PERBANDINGAN ALGORITMA *K-NEAREST NEIGHBOR (KNN)* DAN *ROCCHIO* DALAM KLASIFIKASI TUGAS AKHIR UNIVERSITAS MUHAMMADIYAH JEMBER STUDI KASUS: FAKULTAS TEKNIK

Farid Achmad Arif Adani[1], Deni Arifianto[2], Habibatul Azizah Al Faruq[3]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
faridhimajo@gmail.com[1], deniarifianto@unmuhjember.ac.id[2], habibatulazizah@unmuhjember.ac.id[3]

ABSTRAK

Tugas akhir merupakan salah satu syarat kelulusan yang harus dipenuhi oleh mahasiswa untuk menyelesaikan pendidikan di perguruan tinggi. Semakin bertambahnya mahasiswa tiap tahunnya, maka semakin banyak pula koleksi dokumen tugas akhir. Semakin banyaknya dokumen tugas akhir menyebabkan sulitnya mengkategorikan dokumen tugas akhir jika harus dilakukan secara manual. Mengelompokan tugas akhir berdasarkan judul saja kurang efektif karena ada judul tugas akhir yang dapat dikategorikan kedalam lebih dari satu program studi khususnya di lingkungan fakultas teknik. Pada penelitian ini dilakukan klasifikasi dokumen terhadap abstrak dan bab 1 Tugas Akhir mahasiswa Fakultas Teknik Universitas Muhammadiyah Jember. Data yang digunakan dalam penelitian ini adalah abstrak dan bab 1 Tugas Akhir pada program studi Teknik Elektro, Teknik Sipil, Teknik Informatika, Teknik Mesin, dan Manajemen Informatika. Metode klasifikasi pada penelitian ini adalah metode *K-Nearest Neighbor* (KNN) dan *Rocchio.* Pengujian akurasi pada penelitian ini dilakukan dengan *Cross Validation* dan evaluasi data uji dengan *Confusion Matrix.* Dari penelitian ini didapatkan hasil pada 150 data tugas akhir, metode *Rocchio* menghasilkan nilai akurasi yang sama dengan KNN yaitu 96%, sedangkan presisi dan recall menghasilkan nilai lebih baik yaitu presisi sebesar 97% dan *recall* sebesar 97%, pengujian pada metode KNN menghasilkan nilai presisi sebsar 95% dan *recall* sebesar 95% dengan nilai K=19.

Kata Kunci : Klasifikasi dokumen, Tugas Akhir, *K-Nearest Neighbor, Rocchio.*

PENDAHULUAN

Latar Belakang

Tugas akhir merupakan salah satu syarat kelulusan yang harus dipenuhi oleh mahasiswa untuk menyelesaikan pendidikan di perguruan tinggi. Ilmu yang didapat selama perkuliahan dituangkan ke dalam suatu penelitian yang akan menghasilkan luaran berupa dokumen tugas akhir (Yusra dkk, 2016).

Semakin bertambahnya mahasiswa tiap tahunnya, maka semakin banyak pula koleksi dokumen tugas akhir. Semakin banyaknya dokumen tugas akhir menyebabkan sulitnya mengkategorikan dokumen tugas akhir jika harus dilakukan secara manual. Mengelompokan tugas akhir berdasarkan judul saja kurang efektif karena ada judul tugas akhir yang dapat di kategorikan ke dalam lebih dari satu program studi khususnya di lingkungan fakultas teknik. Salah satu cara yang berhasil dalam mengkategorikan dokumen tugas akhir dalam jumlah banyak yaitu dengan klasifikasi dokumen tugas akhir.

Penelitian klasifikasi tugas akhir sendiri sudah pernah dilakukan antara lain yaitu Yusra dkk (2016) melakukan penelitian berjudul Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode *Naïve Bayes Classifier* dan *K-Nearest Neighbor*, menghasilkan kesimpulan bahwa metode *Bayes Classifier* dan *K-Nearest Neighbor* mampu mengklasifikasikan tugas akhir mahasiswa jurusan Teknik Informatika. Lestari dkk (2019) melakukan penelitian berjudul Klasifikasi Teks Berbasis Ontologi Untuk Dokumen Tugas Akhir Berbahasa Indonesia, menghasilkan kesimpulan bahwa pengujian yang dilakukan kepada sistem yang telah dibuat menghasilkan nilai akurasi 87%.

Universitas Muhammadiyah Jember (UM JEMBER) merupakan salah satu perguruan tinggi swasta yang berada di Kota Jember. salah satu fakultas yang terdapat di UM JEMBER adalah Fakultas Teknik yang di dalamnya terdapat lima jurusan yaitu Teknik Informatika, Teknik Sipil, Teknik Mesin, Teknik Elektro dan Manajemen Informatika. Sebagai syarat kelulusan, mahasiswa diwajibkan untuk menyusun tugas akhir. Namun tugas akhir yang ada di Fakultas Teknik Universitas Muhammadiyah Jember belum ada yang mengklasifikasikan berdasarkan program studi.

Dalam *data mining* terdapat beberapa metode pengklasifikasian salah satu diantaranya yaitu KNN (*K-Nearest Neighbor*) & *Rocchio*. Penelitian – penelitian menggunakan metode KNN & *Rocchio* telah banyak diterapkan dan menyebutkan hasil dari dua algoritma tersebut memiliki akurasi yang cukup bagus dalam metode klasifikasi, berdasarkan penelitian yang dilakukan oleh Moldagulova dan Sulaiman (2017), yang membahas tentang penggunaan algoritma KNN untuk klasifikasi dokumen tekstual. Dimana dalam penelitiannya algoritma KNN menghasilkan akurasi sebesar 97,14 % untuk K = 5. Begitu juga dengan metode *Rocchio*, berdasarkan penelitian yang dilakukan oleh Afriza dan Adisantoso (2018), yang membahas tentang metode klasifikasi *Rocchio* untuk analisis *hoax* dengan algoritma *multinomial naive bayes* sebagai pembanding dimana algoritma *Rocchio* menghasilkan akurasi yang lebih besar yaitu 83,501 %.

Berdasarkan uraian di atas, maka dalam penelitian ini akan dilakukan perbandingan antara metode KNN dan *Rocchio* untuk mengklasifikasikan tugas akhir berdasarkan abstrak dan bab 1 di Fakultas Teknik Universitas Muhammadiyah Jember kedalam 5 program studi yaitu Teknik Sipil, Teknik Informatika, Teknik Mesin, Teknik Elektro, dan Manejemen Informatika dengan judul "PERBANDINGAN ALGORITMA *K-NEARST NEIGHBOR* (KNN) DAN *ROCCHIO* DALAM KLASIFIKASI TUGAS AKHIR UNIVERSITAS MUHAMMADIYAH JEMBER STUDY KASUS FAKULTAS TEKNIK"

Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan di atas, maka permasalahan yang akan dibahas dalam penelitian ini adalah sebagai berikut: Berapakah tingkat akurasi, presisi, dan *recall* perbandingan algoritma *K-Nearest Neighbor (*KNN*) dan Rocchio* dalam klasifikasi tugas akhir Fakultas Teknik Universitas Muhammadiyah Jember berdasarkan abstrak dan bab 1?

Batasan Masalah

Agar permasalahan tidak menyimpang pada tujuan penelitian, maka berikut beberapa batasan yang perlu dibuat, yaitu:

1. Jumlah dataset yang digunakan dalam penelitian ini sebesar 150 *record* dengan masing – masing 30 *record* untuk tiap kelas.
2. Dokumen yang digunakan berupa abstrak dan bab 1 pada tugas akhir Fakultas Teknik yang ada di *Repository* Universitas Muhammadiyah Jember yang diambil secara acak .
3. Bahasa pemrograman yang digunakan dalam penelitian ini adalah *Python.*
4. *Tools* yang digunakan dalam penelitian ini adalah *Jupyter Notebook.*
5. Apabila terdapat *record* data yang memiliki lebih dari satu kelas pada KNN maka diasumsikan benar jika salah satu kelas prediksi sama dengan kelas sebenarnya.
6. Objek penelitian ini berupa kelas SIPIL, MI, TI, MESIN, & ELEKTRO.
7. Merupakan klasifikasi *single class.*
8. Menghitung nilai akurasi, presisi, dan *recall*.
9. Menggunakan modul *stemming* Sastrawi.
10. Menggunakan modul *NLTK* untuk *Stopword.*
11. Menggunakan modul *sklearn* untuk algoritma *KNN* dan *Rocchio*

Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

Membandingkan tingkat akurasi, presisi, dan *recall* dari algoritma *K-Nearest Neighbor (*KNN*)* dan *Rocchio* dalam klasifikasi tugas akhir Fakultas Teknik Universitas Muhammadiyah Jember.

Manfaat Penelitian

Manfaat dari penelitian ini yang dilakukan sebagai berikut:

1. Bagi penulis:
   Hasil dari penelitian ini diharapkan dapat menambah wawasan dalam bidang *text mining*, serta sebagai syarat kelulusan penulis.
2. Bagi instansi:

Diharapkan dapat memberikan kontribusi secara keilmuan berupa hasil perbandingan algoritma *K-Nearest Neighbor* dan *Rocchio* dalam klasifikasi tugas akhir mahasiswa Fakultas Teknik Universitas Muhammadiyah Jember.

3. Bagi peneliti lain:
   Hasil penelitian ini dapat digunakan sebagai referensi untuk penelitian-penelitian berikutnya di bidang klasifikasi dokumen dengan metode *text mining*.

METODE PENELITIAN

Term Weighting TF-IDF

Achmad dan Ilham (dalam Lestari dkk. 2019) menjelaskan bahwa *Term frequency – inverse document frequency* atau TF-IDF adalah metode pembobotan kata dengan menghitung nilai TF dan juga menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan dan menghitung inverse frekuensi dokumen yang mengandung kata (IDF).dengan rumus:

$$W_{td} = tf_{td} * \log\left(\frac{N}{df_t}\right) + 1 \qquad (2.1)$$

Keterangan:
$w(t,d)$ = hasil pembobotan TF-IDF pada term *t* di dalam data *d*.
$tf(t,d)$ = nilai kemunculan *term t* di dalam data *d*.
N = total dokumen
df_t = jumlah data komentar yang mengandung *term t*.

K-Nearest Neighbor

Metode *K-Nearest Neighbor* (KNN) bekerja dengan cara mencari sejumlah k pola (diantara semua pola latih yang ada di semua kelas) yang terdekat dengan pola masukan, kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak di antara k pola tersebut (voting) (Suyanto, 2018). Dekat atau jauhnya lokasi (jarak) bisa dihitung melalui salah satu dari besaran jarak yang telah ditentukan yakni jarak *Euclidean* & jarak *Minkowski*. Namun dalam penerapannya seringkali digunakan jarak *Euclidean* karena memiliki tingkat akurasi dan juga *productivity* yang tinggi (Asiyah dan Fithriasari, 2016). Rumus jarak *Euclidean* adalah sebagai berikut:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{p} (x_{ip} - x_{jp})^2} \qquad (2.2)$$

$d(x_i, x_j)$ merupakan jarak *euclidean* dari data uji dengan data latih sedangkan $x_{ip}\ dan\ x_{jp}$ merupakan data *testing* ke i dan data *training* ke j.

Rocchio

Manning dkk. (dalam Afriza dan Adisantoso, 2018) menjelaskan bahwa klasifikasi Rocchio merupakan klasifikasi dengan bentuk linear, bahwa dokumen dalam suatu kelas yang sama tidak akan terjadi *overlap* dengan kelas yang berbeda. Nilai *centroid* pada *Rocchio* diperoleh dengan mencari rata – rata *vector* pada tiap-tiap dokumen data *training* untuk masing – masing kelas. *Centroid* tiap kelas dihitung dengan rumus sebagai berikut:

$$\vec{u}(c) = \frac{1}{dc} \sum_{d \in Dc} \vec{v}\ (d) \qquad (2.3)$$

Dengan dc merupakan total dokumen di kelas c, $\vec{v}(d)$ adalah *vector* kata – kata dalam kelas c, dan $\vec{u}(c)$ adalah *centroid* masing – masing kelas. Salah satu cara untuk menentukan kedekatan data uji

dengan data latih adalah dengan menggunakan persamaan *cosine similarity* antar kedua titik c1 dan d2. Rumus *cosine similarity* adalah sebagai berikut:

$$sim(c1, d2) = \frac{\vec{v}(c1).\vec{v}(d2)}{\sqrt{c1^2.d2^2}} \quad (2.4)$$

Dengan $\vec{v}(c1)$ merupakan nilai *vector centroid* untuk tiap *class* dan $\vec{v}(d2)$ merupakan nilai *vector* data uji, sedangkan $c1^2$ merupakan nilai panjang *vector centroid* dan $d2^2$ merupakan nilai panjang *vector* data uji.

HASIL DAN PEMBAHASAN

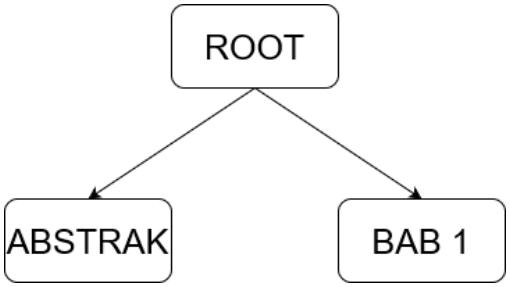
Data

Data yang dikumpulkan berasal dari abstrak dan bab 1 tugas akhir Fakultas Teknik Universitas Muhammadiyah Jember yang ada di *repository* Universitas Muhammadiyah Jember. Teknik pengumpulan data menggunakan studi literatur dengan cara mengambil abstrak dan bab 1 tugas akhir mahasiswa Fakultas Teknik saja. Dari hasil studi literatur didapatkan data abstrak tugas akhir mahasiswa Fakultas Teknik berjumlah 150 data tugas akhir, dengan masing-masing kelas atau prodi memiliki data berjumlah 30 *record*.

Hasil Klasifikasi

Klasifikasi dilakukan menggunakan algoritma *KNN* dan *Rocchio*, untuk pengujianya menggunakan *Kfold* dengan nilai K yaitu 2, 3, 5, 6, 10.

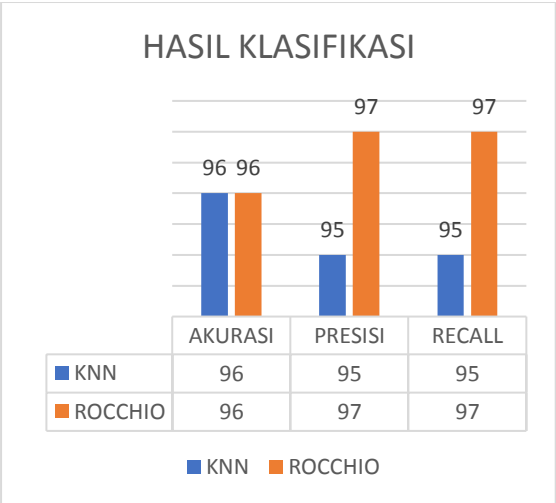
Partisi Data Abstrak Dan Bab 1



Gambar 1. Tree Data Tugas Akhir.

Berdasarkan gambar di atas, pada data tugas akhir dibagi menjadi 2 skenario yang pertama abstrak yang berkontribusi sebanyak 40% dan bab 1 sebanyak 60% dan skenario yang kedua yaitu abstrak berkontribusi sebanyak 30% dan bab1 sebanyak 70%. Bab 1 memiliki porsi yang lebih besar karena memiliki isi yang lebih banyak dari pada abstrak.

Berdasarkan hasil klasifikasi dengan menggunakan data Abstrak dan Bab1 Tugas Akhir untuk algoritma *K-Nearest Neighbor (KNN)* didapatkan hasil akurasi sebesar 96%, presisi 95%, dan *recall* 95% pada *Kfold* 6. Sedangkan pada algoritma *Rocchio* didapatkan hasil akurasi sebesar 96%, presisi 97%, dan *recall* 97% dengan menggunakan *Kfold* 6 .



KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Hasil akurasi paling tinggi dalam klasifikasi Tugas Akhir dengan abstrak dan bab 1 menggunakan algoritma KNN didapatkan hasil sebesar 96% pada *Kfold* 6, dan akurasi paling tinggi algoritma *Rocchio* yaitu sebesar 96% pada *Kfold* 6.
2. Hasil presisi paling tinggi dalam klasifikasi Tugas Akhir dengan abstrak dan bab 1 menggunakan algoritma KNN didapatkan hasil sebesar 95% pada *Kfold* 3, 6, dan 10, Untuk hasil presisi paling tinggi dengan menggunakan algoritma *Rocchio* didapatkan hasil sebesar 97% pada *Kfold* 6.
3. Hasil *recall* paling tinggi dalam klasifikasi Tugas Akhir dengan abstrak dan bab 1 menggunakan algoritma KNN didapatkan hasil sebesar 95% pada *Kfold* 6, Untuk hasil *recall* paling tinggi dengan menggunakan algoritma *Rocchio* didapatkan hasil sebesar 97% pada *Kfold* 6 .
4. Hasil pembagian data yang lebih baik yaitu dengan menggunakan 30% kontribusi data abstrak dan 70% kontribusi data bab 1.
5. Secara keseluruhan kinerja dari dua algoritma ini dalam mengklasifikasikan tugas akhir berdasarkan abstrak dan bab 1 sudah sangat baik, terbukti dengan menghasilkan nilai akurasi, presisi, dan recall yang cukup tinggi, tetapi Rocchio mampu menghasilkan nilai presisi dan *recall* lebih tinggi dari pada KNN.

Saran

Berdasarkan penelitian yang telah dilakukan, beberapa saran yang dapat dikembangkan untuk penelitian selanjutnya adalah sebagai berikut:

1. Perlu dikembangkan menggunakan jumlah data yang lebih banyak dan tidak hanya mencakup abstrak dan bab 1 saja.
2. Untuk penelitian selanjutnya bisa menggunakan algoritma lainnya.

DAFTAR PUSTAKA

Afriza, A & Adisantoso, J. 2018. "Metode Klasifikasi Rocchio untuk Analisis Hoax". *Jurnal Ilmu Komputer Agri-Informatika* Vol. 5 No.1. Institut Pertanian Bogor, Bogor. Tersedia di http://journal.ipb.ac.id/index.php/jika

Akromunnisa, K dan Hidayat, R 2019. "Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan *K-Nearest Neighbour* ". *JISKa*, Vol. 4, No. 1, Mei 2019, Pp. 69-75. ISSN : 2527-5836. UIN Sunan Kalijaga, Yogyakarta.

Alfiyanti, D.Y., Ratnawati, E.D., dan Anam, S. 2019. "Klasifikasi Fungsi Senyawa Aktif Data Berdasarkan Kode *Simplefied Molecular Input Line Entry System (SMILES)* Menggunakan Metode *Modified K-Nearest Neighbour". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol.

3, No. 4, April 2019, hlm. 3244-3251. e-ISSN: 2548-964X. Universitas Brawijaya, Malang.

Asiyah, S.N. 2016. "Klasifikasi Berita Online Menggunakan Metode *Support Vector Machine* dan *K-Nearest Neighbor*". Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember.

Asril, H., Mustakim., Kamila, I 2019. "Klasifikasi Dokumen Tugas Akhir Berbasis *Text Mining* menggunakan Metode *Naïve Bayes Classifier* dan *K-Nearest Neighbor*". *Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 11*, ISSN (Printed) : 2579-7271. Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau, Pekanbaru.

Feldman, R dan Sanger, J. 2007. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructered Data*. New York: Cambridge University Press.

ISO 214:1976. Documentation — Abstracts for publications and documentation [online]. Tersedia di https://www.iso.org/standard/4084.html. Diakses Januari 2020.

Jupyter Notebook [online]. Tersedia di https://jupyter.org/. Diakses 20 Maret 2020.

Lestari, P.A., Maskur., dan Hayatin Nur 2019. "Klasifikasi Text Berbasis Ontologi Untuk Dokumen Tugas Akhir Berbahasa Indonesia". *REPOSITORI*, Vol. 1, No. 2, Desember 2019, Pp. 79-86 ISSN : 2714-7975 E-ISSN : 2716-1382. Universitas Muhammadiyah Malang, Malang.

Kamus Besar Bahasa Indonesia (KBBI) [online]. Tersedia di https://kbbi.web.id/abstrak. Diakses 22 Nopember 2019.

Manning, C., Raghavan, P., dan Schutze, H. 2009. *Introduction to Information Retrieval*, Cambridge University Press.

Melita, R., Suseno, B.H., dan Dirjam, T. 2018. "Penerapan Metode *Term Frequency Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity* Pada Sistem temu Kembali Informasi Untuk Mengetahui Syarah *Hadist* Berbasis Web (Studi Kasus: *Syarah Umdatil Akram*). *JURNAL TEKNIK INFORMATIKA* Vol. 11 No. 2, Oktober 2018. Universitas Islam Negeri Syarif Hidayatullah, Jakarta.

Moldagulova, A. dan Sulaiman, R.B. 2017. "Using KNN Algorithm for Classification of Textual Documents". *International Conference on Information Technology (ICIT).* University Tenaga Nasional, Malaysia.

Muslimah, N., Indriati., dan Wihandika, C.R. 2019. "Klasifikasi Film Berdasarkan Sinopsis Dengan Menggunakan *Improved K-Nearest Neighbour*". *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol. 3, No. 1, Januari 2019, hlm. 196-204. e-ISSN: 2548-964X. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Naffisah, S.M. dan Surjandari, I. 2014. "Pengguanaan *Text Mining* Pada Analisis Sentimen Masyarakat Terhadap Perubahan Harga Bahan Pokok Melalui Twitter". Fakultas Teknik, Universitas Indonesia, Depok.

Puspitasari, A.A., Santoso Edy., dan Indriati 2018. "Klasifikasi dokumen tumbuhan obat menggunakan metode *improved k-Nearest Neighbour". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol. 2, No. 2, Februari 2018, hlm. 486-492. e-ISSN: 2548-964X. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Putri, K.E dan Setiadi Tedy. 2014 "Penerapan *Text Mining* Pada Sistem Klasifikasi Email Spam Menggunakan *Naive Bayes*". *Jurnal Sarjana Teknik Informatika,* Volume 2 Nomor 3, Oktober 2014. e-ISSN: 2338-5197. Universitas Ahmad Dahlan, Yogyakarta.

Repository Universitas Muhammadiyah Jember [online]. Tersedia di http://repository.unmuhjember.ac.id/view/divisions/information/. Diakses Nopember 2019.

Wibawa, W,D., Nasrun, M., dan Setianingsih, C. 2018. "*Sentiment Analysis On User Satisfacation Level Of Cellular Data Service Using The K-Nearest Neighbor (K-NN) Algorithm*". *The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC).* Telkom University, Indonesia.

Yusra, Olivita, D, dan Vitriani, Y. 2016. " Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode *Naive Bayes Classifier* dan *K-Nearest Neighbour". Jurnal Sains, Teknologi dan Industri*, Vol. 14, No. 1, Desember 2016, pp. 79 – 85. ISSN 1693-2390 print/ISSN 2407-0939 online. Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau.

Zakky. 2018. Pengertian Abstrak Menurut Para Ahli, KBBI dan Secara Umum [online]. Tersedia di https://www.zonareferensi.com/pengertian-abstrak/. Diakses Januari 2020.