ID JTIIK 1323

by 1323 Jtiik

Submission date: 05-Dec-2018 03:49PM (UTC+0700)
Submission ID: 1051059494
File name: 1323-4251-1-SM.docx (154.67K)
Word count: 2476
Character count: 16039

Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)
Vol. x, No. x, April 2014, hlm. x-x
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017

DOI: 10.25126/jtiik...
p-ISSN: 2355-7699
e-ISSN: 2528-6579

ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI SUMBER BELAJAR BERBASIS TEKS PADA MATA PELAJARAN PRODUKTIF DI SMK RUMPUN TEKNOLOGI INFORMASI DAN KOMUNIKASI

(Naskah masuk: dd mmm yyyy, diterima untuk diterbitkan: dd mmm yyyy)

Abstrak

Salah satu komponen esensial dalam kegiatan pembelajaran di Sekolah Menengah Kejuruan Rumpun Teknologi Informasi dan Komunikasi (SMK TIK) adalah ketersediaan sumber belajar mata pelajaran produktif. Tujuan penelitian adalah untuk mendeskripsikan hasil klasifikasi dan evaluasi kualitas klasifikasi sumber belajar berbasis teks dengan menggunakan Algoritma Naïve Bayes. Tahapan penelitian yang dilakukan adalah pengoleksian data set, pemrosesan awal dengan *text mining*, pembobotan Tf-Idf, pengklasifikasian Naïve Bayes, dan evaluasi akurasi. Pengklasifikasian teks menghasilkan sembilan kelompok mata pelajaran produktif dan pengujian menghasilkan nilai akurasi tertinggi 81,48%, sedangkan nilai akurasi terendah sebesar 79,63%.

Kata kunci: *sumber belajar, text mining, algoritma naïve bayes, sekolah menengah kejuruan, teknologi informasi dan komunikasi*

NAÏVE BAYES ALGORITHM FOR TEXT BASED LEARNING RESOURCES CLASSIFICATION IN PRODUCTIVE SUBJECT AT INFORMATION AND COMMUNICATION TECHNOLOGY VOCATIONAL HIGH SCHOOL

Abstract

The availability of learning resources for productive subjects is one of the essential components in learning activi for Vocational High Schools, especially for Information and Communication Technology competence field. The purpose of the study was to cribe the results of the classification and classification quality evaluation of text-based learning sources using the Naïve Bayes Algorithm. The stages of the research carried out are collecting data sets, pre-processing with text mining, Tf-Idf weighting, Naïve Bayes classifying, and accuracy evaluation. Text classification results shows that there are nine productive subject groups and based on testing results shows that the highest accuracy value was 81,48%, while the lowest accuracy value was 79,63%.

Keywords: *learning resource, text mining, naïve bayes algorithm, vocational high school, inforaton and communication technology*

1. PENDAHULUAN

Kebutuhan calon pekerja bidang Teknologi Informasi dan Komunikasi (TIK) lulusan Sekolah Menengah Kejuruan (SMK) semakin meningkat dengan bukti bahwa Direktorat Pembina SMK menambah satu rumpun keilmuan SMK TIK di tahun 2018, yaitu jurusan Sistem Informatika, Jaringan, dan Aplikasi (SIJA) (Kurikulum, 2017). Artinya, jurusan Sistem Informatika dapat melengkapi kebutuhan penyedia lapangan kerja bidang TIK untuk lulusan SMK yang semula hanya terbagi menjadi tiga jurusan, yaitu jurusan Rekayasa Perangkat Lunak (RPL), Teknik Komputer dan Jaringan (TKJ), dan Multimedia (MM).

Penciptaan calon pekerja bidang TIK yang memiliki relevansi dengan kebutuhan penyedia lapangan kerja harus didukung oleh institusi SMK dengan cara menyajikan kegiatan pembelajaran yang bermutu. Perkembangan teknologi informasi dan komunikasi (TIK) dewasa ini memunculkan sumber belajar yang mudah diakses sehingga dapat memperkaya interaksi antara guru dan siswa selama ses pembelajaran. Media internet atau online merupakan salah satu bentuk sumber belajar media elektronik yang dapat digunakan oleh siswa dan

guru melalui jaringan internet. Salah satu bentuk dokumen yang menggunakan media online adalah halaman web berformat .html (*Hypertext Markup Language*). Media *online* memiliki beberapa keunggulan, yaitu, informasinya senantiasa *up to date* (senantiasa terbaru), informasinya bersifat *real time*, dan informasinya bersifat praktis.

Ketersedian sumber belajar berbentuk halaman web dan berformat .html sangat melimpah. Hal ini dapat dibuktikan melalui *Google Search Engine* dengan kata kunci "Sistem Operasi untuk SMK RPL" dan menghasilkan lebih dari 70.000 dokumen berformat .html. Apabila dokumen tersebut tidak diatur berdasarkan isi esensial setiap materi mata pelajaran, maka berpotensi akan menyulitkan guru dan siswa dalam kegiatan memilih materi pelajaran yang relevan dengan kebutuhan pembelajaran. Pengaturan dokumen teks dapat dilakukan dengan klasifikasi teks secara otomatis pada masing-masing dokumen tersebut berdasarkan kriteria atau ciri esensial setiap mata pelajaran produktif di SMK TIK.

Pengklasifikasian teks adalah sebuah area dimana algoritma klasifikasi digunakan pada dokumen-dokumen teks (Sriram et al., 2010). Pada saat ini terdapat bermacam-macam algoritma klasifikasi yang dapat digunakan untuk klasifikasi dokumen teks. Salah satunya adalah algoritma klasifikasi Naïve Bayes. Naïve Bayes merupakan algoritma klasifikasi sederhana yang menerapkan teorema Bayes dengan menganggap semua fitur saling tidak berhubungan. Penggunaan algoritma ini menggunakan keseluruhan probabilitas, yaitu probabilitas dokumen terhadap kategori (*prior*). Kemudian teks akan terkategori berdasarkan probabilitas maksimumnya (*posterior*). Dengan kata lain, algoritma ini mengasumsikan bahwa ada atau tidaknya fitur tertentu dari kelas tidak berhubungan dengan ada tidaknya fitur yang lain (Yuan, 2010).

Algoritma klasifikasi Naive Bayes untuk klasifikasi dokumen teks telah dilakukan oleh beberapa peneliti. Hamzah (2012) meneliti sejauh mana kinerja algoritma klasifikasi Naive Bayes dalam kategorisasi teks yang berupa teks berita dan teks akademis berupa abstrak akademis dari berbagai disiplin ilmu. Penelitian tersebut menggunakan 1000 dokumen berita dan 450 dokumen abstrak akademik. Hasil penelitian menunjukkan pada dokumen berita akurasi maksimal dicapai 91% sedangkan pada dokumen abstrak akademik 82%. Seleksi kata dengan minimal muncul pada 4 atau 5 dokumen memberikan akurasi yang paling tinggi. Penelitian oleh Chandra, Indrawan, & Sukajaya (2016) menggunakan algoritma klasifikasi Naive Bayes untuk melakukan klasifikasi dokumen menjadi berita politik, ekonomi, news, edukasi, kesehatan, travel, dan olahraga pada portal www.kompas.com. Hasil penelitian tersebut menunjukkan hasil akurasi sebesar 78.66% untuk data uji berita ekonomi, news, edukasi, kesehatan, olahraga, entertainment, dan lain-lain dalam Bahasa Indonesia.

Perlu ada penelitian terkait implementasi algoritma Naïve Bayes untuk mengklasifikasikan sumber belajar. Sumber belajar merupakan segala sumber daya yang bisa digunakan oleh siswa untuk mendukung proses pembelajarannya (Hillier, 2005).

Sumber belajar utama yang dapat digunakan adalah berupa sumber belajar elektronik yang didukung dengan teknologi jaringan komputer atau internet. Materi pembelajaran berkualitas yang tersedia di internet dapat dijadikan sebagai sumber belajar primer untuk siswa (Hillier, 2005). Akses ke internet sangat direkomendasikan di dalam pembelajaran rumpun TIK, karena akan membantu siswa untuk memperoleh sosok materi yang dipelajari dan memperkaya referensi siswa untuk memecahkan masalah (Malmi & Korhonen, 2008).

Sumber belajar di ineternet banyak disajikan dalam bentuk teks, sehingga metode pencarian pola teks dalam proses pengklasifikasian yang dapat digunakan adalah *text mining*. *Text mining* adalah suatu penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu tidak diketahui sebelumnya atau menemukan suatu informasi yang tersirat secara implisit yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Fieldman et al., 2006). *Text preprocessing* adalah tahap awal dari *text mining* untuk merubah data yang tidak terstruktur menjadi data terstruktur (Torunoğlu et al., 2011). Proses yang dilakukan pada tahap ini seperti *casefolding*, *cleaning term* (pembersihan kata), *tokenizing*, *filtering* (*stopword removal*), dan *stemming* (Agusta, Kristen, & Wacana, 2009; Nurjanah, Hamdani, & Astuti, 2013).

2. METODOLOGI

Tahapan penelitian yang telah dilakukan terdiri dari kegiatan pengoleksian data set, pemrosesan awal dengan *text mining*, pembobotan Tf-Idf, pengklasifikasian Naïve Bayes, dan evaluasi akurasi. Pada tahap pengoleksian dataset, peneliti melakukan pengumpulan data berupa dokumen teks sumber belajar. Data yang digunakan pada penelitian ini adalah dokumen teks atau artikel teks yang diperoleh dari website yang berisikan materi-materi mata pelajaran produktif dari SMK Rumpun Teknologi Informasi dan Komunikasi. Dokumen teks yang diambil hanya materi yang terbagi pada Rekayasa Perangkat Lunak (RPL), Teknik Komputer dan Jaringan (TKJ), Multimedia (MM), dan Sistem Informatika, Jaringan, dan Aplikasi. Data diambil dengan menggunakan *crawling website* selama 2 bulan, yaitu antara bulan Maret sampai April tahun 2018. Data dikelompokkan menjadi sembilan kelas, yaitu K1 = Animasi 2D, K2 = Animasi 3D, K3 = Desain Multimedia, K4 = Basis Data, K5 = Pemodelan Perangkat Lunak, K6 = Pemrograman

Dasar, K7 = Komputer Terapan Jaringan, K8 = Komunikasi Data, dan K9 = Sistem Operasi.

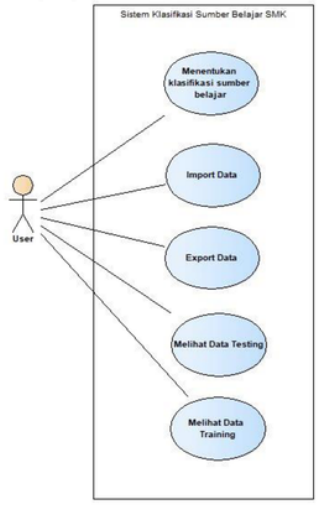
Kemudian dilakukan tahap perancangan sistem yang terdiri tiga bagian, yaitu tahap preproses, tahap pembobotan kata (*tf-idf*), dan klasifikasi dokumen menggunakan Algoritma Naïve Bayes. Tahap praproses dalam penelitian ini terdiri empat bagian, yaitu pemenggalan kata (*tokening term*), pembersihan kata (*cleaning term*), penghapusan *stopword stopword removal*), dan *stemming* (kata dasar). Setelah dapat kata-kata dalam dokumen teks, selanjutnya menghitung bobot kemunculan kata dengan teknik *Term Frequency Inverse Document Frequency* (Tf–Idf). Tf-Idf adalah konsep pembobotan *term* pada sebuah dokumen. Pembototan Tf–Idf diterapkan pada lingkup kalimat, maka sebuah kalimat akan diberlakukan sebagai dokumen. Selanjutnya data bobot Tf-Idf diklasifikasikan menggunakan Algoritma Naïve Bayes. Pada tahap pengujian peneliti melakukan pengujian dengan menghitung akurasi dari evaluasi Algoritma Naïve Bayes. Dataset yang nanti didapatkan akan dibagi menjadi *data training* dan *data testing*. Standar yang digunakan dalam melakukan evaluasi yaitu dengan mengukur kinerja sistem yaitu akurasi. Akurasi merupakan seberapa dekat suatu angka hasil pengukuran terhadap angka sebenarnya (*true value* atau *reference value*). Dalam penelitian ini pengujian akurasi dilakukan untuk mengetahui performa dari sistem dalam memberikan kesimpulan prediksi. Perhitungan nilai akurasi dilakukan dengan menggunakan Persamaan (1).

$$Nilai\ Akurasi = \frac{jumlah\ data\ akurat}{jumlah\ seluruh\ data} \times 100\% \quad (1)$$

3. HASIL

Analisis kebutuhan sistem pada penelitian ini dilakukan untuk mengidentifikasi permasalah yang terjadi di 2 SMK TIK di Kota Malang untuk selanjutnya dapat menghasilkan perancangan sistem yang sesuai dengan kebutuhan SMK TIK. Sistem disini diposisikan sebagai test-bed implementasi Algoritma Naïve Bayes. Setelah melakukan wawancara dengan pihak sebanyak 2 SMK TIK yang ada di Kota Malang (SMK 12 Malang dan SMK Widyagama Malang), dapat disimpulkan bahwa siswa masih merasa kesulitan dalam mencari materi-materi di website. Siswa juga tidak tahu materi tersebut masuk dalam bidang keminatan yang mana dalam SMK TIK. SMK TIK memiliki beberapa mata pelajaran yang terbagi dalam tiga keminatan, yaitu Rekayasa Perangkat Lunak (RPL), Teknik Komputer dan Jaringan (TKJ), Multimedia (MM), dan Sistem Informatika, Jaringan, dan Aplikasi (SIJA).

Terdapat lima (5) *Use Case* yang menggambarkan fungsi sistem dalam penelitian ini, yaitu Menghitung proses klasifikasi Naïve Bayes; Import Data; Export Data; Melihat Data Training, dan Melihat Data Testing. Actor pada aplikasi terdapat dua orang yaitu siswa atau guru. Lima (5) fungsi yang didefinisikan tersebut merupakan jawaban dari kebutuhan yang muncul dari hasil identifikasi masalah. Gambar 1 mendefinisikan *Use Case* yang terdapat pada sistem.



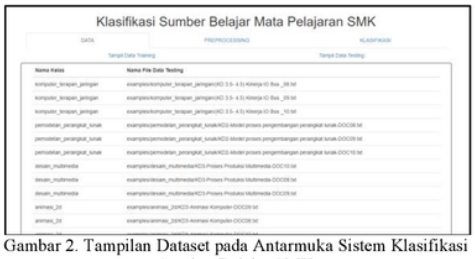
Gambar 1. Use Case Sistem Klasifikasi Sumber Belajar SMK

Dataset yang berhasil diperoleh selama 2 bulan adalah sebanyak 235 data dalam bentuk file notepad (txt). Dataset dibagi sesuai kelasnya masing-masing, yaitu K1 sebanyak 32 data, K2 sebanyak 32 data, K3 sebanyak 32 data, K4 sebanyak 21 data, K5 sebanyak 21 data, dan K6 sebanyak 21 data, K7 sebanyak 32 data, K8 sebanyak 22 data, dan K9 sebanyak 22 data. Dataset dijadikan sebagai *data training* dan *data testing* untuk Algoritma Naïve Bayes. *Data training* dan *data testing* ini adalah data yang berbeda-beda.

Pengujian dilakukan sebanyak 4 kali dengan perbandingan *data training* dan *data testing* yang sama. *Data testing* pertama menggunakan 20 data dengan jumlah tiap kelas. *Data testing* pertama menggunakan 36 data dari jumlah semua kelas. *Data testing* kedua menggunakan 54 data dari jumlah semua kelas. *Data testing* ketiga menggunakan 72 data dari jumlah semua kelas. *Data testing* keempat menggunakan 120 data dari jumlah semua kelas. *Data training* yang digunakan berjumlah 55 data. Jumlah data training dibuat tetap jumlahnya dalam setiap percobaan karena untuk menguji apakah jika menggunakan data testing yang berbeda akan mempengaruhi hasil akurasi prediksi sistem. Gambar 2 menunjukkan bahwa dataset dapat ditampilkan pada sistem.

Semua data training menggunakan data sebanyak 55 data. Percobaan ke-1 menggunakan data testing sebanyak 36 data. Dari hasil percobaan

ke-1 diperoleh nilai akurasi sebesar 80,56%. Sebanyak 29 data yang tepat atau sesuai dengan data target. Sedangkan 7 data dari sistem tidak sesuai dengan data target. Percobaan ke-2 menggunakan data testing sebanyak 54 data. Dari hasil percobaan ke-2 diperoleh nilai akurasi sebesar 79,63%. Sebanyak 43 data yang tepat atau sesuai dengan data target. Sedangkan 11 data dari sistem tidak sesuai dengan data target. Percobaan ke-3 menggunakan data testing sebanyak 72 data. Dari hasil percobaan ke-3 diperoleh nilai akurasi sebesar 80,56%. Sebanyak 58 data yang tepat atau sesuai dengan data target. Sedangkan 14 data dari sistem tidak sesuai dengan data target. Percoabaan ke-4 menggunakan data testing sebanyak 120 data. Dari hasil percobaan ke-4 diperoleh nilai akurasi sebesar 81,48%. Sebanyak 98 data yang tepat atau sesuai dengan data target. Sedangkan 22 data dari sistem tidak sesuai dengan data target. Dari hasil pengujian dapat diketahui nilai akurasi untuk setiap percobaan. Nilai akurasi tertinggi dari semua percobaan yaitu 81,48%, sedangkan nilai akurasi terendah dari semua percobaan yaitu 79,63%. Perbandingan nilai akurasi untuk setiap percobaan disajikan pada Tabel 1.



Gambar 2. Tampilan Dataset pada Antarmuka Sistem Klasifikasi Sumber Belajar SMK

Tabel 1. Hasil Perbandingan Akurasi Pengujian

Pengujian Ke-	Jumlah Data Testing	Jumlah Data Yang Tepat	Jumlah Data Yang Salah	Akurasi
1	36 Data	29 Data	7 Data	80,56%.
2	54 Data	43 Data	11 Data	79,63%
3	72 Data	58 Data	14 Data	80,56%
4	120 Data	98 Data	22 Data	81.48%

4. SIMPULAN

Kelompok data sumber belajar yang diperoleh sebanyak 235 data yang sudah terklasifikasi berdasarkan masing-masing kelas. Data akan dikelompokkan menjadi 9 kelas yaitu K1 = Animasi 2D, K2 = Animasi 3D, K3 = Desain Multimedia, K4 = Basis Data, K5 = Pemodelan Perangkat Lunak, K6 = Pemrograman Dasar, K7 = Komputer Terapan Jaringan, K8 = Komunikasi Data, dan K9 = Sistem Operasi. Data testing pertama menggunakan 36 data dengan jumlah tiap kelas. *Data testing* kedua menggunakan 54 data dengan jumlah tiap kelas. *Data testing* ketiga menggunakan 72 data dengan jumlah tiap kelas. *Data testing* keempat menggunakan 120 data dengan jumlah tiap kelas berbeda. *Data training* yang digunakan berjumlah 55 data. Pengujian pada data sumber belajar menggunakan algoritma Naive Bayes menghasilkan nilai akurasi tertinggi 81,48, sedangkan nilai akurasi terendah sebesar 79,63%.

Penelitian selanjutnya perlu mempertimbangkan pemilihan *data training*, karena pola *data training* akan dijadikan sebagai rule untuk menentukan kelas pada data testing dan memperngaruhi nilai akurasi. Penelitian selanjutnya juga perlu mempertimbangkan kompleksitas/ketelitian pada tahap *pre-processing*, karena kualitas data yang dihasilkan dari tahap *pre-processing* juga bisa menentukan nilai akurasi.

5. DAFTAR RUJUKAN

AGUSTA, L., KRISTEN, U., & WACANA, S. 2009. Perbandingan Algoritma *Stemming* Porter Dengan Algoritma Nazief & Adriani Untuk *Stemming* Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika 2009*, (KNS&I09-036), pp.196–201.

ALLAN, J., KUMARAN, G., & STORAGE, H.I. 2003. *Stemming in the Language Modeling Framework*. (June), pp.455–456.

CHANDRA, D. N., UTAMA, I.G.B.B., & SUKAJAYA, I.N. 2016. Klasifikasi Berita Lokal Radar Malang menggunakan Metode Naive Bayes dengan Fitur N-Gram. *Jurnal Ilmiah Teknologi dan Informasi ASIA (JITIKA)*. 10(1): hlm. 11-19

FELDMAN, R., & JAMES, S. 2006. *The Text mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

HAMZAH, A. 2012. Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokkan Teks Berita dan Abstract Akademis. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*. ISSN : 1979-9111, Vol. 3.

HILLIER, Y. 2005. *Reflective Teaching in Further and Adult Education: Second Edition*. London: Continuum.

KURIKULUM. 2017. *Kompetensi Inti dan Kompetensi Dasar (KI & KD) SMK/MAK*, (*Online*), (http://psmk.kemdikbud.go.id/kikd2017), diakses 1 Februari 2017.

MALMI, L. & KORHONEN, A. 2008. Active Learning and Examination Methods in a Data Structures and Algorithms Course. Dalam Caspersen, E. M. & Kolling, M. (Eds.). *Reflections on the Teaching of*

Programming: Methods and Implementions (hlm. 210-227). Berlin: Springer.

NURJANAH, M., HAMDANI, & ASTUTI, I. F. 2013. Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) untuk *Text mining*. *Jurnal Informatika Mulawarman*. 8 (3): hlm. 110-113.

TORUNOĞLU, D., ÇAKIRMAN, E., GANIZ, M.C., AKYOKUŞ, S., & GÜRBÜZ, M.Z. 2011. Analysis Of Preprocessing Methods On Classification of Turkish Texts. *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications*, pp.112–117.

YUAN, L. 2010. An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing. *Proceedings of the Third International Symposium on Computer Science and Computational Technology* (ISCSCT '10), Jiaozuo, P. R. China, 14-15, August 2010, pp. 267-269.

ID JTIIK 1323

ORIGINALITY REPORT

26%	25%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	media.neliti.com Internet Source	4%
2	es.scribd.com Internet Source	2%
3	repository.usu.ac.id Internet Source	2%
4	etheses.uin-malang.ac.id Internet Source	2%
5	ijasret.com Internet Source	1%
6	jurnaleeccis.ub.ac.id Internet Source	1%
7	openaccess.dogus.edu.tr Internet Source	1%
8	cipto-suparno.blogspot.com Internet Source	1%
9	lp3m.asia.ac.id Internet Source	1%

10	jurnal.umk.ac.id Internet Source	1%
11	si.its.ac.id Internet Source	1%
12	ie.akprind.ac.id Internet Source	1%
13	journal.unnes.ac.id Internet Source	1%
14	filkom.ub.ac.id Internet Source	1%
15	Eva Y. Puspaningrum, Lailly S. Qolby, Yisti V. Via. "OPTIMASI JARINGAN SARAF TIRUAN UNTUK DIAGNOSIS PENYAKIT DIABETES INDIAN PIMA", Teknologi, 2016 Publication	1%
16	jurnal.uns.ac.id Internet Source	1%
17	anzdoc.com Internet Source	1%
18	Agung Nugroho. "Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstrasi Fitur N-Gram", J-SAKTI (Jurnal Sains Komputer dan Informatika), 2018 Publication	1%

19	www.education.umd.edu Internet Source	1%
20	jtiik.ub.ac.id Internet Source	1%
21	student.brighton.ac.uk Internet Source	1%
22	www.theseus.fi Internet Source	<1%
23	modulkomputer.com Internet Source	<1%
24	Mohammad Sadeghi, Jesús Vegas. "How well does Google work with Persian documents?", Journal of Information Science, 2016 Publication	<1%
25	sentrin.filkom.ub.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		