Wordcloud

by Hap1 Hap2

Submission date: 31-Dec-2020 08:54PM (UTC+0700)
Submission ID: 1482259041
File name: blindreviewjurnal_wordcloud.docx (6.61M)
Word count: 2980
Character count: 18797

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

WORD CLOUD DATA SCIENCE UKSW

Abstract- Suatu universitas mempunyai data terkait kegiatan publikasi riset dosen yang dapat tergolong *big data*. Untuk dapat memperoleh informasi riset unggulan dari tiap dosen ataupun tiap fakultas tidak dapat dilakukan secara manual. Oleh karena itu, artikel ini memberikan hasil penelitian dalam memberikan informasi yang diperlukan tersebut dengan menggunakan *machine learning* khususnya untuk pengolahan data teks. Pada penelitian ini ditunjukkan *data science* terkait analisa data riset para dosen di Universitas Kristen Satya Wacana (UKSW) berdasarkan data dari Google Scholar. Metode yang digunakan adalah metode *Machine Learning* dalam membaca teks yang ada pada data yang dihimpun. Setelah itu dilakukan deteksi atau analisa dengan menggunakan *Word cloud*. *Word cloud* merupakan himpunan kata yang dikoleksi dengan *machine learning* sehingga tingkat keseringkan kata yang muncul dalam Google Scholar akan menentukan dominasi munculnya kata dalam *Word cloud*. Algoritma diuji cobakan pada data suatu program studi di UKSW. Setelah mendapatkan konfirmasi hasil yang diperoleh, algoritma diimplementasikan pada setiap dosen di UKSW. Untuk mengenali keunggulan tiap fakultas, data diklasifikasi pada tiap fakultas. Dengan hasil ini dapat diperoleh keunggulan tiap dosen, tiap program studi maupun tiap fakultas di UKSW. Sedangkan kata unggulan yang muncul untuk seluruh data riset di UKSW yang menonjol adalah pada kata "Student","Informasi","Pendidikan","Development","WEB" dan "Evaluasi". Dapat disimpulkan bahwa di UKSW mempunyai unggulan penelitian dan riset dengan mengangkat tema atau dari kata tersebut.

Kata Kunci: *machine learning, Word cloud, nlp, google scholar*

1. Pendahuluan

Upaya pendataan terhadap hasil -hasil penelitian pada suatu universitas seringkali mengalami kesulitan karena data tidak terdokumentasi scara terpadu. Selain itu, untuk menelusuri kompetensi penelitian dosen tidak dapat dilakukan dengan menelusuri dokumen secara manual untuk dapat membuat kesimpulan global. Salah satu teknik yang digunakan untuk melakukan otomatisasi dokumen digital dalam riset adalah dengan melakukan klasifikasi teks [1][2]. Pada literatur tersebut teknik-teknik klasifikasi dibandingkan untuk mempelajari akurasi dari beberapa teknik klasifikasi dalam klasifikasi data berupa teks. Hal ini dianggap kurang mudah untuk dibaca. Untuk itulah pada penelitian ini digunakan *Word cloud* untuk dapat memberikan hasil klasifikasi dengan lebih mudah dimana cara ini pernah juga dilakukan oleh author yang lain dalam melakukan analisa teks [3] [4]. Penggunaan *Word cloud* program dengan machine learning merupakan salah satu cara cepat yang dapat dilakukan dalam menelusuri dokumen yang terdigitalisasi pada google scholar. Hasil yang diperoleh akan dapat membantu pimpinan dalam mengambil kebijakan dalam pengembangan penelitian dosen dan mahasiswa.

Pembuatan Word cloud ini dilakukan dengan menggunakan algoritma machine learning. *Machine Learning* (ML) telah digunakan pada berbagai aplikasi dalam pengolahan big data seperti pada pembuatan kecerdasan buatan, pengolahan data COVID-19 pada laju kematian di negara Korea Selatan [5], pengolahan data diabetes [6] dan berbagai aplikasi lainnya. Demikian pula pada pengolahan data yang ada di internet, ML mampu mengidentifikasi karakteristik pengguna internet. Hal ini menyebabkan kemudahan sistem dalam mencermati karakteristik data dari berbagai sumber untuk memberikan kesimpulan tertentu.

Machine Learning (ML) diterjemahkan sebagai mesin pembelajaran pada artikel ini adalah bagian dari ilmu pendataan misalkan studi tentang analisa teks dalam jumlah besar (*big data*)[7],pengenalan pola (*pattern recognition*) dan teori komputasi dalam intelegensi buatan (*artificial intelegence*) [8] . ML melakukan konstruksi dan studi algoritma yang mempelajari dari dan membuat prediksi data. Algoritma demikian membangun model dari contoh input yang ada untuk membuat keputusan atau prediksi yang dibangun data dimana program statis pada umumnya hanya mengikuti instruksi program. Seringkali pula ML berkaitan dan beririsan dengan statistika komputasi dimana statistika komputasi juga mengkhususkan pada pembuatan prediksi yang terkait erat dengan optimasi matematika yang membangun teori ,metode dan aplikasi terkait domain pada big data [9]. Demikian pula ML juga sering terkait dengan data mining dimana data mining lebih mengeksplor analisis data teks[10] .

Pada tahun 1959 disebutkan oleh Arthur Samuel [8] bahwa ML sebagai suatu studi untuk mempelajari sendiri tanpa secara eksplisit

diprogram. Diinspirasi oleh sifat natural tersebut, misalkan kita dapat membangun algoritma untuk mengkoleksi email yang tergolong spam [11]. Mesin akan mengingat bagaimana berdasarkan pengetahuan sebelumnya bahwa email dikatakan spam oleh pengguna maka email yang datang selanjutnya dapat tergolong spam dan bukan spam. Jadi pendekatan cara kerja demikian dikatakan 'pembelajaran karena pengingat'. Hal ini mempunyai kekurangan dalam aspek pembelajaran yaitu kemampuan memberi label pada pesan email yang tidak terlihat. Suatu pembelajaran berhasil jika dapat melakukan kemajuan secara individu dalam melakukan perumuman yang lebih luas. Untuk mencapai perumuman dalam tugas melakukan filterisasi email pada spam, pembelajar dapat menelusur email-email yang sebelumnya dilihat dan melakukan ekstraksi kata-kata dalam pesan yang terindikasi spam. Ketika email baru datang, mesin dapat menguji apakah kata-kata dalam email tersebut sebagai spam dan menduga labelnya [11]. Sistem yang demikian dapat menduga secara benar pelabelan dalam email-email yang tidak terlihat. Dengan dasar pengetahuan ini kemudian dilakukan klasifikasi data google scholar dimana klasifikasi yang dibentuk lebih dari 2 berdasarkan *Word cloud*. Telah diketahui bahwa Word cloud sebagai teknik yang memudahkan pengguna dalam melakukan analisa teks dimana analisa tergantung pada frekuensi teks tersebut muncul [4].

2. Metode

2.1 Tahap pembuatan *Word cloud*

Pada literatur terdapat beberapa *Word cloud generator*. Akan tetapi algoritma yang ada perlu disesuaikan dengan kebutuhan dalam penelitian ini. Pada prinsipnya, ukuran huruf sebuah kata pada Word cloud ditentukan oleh besarnya frekuensi kemunculannya. Untuk frekuensi yang lebih kecil, ukuran huruf dapat langsung digunakan. Sebutlah ukuran mula-mula adalah s_0. Untuk nilai frekuensi yang lebih besar, maka huruf dilakukan penskalaan, dinormalisai linier. Sebutlah nilai t_i adalah hitungan ke-i, t_{max} adalah hitungan maksimum, sedangkan t_{min} adalah hitungan minimum, maka ukuran huruf diskala dalam bentuk formulasi [12] :

$$s_i = \begin{cases} \left[\frac{f_{max}(t_i - t_{\min})}{(t_{max} - t_{min})}\right], t_i > t_{min}. \\ 1 \quad , \quad \text{sebaliknya.} \end{cases}$$

Algoritma yang digunakan adalah algoritma NLP (*Natural Language Process*) yaitu cabang dari kecerdasan buatan yang difokuskan untuk memungkinkan komputer memahami dan menafsirkan bahasa manusia. Proses NLP ditunjukkan pada tahapan berikut ini.

2.1.1 Input data

Pada tahap ini data diinput dengan menyesuaikan sistem yang diacu. Demikian pula terdapat langkah pembangkitan bilangan random agar hasil yang diperoleh konsisten. Pada ML hal ini disediakan menu yang disebut *random seed*.

2.1.2 Preproses data

Pada langkah ini kita mentransformasi data mentah menjadi format yang dikenali untuk model NLP Data biasanya tidak lengkap, tidak konsisten dan/atau juga kekurangan sesuatu atau trend dan juga memuat error.

A. Langkah Tokenisasi

Langkah Tokenisasi (*Tokenization*) yaitu prose memenggal kata/text menjadi kata, frase ataupun elemen yang bermakna yang disebut token. Daftar token kemudian digunakan sebagai proses lebih lanjut. Pada library *nltk* mempunyai *word_tokenize* dan *sent_tokenize* untuk memudahkan daftar kalimat untuk dipenggal menjadi kata atau kalimat-kalimat.

B. Langkah Lemmatisasi

Langkah Lemmatisasi (*Word Stemming/Lemmatization*) bertujuan sama dengan proses diatas yaitu mereduksi bentuk-bentuk infleksi dari tiap kata menjadi bentuk dasar atau akar kata. Misalkan : kata *eating* menjadi *eat*. *Lemmatization* hampir dekat dengan *stemming* : perbedaannya adalah bahwa *stemmer* mengoperasikan suatu kata tanpa ada pengetahuan tentang konteks dan oleh karena itu tidak dapat membedakan kata yang mempunyai perbedaan arti (misal roti dimakan dengan dimakan roti jelas berbeda, tetapi disini tidak dibedakan). *Stemmer* biasanya lebih mudah diimplementasikan dan lebih cepat dan penurunan akurasi dapat menjadi tidak penting untuk beberapa aplikasi . Berikut ini adalah langkah lengkap yang menyusun tahapan *preprocessing* :

a. Membuang baris yang blank dari data jika ada
b. Merubah semua huruf dalam huruf kecil
c. Men-token kata
d. Membuang *stop word*
e. Membuang text *non-alpha*
f. Meng-lemma Kata

2.1.3 Menyiapkan data latih dan data uji

Sebagaimana pada langkah *Machine learning*, maka kita perlu memisahkan data menjadi data latih dan data uji. Pada bagian ini akan ditunjukkan langkah-langkahnya.

2.1.4 Encoding

Dengan *encoding* berarti memberikan pelabelan pada variabel target. Hal ini dilakukan dengan mentransformasikan data kategori bertipe pada tipe string dalam data menjadi data bernilai numerik.

2.1.5 Vektorisasi *Word*

Langkah ini merupakan proses umum yang merubah koleksi dokumen teks menjadi vektor-vektor fitur.` Terdapat banyak metode untuk itu, tetapi yang popular disebut TF-IDF ("*Term Frequency — Inverse Document" Frequency*) yang memberikan skor /nilai pada tiap kata.

- *Term Frequency*: Ini meringkas seberapa sering suatu kata muncul dalam dokumen

- *Inverse Document Frequency:*
- Ini memberikan penyekalaan (secara menurun) terhadap kata yang muncul dalam dokumen.

Tanpa detail matematika, TF-IDF adalah nilai frekuensi kata (*word*) yang diupayakan menonjolkan lebih menarik :misal muncul sangat sering dokumen tapi tidak pada seluruh dokumen. TF-IDF membangun koleksi kata yang telah dipelajari dari data corpus dan yang menunjuk bilangan bulat secara tunggal kepada kata-kata ini. Terdapat 5000 kata/fitur yang tunggal yang ditunjukkan pada parameter *max_features*=5000. Akhirnya kita akan mentransformasi Train_X dan Test_X yang divektorisasi Train_X_Tfidf dan Test_X_Tfidf. Kemudian kita dapat menggunakan algoritma ML untuk memprediksi luaran misal dengan menggunakan Naïve Bayes sebagaimana ditunjukkan pada bagian selanjutnya.

2.2 Klasifikasi dengan *Word cloud*

Pada bagian ini akan dicoba menggambar statistik kata-kata yang sudah disusun. Untuk itu diperlukan *Word cloud* sehingga perlu install *Word cloud*. Jika pada anaconda, kita dapat install pada menu anaconda yaitu : *pip install Word cloud*. Di bawah ini dilakukan sesudah *Word cloud* diinstal. Pada Gambar 2.1 adalah *Word cloud* pada klasifikasi email dalam *spam*. Sedangkan pada Gambar 2.2 merupakan *Word cloud* pada klasifikasi email yang tergolong *nonspam*. Sedangkan pada Gambar 2.3 menggabungkan hasil email *spam* dan *nonspam*.

Gambar 2.1 Hasil *Word cloud* untuk email yang terklasifikasi *spam* .

Gambar 2.2 Hasil *Word cloud* untuk email yang terklasifikasi *nonspam*.

Gambar 2.3 Hasil *Word cloud* untuk email yang terklasifikasi spam dan *nonspam*

2.3 Melakukan Vektorisasi Token

Kita akan membuat 4 *DataFrame*, yaitu:

- DataFrame token_vek, berisi token setiap email yang sudah dibersihkan, dalam sebuah tabel jumlah token tiap email. Ini menjadi fitur data.
- DataFrame klasifikasi_vek, berisi vektor klasifikasi spam atau nonspam.
- DataFrame email_vek, berisi token email yang sudah dirangkai menjadi jadi satu, untuk masing-masing email
- DataFrame vokab_vek, berisi vokabulari yang dipakai untuk menyusun tabel DataFrame token_vek.

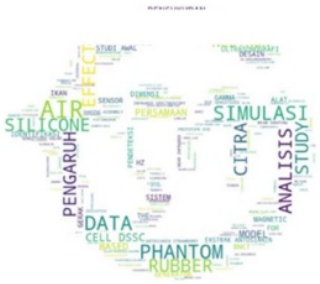
Hasil *Word cloud* memberikan pengaturan tata letak, ukuran dan warna dimana kata dengan frekuensi terbesar akan muncul paling besar dan warna yang lebih dominan sehingga mudah dikenali.

3. Hasil dan Diskusi

3.1 Data kasus *Word cloud* dosen Magister Sains Data

Sebagai kasus awal yang dipelajari adalah data penelitian dosen Magister Sains Data di Fakultas Sains dan Matematika UKSW. Dengan menggunakan machine learning maka diperoleh hasil *Word cloud* menurut Gambar 3.1-3.3.

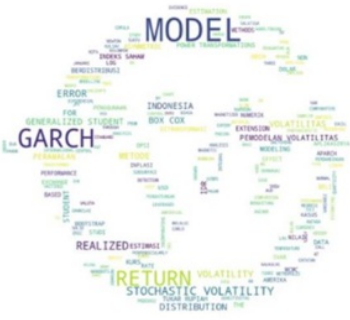
Gambar 3.1 *Word cloud* untuk data penelitian Dr. Adi Setiawan, MS berdasarkan data hingga Juli 2020.



Gambar 3.2 *Word cloud* untuk data penelitian Dr.Suryasatriya Trihandaru,MSc berdasarkan data hingga Juli 2020.

Gambar 3.3 *Word cloud* untuk data penelitian Didit Budi Nugroho, Msi,DSc berdasarkan data hingga Juli 2020.

Gambar 3.4 Word cloud untuk data penelitian Dr. Hanna Arini Parhusip berdasarkan data hingga Juli 2020.



Gambar 3.5 *Word cloud* untuk data penelitian Dr. Bambang Susanto,MS berdasarkan data hingga Juli 2020.

3.2 Bagaimana melakukan Analisa?

Sejauh ini *Word cloud* dibuat belum diteliti akurasi model yang diperoleh. Akan tetapi secara sekilas kata yang menonjol dari 5 orang sampel adalah kata 'Model' dan kata 'Data'. Kata dominan yang lain adalah 'Analisis'. Secara garis besar maka dapat disimpulkan bahwa riset yang ditunjukkan oleh 5sampel pada Gambar 3.1-3.5 di atas adalah tentang modelling data serta analisanya. Salah satu kata yang menonjol adalah Garch. Dari status jurnal yang dimunculkan ini telah diketahui termasuk dalam Scopus sehingga mendapatkan dominasi pada *Word cloud*. Studi kasus selanjutnya akan dilakukan penelusuran riset unggulan UKSW dengan *Word cloud* bagi seluruh dosen UKSW melalui penelitian ini.

3.3 Hasil Penelitian data kasus *Word cloud* UKSW

Penelusuran riset unggulan UKSW telah dilakukan menggunakan *Word cloud* data seluruh dosen UKSW pada masing-masing Fakultas yang ada di UKSW. Dengan menggunakan *machine learning*, maka diperoleh hasil *Word cloud* pada beberapa Fakultas yang ditunjukkan pada Gambar 3.6-3.12.

Gambar 3.6 *Word cloud* untuk data penelitian Fakultas Bahasa dan Sastra (kiri) dan Fakultas Biologi (kanan)

Gambar 3.7 Word cloud untuk data penelitian Fakultas Ekonomi dan Bisnis (kiri) dan Fakultas Hukum (kanan)

Gambar 3.8 *Word cloud* untuk data penelitian Fakultas Ilmu Sosisal dan Komunikasi (kiri) dan Fakultas Interdisiplin (kanan)

Gambar 3.9 *Word cloud* untuk data penelitian Fakultas Pertanian dan Bisnis (kiri) dan Fakultas Psikologi (kanan)

Gambar 3.10 *Word cloud* untuk data penelitian Fakultas Sains dan Matematika (kiri) dan Fakultas Teknik Elektro dan Komputer (kanan)

Gambar 3.11 *Word cloud* untuk data penelitian Fakultas Teknologi dan Informatika dan Fakultas Teologi

Gambar 3.12 *Word cloud* untuk data penelitian UKSW .

Pada penelitian *Word cloud* yang ada pada riset unggulan UKSW, di tiap fakultasnya memiliki kata yang menonjol berbeda-beda. Hal ini berarti di tiap fakultas memiliki penelitian dan hasil yang berbeda dari fakultas satu ke fakultas yang lain. pada Fakultas Bahasa dan Sastra misalnya, kata yang paling menonjol adalah "*English*","*Student*","*Teacher*","*Teaching*", dan "Musik" ini berarti *Word doud* dapat menunjukkan bahwa ada penekanan kata yang terlihat dari jumlah frekuensi kata yang berulang pada penelitian yang ada di Fakultas Bahasa dan Sastra UKSW. Ditunjukkan pula semakin besar hurufnya berarti semakin populer topik / laporan yang mengandung kalimat kalimat tersebut yang diangkat pada setiap riset yang ada di fakultas tersebut.

Kemudian di Fakultas Biologi kata yang paling sering muncul adalah "SMA","Daun","SMP","Tempe", dan "Bakteri". Sehingga dapat dikatakan bahwa di fakultas biologi lebih sering mengangkat penelitian/riset yang menggunakan kata-kata tersebut atau bahkan

penelitian dengan tema yang mengangkat kata tersebut seperti Tempe atau Bakteri. Berikutnya di Fakultas Ekonomi dan Bisnis kata yang sering muncul adalah "Akuntansi", "Perusahaan", "*Corporate Social*","Bank", dan "*Social Responsibilty*". Ini menunjukkan bahwa di Fakultas Ekonomi dan Bisnis kata pada riset atau penelitian yang muncul adalah seputar kata-kata tersebut. Dan yang paling sering digunakan adalah kata Perusahaan dan Akuntansi karena kata yang paling menonjol dan paling besar adalah pada kedua kata tersebut. Di Fakultas Hukum kata-kata yang paling sering muncul adalah "Hukum","Internasional',dan "undang-undang". Hal ini menunjukkan bahwa pada riset penelitian unggulan yang sering dilakukan mengandung kata-kata tersebut, tidak terlepas dari materi yang ada di Fakultas Hukum. Di Fakultas Ilmu Sosial dan Komunikasi frekuensi kata yang sering muncul adalah "Masyarakat","Sosial","Negara","Komunikasi" serta "peran" ini berarti kata frekuensi yang sering digunakan dalam riset atau penelitian di fakultas ini adalah seputar kata-kata tersebut. Di fakultas Interdisiplin kata yang sering muncul adalah "Batik" karena kata yang menonjol ada pada kata Batik ini berarti di fakultas ini sering menggunakan kata / riset yang mengandung kata Batik. Kemudiaan di Fakultas Pertanian dan Bisnis frekuiensi kata yang sering muncul adalah pada kata "Desa","Petani","Pertanian", dan "Triticum Aestivum". Hal ini menunjukkan bahwa riset atau penelitian yang sering dilakukan oleh Fakultas pertanian dan bisnis adalah menangkat tema pada kata yang menonjol pada *Word cloud* tersebut. Kemudian di Fakultas Psikologi kata yang sering muncul adalah "Remaja","Kerja","Gambaran","Guru", dan "perilaku" hal ini dibuktikan dengan adanya kata yang paling menonjol yang memiliki ukuran yang paling besar, berarti di fakultas psikologi pada riset dan penelitiannya sering mengangkat tema atau kata yang menggunakan kalimat tersebut. Kemudian di Fakultas Sains dan Matematika frekuensi kata yang sering muncul adalah pada kata "Materi","Fisika","Stevia rebaudiana", dan "Identifikasi" hal ini menunjukkan bahwa di Fakultas Sains dan matematika sering menggunakan kata kata tersebut sebagai riset atau penelitian yang dilakukan. Kemudian di Fakultas Teknik Elektro dan Komputer kata yang paling menonjol adalah pada kata "Robot","Jaringan","Digital","*Support Vector*", dan "*Vector Machine*" ini menunjukkan bahwa Fakultas teknik Elektro dan Komputer sering menggunakan kata atau tema untuk riset penelitiannya. Kemudian di Fakultas Teknologi dan Informatika frekuensi kata yang sering muncul adalah pada kata "Informasi","Framework cobit","WEB","Komunikasi" dan "Android" ini mwnunjukkan bahwa dalam penelitian dan risetnya fakultas teknologi dan informatika sering menggunakan kata-kata tersebut. Kemudian di Fakultas Teologi pada risetnya sering menggunakan kata "Sosial","Masyarakat","Ritual","Kesehatan Mental", dan "Perempuan". Berikutnya untuk UKSW sendiri kata yang menonjol adalah pada kata "Student","Informasi","Pendidikan","Development","WEB" dan "Evaluasi". Ini menunjukkan bahwa di UKSW sering dilakukan penelitian dan riset dengan mengangkat tema atau dari kata tersebut.

4. Kesimpulan

Pada penelitian ini ditunjukkan tentang pembuatan *Word cloud data scince* Universitas Kristen Satya Wacana (UKSW) yaitu pembacaan data riset pada tiap dosen di UKSW. Hal ini dilakukan karena untuk upaya pendataan terhadap hasil penelitian UKSW seringkali mengalami kesulitan karena data tidak terdokumentasi secara terpadu. Selain itu, untuk menelusuri kompetensi penelitian dosen tidak dapat dilakukan dengan menelusuri dokumen secara manual untuk dapat membuat kesimpulan global. Penggunaan *Word cloud* program dengan *machine learning* merupakan salah satu cara cepat yang dapat dilakukan dalam menelusuri dokumen yang terdigitalisasi pada google scholar. Hasil yang diperoleh dari penelitian ini adalah hasil *Word cloud* yang dapat memudahkan untuk mengetahui riset dan penelitian apa saja yang telah dilakukan para dosen di fakultas, maupun di universitas. *Word cloud* juga memudahkan dalam pendataan terhadap hasil -hasil penelitian UKSW dan riset apa saja yang telah dilakukan masing-masing fakultas. Dari hasil yang diperoleh, tiap fakultas menunjukkan sesuai dengan karakteristik dari fakultas tersebut. Sedangkan hasil yang menonjol dari tiap dosen ditunjukkan oleh jurnal yang muncul dengan reputasi internasional terindeks Scopus akan mendapatkan tampilan pada *Word cloud* lebih utama. Jika memperhatikan kata dengan frekuensi tertinggi dalam hasil data riset dosen di UKSW, maka kata yang dominan muncul adalah "Student","Informasi","Pendidikan","Development","WEB" dan "Evaluasi".

Persantunan

Penelitian ini didanai oleh pusat penelitian UKSW dengan penelitian internal tahun anggaran 2020/2021 dengan penelitian yang berjudul 'Analisa Riset Unggulan UKSW menggunakan Machine Learning dan Data Google Schoolar'.

Daftar Pustaka

[1] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, "Text Classification for Organizational Researchers: A Tutorial," *Organ. Res. Methods*, vol. 21, no. 3, pp. 766–799, 2018, doi: 10.1177/1094428117719322.

[2] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, "Short Text Classification: A Survey," *J. Multimed.*, vol. 9, no. 5, pp. 635–643, 2014, doi: 10.4304/jmm.9.5.635-643.

[3] B. Tessem, S. Bjørnestad, W. Chen, and L. Nyre, "Word cloud visualisation of locative information," *J. Locat. Based Serv.*, vol. 9, no. 4, pp. 254–272, 2015, doi: 10.1080/17489725.2015.1118566.

[4] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on Word cloud," in

Proceedings of the Annual Hawaii International Conference on System Sciences, 2014, pp. 1833–1842, doi: 10.1109/HICSS.2014.231.

[5] C. An, H. Lim, D. W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-75767-2.

[6] J. Beschi Raja, R. Anitha, R. Sujatha, V. Roopa, and S. Sam Peter, "Diabetics prediction using gradient boosted classifier," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 3181–3183, 2019, doi: 10.35940/ijeat.A9898.109119.

[7] H. Qian, "Big data Bayesian linear regression and variable selection by normal-inverse-gamma summation," *Bayesian Anal.*, vol. 13, no. 4, pp. 1007–1031, 2018, doi: 10.1214/17-BA1083.

[8] P. Dönmez, "Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95 + 584 pages.," *Nat. Lang. Eng.*, vol. 19, no. 2, pp. 285–288, 2013, doi: 10.1017/s1351324912000290.

[9] L. Demidova, E. Nikulchev, and Y. Sokolova, "Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 294–312, 2016, doi: 10.14569/ijacsa.2016.070541.

[10] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, 2012, pp. 415–463.

[11] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[12] Y. Jin, "Development of Word Cloud Generator Software Based on Python," in *Procedia Engineering*, 2017, vol. 174, pp. 788–792, doi: 10.1016/j.proeng.2017.01.223.

Wordcloud

ORIGINALITY REPORT

10%	9%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

3%

★ journals.ums.ac.id

Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		