2019). Peringkasan Teks Otomatis Pada Artikel Berita Hiburan Berbahasa Indonesia Menggunakan Metode BM25. *JPTIK (Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer), 3(3), 2603-2610. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4761*

Arisandi, D., & Sutrisno, T. (2022). Aplikasi peringkasan dokumen menggunakan metode maximum marginal relevance (mmr). *Jurnal Ilmu Komputer dan Sistem Informasi, 10(1). https://doi.org/10.24912/jiksi.v10i1.17820*

Dimas, F., Al-Hafiidh, F., Rozi, I., & Kusumaning, P. (2022). Peringkasan teks otomatis pada portal berita olahraga menggunakan metode maximum marginal relevance. *JIP (Jurnal Informatika Polinema), 8(1), 21-30. http://repota.jti.polinema.ac.id/id/eprint/464*

Elbarougy, R., Behery, G., & el Khatib, A. (2020). Extractive Arabic Text Summarization Using Modified PageRank Algorithm. *Egyptian Informatics Journal*, *21(2), 73-81. https://doi.org/10.1016/j.eij.2019.11.001*

Fauzi, A. (2022). Penerapan Algoritma Text Mining dan Lexrank dalam Meringkas Teks Secara Otomatis. *Bulletin of Data Science*, *1(2), 65-72. https://ejurnal.seminar-id.com/index.php/bulletinds/article/view/1359*

Hermawan, L., Ismiati, M. B., Bangau, J., 60, N., & Charitas, M. (2020). Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. *TRANSFORMATIKA*, *17(2), 188-199. http://dx.doi.org/10.26623/transformatika.v17i2.1705*

Lin, N., Li, J., & Jiang, S. (2022). A simple but effective method for Indonesian automatic text summarisation. *Connection Science, 34(1), 29-43. https://doi.org/10.1080/09540091.2021.1937942*

Rifano, E. J., Fauzan, Abd. C., Makhi, A., Nadya, E., Nasikin, Z., & Putra, F. N. (2020). Text Summarization Menggunakan Library Natural Language Toolkit (NLTK) Berbasis Pemrograman Python. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, 2(1), 8-17, *https://doi.org/10.28926/ilkomnika.v2i1.32*

Riyani, A., Zidny Naf'an #2, M., & Burhanuddin, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Jurnal Linguistik Komputasional*, *2(1), 23-27. https://doi.org/10.26418/jlk.v2i1.17*

Rofiqi, M. A., Fauzan, Abd. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics, 1(2), 58-64. https://doi.org/10.28926/ilkomnika.v1i2.18*

R., Kurniawan, A., Irsan Humaidy, M. (2022). Penerapan Algoritma Maximum Marginal Relevance Dalam Peringkasan Teks Secara Otomatis. *Bulletin of Data Science*, 1*(2), 49-56. https://ejurnal.seminar-id.com/index.php/bulletinds/article/view/1358*

Sari Yunita, & Fatonah Nenden. (2021). Peringkasan Teks Otomatis pada Modul Pembelajaran Berbahasa Indonesia Menggunakan Metode Cross Latent Semantic Analysis (CLSA). *JEPIN (Jurnal Edukasi Dan Penelitian Informatika), 7(2), 153-159. http://dx.doi.org/10.26418/jp.v7i2.47768*

Shiddiqi, A. M., Ijtihadie, R. M., Ahmad, T., Wibisono, W., Anggoro, R., Bagus, D., & Santoso, J. (2020). Penggunaan Internet dan Teknologi IoT untuk Meningkatkan Kualitas Pendidikan. *Jurnal Direktorat Riset dan Pengabdian Kepada Masyarakat-DRPM ITS, 4(3), 235-240. https://journal.its.ac.id/index.php/sewagati/article/view/369*

Tuhpatussania, S. (2022). Perbandingan Algoritma LexRank dan Maximum Marginal Relevance pada *peringkasan* teks berbahas Indonesia pada portal berita online, 18(2), 187-192. 10.33480/pilar.v18i2.3190

Yulita, W., Priyanta, S., & SN, A. (2019). Automatic Text Summarization Based on Semantic Networks and Corpus Statistics. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(2), 137-148. 10.22146/ijccs.38261*