e-Bangi Faculty of Social Sciences and Humanities Universiti Kebangsaan Malaysia
Journal of Social Sciences and Humanities

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

PERBANDINGAN ALAT PENGEKSTRAKAN DATA TEKS JANAAN PENGGUNA

(A Comparison of Text Data Extraction Tool Using User Generated Content)

Fatimah Rahmat, Zuraihah Ngadengon & Nurul Shakirah Mohd Zawawi

ABSTRAK

Era Big Data dan perlombongan data telah mewarnai dunia penyelidikan data teks yang dijana pengguna. Peningkatan pengguna media sosial setiap tahun bermaksud pertambahan data dan maklumat yang dijana pengguna memenuhi ruang pelayan di kerangka utama laman sesawang terlibat. Data dan maklumat ini amat bernilai sekiranya digunakan untuk tujuan penyelidikan. Namun begitu, bagaimanakah untuk mengekstrak bilangan data teks dalam jumlah yang banyak dengan mudah? Terdapat alat-alat pengesktrakan data teks yang telah dicipta untuk menyelesaikan masalah ini. Malah, banyak kajian terdahulu yang menggunakan data teks sebagai data utama dalam kajiannya tetapi tiada penerangan jelas tentang cara menggunakan alat pengekstrakan data teks tersebut. Oleh itu, kajian ini membincangkan berkenaan lima alat pengesktrakan data teks, ciri-ciri alat pengekstrakan data teks dan perbandingan terhadap 5 alat tersebut. Kajian ini telah melalui fasa penerokaan penting iaitu pemasangan perisian, pengujianan dan hasil output bagi setiap alat tersebut. Hasil kajian ini mendapati bahawa, setiap penggunaan alat pengekstrakan data teks mempunyai ciri istimewa yang tersendiri iaitu jenis perisian, tahap penggunaan, asas pengetahuan pengguna dan jenis laman sesawang yang boleh diekstrak. Selepas melalui proses pengujian bagi setiap alat, kajian ini mendapati bahawa alat pengekstrakan data teks ini sangat memudahkan pengguna untuk mendapatkan data teks dalam kuantiti yang banyak secara sistematik. Oleh itu, semoga kajian ini dimanfaatkan sepenuhnya bagi membolehkan para penyelidik berinteraksi dan berkongsi idea dengan ramai orang serta menjadi rujukan untuk penyelidikan akan datang.

Keywords: data teks, data yang dijana pengguna, media sosial, big data, perlombongan data

ABSTRACT

The era of Big Data and Data Mining has colored the world of user-generated content research in text data. The rising number of social media users each year results in the expansion of data and information generated by users whom filling-up the server space on the main homepage. This data and information is very valuable when it is being used for research purposes. However, how to easily extract large amounts of text data? A data text extraction tool has been created to solve this problem. In fact, many previous studies used text data as the primary data in their study but it was unclear how to use the data extraction tool. Therefore, this study discusses the 5 text data extraction tools, their features and comparisons across the tools. This study went through an important exploration phase of software installation, testing and output of each of these tools.

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

The results show that every use of data text extraction tool has its own unique features namely software type, usage level, user basic knowledge and extracted web site type. After going through the testing process for each tool, this study found that this text data extraction tool made it very easy for users to systematically obtain large amounts of text data. Therefore, it is hoped that this study will be fully utilized to enable researchers to interact and share ideas with many people and served as a reference for future research.

Keywords: Text data, user generated content, social media, big data, data mining

PENGENALAN

Setiap manusia dikategorikan sebagai pengguna apabila menggunakan sesuatu barang atau perkhidmatan. Pengguna ini pula terbahagi kepada pengguna dunia relaiti dan pengguna alam maya. Pengguna realiti adalah pengguna yang menggunakan barang atau perkhidmatan secara fizikal dan dapat dilihat orang lain. Dengan perkembangan pesat teknologi maklumat pada masa kini, wujud pengguna maya atas talian di mana pengguna ini melawat dari satu laman web ke satu laman web untuk mengikut tujuan masing-masing seperti laman sesawang www.tonton.com.my untuk menonton drama dan berita, laman sesawang www.shopee.com untuk membeli barang keperluan harian, laman sesawang www.youtube.com untuk belajar memasak dan laman sesawang https://web.whatsapp.com untuk berbual secara maya dengan rakan-rakan. Laman sesawang ini membenarkan pengguna berinteraksi dengan pengguna lain secara maya dengan menaip teks dalam ruangan yang disediakan dan membolehkan pengguna maya lain turut serta berbincang terhadap topik tertentu. Pengguna boleh berkongsi kandungan maklumat dalam bentuk teks, imej, video atau animasi ringkas dengan pengguna yang lain semasa berinteraksi. Situasi perkongsian data dan maklumat ini setiap masa berlaku di kalangan pengguna alam maya. Data dan maklumat ini merupakan kandungan yang dijana oleh pengguna atau dikenali sebagai "User Generated Content (UGC)".

Menurut Krumm, Davies, & Narayanaswami (2008), UGC adalah kandungan data dan maklumat yang datang daripada pengguna secara sukarela berkongsi data berbentuk teks dan media dalam bentuk yang berguna atau yang menghiburkan dalam laman sesawang tertentu. Contoh teks dan media dalam bentuk yang berguna adalah seperti blog dan wiki. Manakala contoh teks dan media yang menghiburkan adalah seperti media sosial Facebook, Youtube, Whatsapp, Twitter dan Instagram. UGC telah berkembang dengan pesat pada setiap masa kerana kemudahan Internet Service Provider (ISP) di Malaysia telah memberikan kemudahan capaian rangkaian Internet yang mampu milik dengan kapasiti jalur lebar yang cukup untuk menampung capaian Internet bagi setiap penggunanya.

Penggunaan aplikasi media sosial seperti Twitter, Instagram dan Facebook sebagai medium interaksi dan perkongsian kandungan juga telah semakin meluas sehingga mencecah 3.8 bilion pengguna media sosial di seluruh dunia (Dubras & Tono 2020). Perubahan dalam dunia aplikasi media sosial telah membolehkan setiap individu menulis status hantaran mereka dalam bentuk perbincangan, mengkritik, membuat ulasan dan juga memberi cadangan atau pendapat dalam pelbagai bidang pengetahuan. Bilangan pengguna internet diseluruh dunia telah meningkat sehingga 4.54 bilion (Dubras & Tono 2020). Dengan peningkatan jumlah bilangan pengguna dan gajet telefon pintar mampu milik yang terdapat di pasaran membolehkan

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

pemegang akaun media sosial memuatnaik hantaran dengan lebih mudah. Hantaran ini mudah dicapai dan sentiasa mengikuti perkembangan berita semasa sama ada dari segi politik, sukan mahu pun hiburan. Justeru, kajian ini bertujuan untuk menerokai apakah alat yang boleh digunakan untuk mengekstrak kandungan yang dijana pengguna (UGC) berbentuk data teks.

Teks merupakan medium utama yang digunakan dalam perkongsian data dan maklumat dalam era Big Data (Kim & Johnson 2016). Teks ini terdiri daripada dua jenis teks iaitu teks berstruktur dan teks tidak berstruktur. Menurut Sharma & Srivastava (2016), teks berstruktur ialah teks yang dipaparkan dalam bentuk baris dan lajur yang memudahkan proses capaian menggunakan alat perlombongan data. Contohnya seperti katalog perpustakaan (mengandungi baris seperti tarikh, nama penulis, tajuk dan penerbit) dan rekod banci (mengandungi baris seperti tarikh lahir, jantina, alamat dan pekerjaan). Manakala teks tidak berstruktur pula ialah teks yang tergolong dalam kumpulan UGC seperti emel, khidmat pesanan ringkas dan status individu yang dimuatnaik dalam media sosial (Sharma & Srivastava 2016). Teks tidak berstruktur ini merupakan sumber maklumat terbesar yang mudah dicapai melalui media sosial.

Lazimnya, individu yang mempunyai akaun media sosial akan memuatnaik hantaran atau komen berbentuk teks setiap hari. Terdapat banyak kajian yang diterbitkan telah menjadikan data teks ini sebagai data utama dan menghasilkan hasil dapatan kajian yang baru dan menarik seperti analisis sentimen terhadap produk dan perkhidmatan seperti mana yang telah diterangkan dalam bab Kajian Literatur. Namun demikian, kajian-kajian ini tidak menerangkan kaedah atau alat untuk mendapatkan data teks tersebut secara terperinci. Jadi, bagi pengkaji yang memerlukan data teks ini dalam kajiannya, memerlukan masa yang lebih untuk melakukan eksplorasi dan eksperimen berkenaan alat yang sesuai untuk digunakan untuk mengekstrak data teks.

Mengapa perlu mengekstrak data teks yang dijana pengguna? Hal ini demikian kerana tidak semua data teks ini digunakan untuk kegunaan kajian ilimiah sahaja. Corak perniagaan atas talian zaman moden pada masa kini telah berlumba-lumba membuat pertandingan yang membolehkan pengguna berpeluang menerima hadiah berbentuk wang mahupun produk perniagaan itu sendiri. Sebagai contoh, pertandingan siapa yang paling kreatif memberi komen atau keterangan berdasarkan imej yang disediakan. Lazimnya, peniaga akan membaca satu persatu komen tersebut dengan skrol ke atas dan ke bawah laman sesawang media sosial tersebut dan mula membuat penilaian secara manual. Proses ini mengambil masa yang lama untuk dianalisa kerana perlu dibuat dengan satu persatu. Jadi, dengan mengetahui alat pengekstrakkan data teks ini, pengguna boleh menganalisa data teks tersebut dalam bentuk paparan format fail yang mudah seperti fail hamparan Microsoft Excel. Justeru, ia memudahkan proses pencarian pemenang dengan cara yang sistematik.

Oleh itu, kajian ini akan menyenaraikan lima alat pengekstrakan data teks yang ada dan boleh digunakan, mengkaji ciri-ciri alat pengekstrakan dan membuat perbandingan terhadap alat pengekstrakan data teks tersebut.

KAJIAN LITERATUR

Kini, kita telah memasuki era Big Data (Amado et al. 2017). Big Data bukan hanya mengenai jumlah data (volume), tetapi juga mengenai pelbagai (variety) dan halaju (velocity) (Amado et al. 2017). Data teks adalah sesuatu yang bernilai dan mempunyai makumat yang boleh digunakan dalam pelbagai aspek seperti pengkelasan data teks (Alsudais, Leroy, & Corso 2014),

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

analisis sentimen (Thakkar & Patel 2013) (Handayani, Bakar, & Abuzuraida 2018) membuat ramalan keputusan (Ahmad, Ismail & Aziz 2015) dan pengoptimuman enjin carian (SEO) (Singh & Maini 2013). Kajian Sharma & Srivastava (2016) menyatakan bahawa perlombongan data ialah satu proses analisis data untuk mendapatkan maklumat tersembunyi yang berguna daripada koleksi set data yang besar. Kajian Alsudais, Leroy & Corso (2014) pula menyatakan, data-data berbentuk teks yang tersembunyi dalam persekitaran ini boleh diekstrak menggunakan alat pengekstrakkan data teks dan pengkaji boleh menggunakan data tersebut untuk proses kajian selanjutnya. Analisis sentimen menentukan pendapat mengenai objek atau subjek dalam sesuatu topik perbincangan yang dihasilkan sendiri oleh individu (Thakkar & Patel 2013). Kajian Ahmad, Ismail & Aziz (2015) menggunakan data teks untuk mengetahui ramalan prestasi pelajarnya untuk membaiki gred matapelajaran mereka dalam kajiannya. Daripada kajian-kajian tersebut, era Big Data ini membolehkan pengguna mencapai data teks sedia ada yang terdapat dalam laman sesawang media sosial tanpa perlu meminta keizinan daripada pemilik data tersebut. Tiada sekatan dalam mendapatkan data teks ini.

Perubahan dalam dunia aplikasi media sosial telah membolehkan setiap individu menulis status mereka dalam bentuk perbincangan, mengkritik, membuat ulasan dan juga memberi cadangan atau pendapat dalam pelbagai bidang pengetahuan. Contohnya, koleksi data set yang besar berkenaan ulasan sesuatu produk boleh dikelaskan pada ulasan positif atau ulasan negatif. Produk yang mendapat banyak reaksi negatif memerlukan penambahbaikan bagi meningkatkan prestasi produk tersebut. Manakala ulasan positif pula membolehkan pengeluar produk meningkatkan proses pengeluaran untuk memenuhi permintaan pengguna. Maklumat ulasan positif dan negatif bagi produk ini penting untuk pengeluar produk mengetahui maklumbalas pengguna secara telus dan terus daripada pengguna. Situasi ini yang katakan analisis sentimen.

Media sosial diguna sebagai medium komunikasi yang membolehkan keluarga dan rakan-rakan terdekat untuk sentiasa berhubung antara satu sama lain dengan berkongsi aktiviti yang dilakukan sepanjang hari (Hiltz & Plotnick 2013). Sebagai contoh, sebuah keluarga yang menetap jauh di luar negara memuatnaik gambar perkembangan anak menggunakan media sosial untuk tatapan keluarga di Malaysia. Tetapi kini, perkongsian itu boleh dikongsi oleh sesiapa sahaja yang berdaftar dengan media sosial. Tambahan pula dengan kemudahan telefon pintar yang menyediakan aplikasi media sosial yang membolehkan capaian Internet dicapai pada bila-bila masa dan di mana sahaja.

Media sosial telah mengubah tingkah laku seseorang individu mahupun organisasi dalam konteks perkongsian maklumat. Berdasarkan kajian Jalonen (2014), media sosial dilihat sebagai perubahan komunikasi secara global yang mana pengguna dengan secara sukarela berkongsi pelbagai pengalaman yang penuh dengan pelbagai emosi dan situasi sama ada dalam bentuk teks, gambar mahupun video. Dengan bilangan pengguna media sosial yang sentiasa bertambah seiring dengan perkongsian kandungan yang dikongsi menyebabkan maklumat yang berlebihan telah berlaku dalam dunia digital media sosial. Dengan perkembangan UGC ini, maklumat berlebihan ini mempunyai potensi yang besar untuk digunakan oleh para penyelidik untuk mendalami tentang bagaimana untuk mengekstrak maklumat tersebut dan digunakan dalam pelbagai aplikasi kajian yang bersesuaian (Huang et al. 2013). Walau bagaimanapun, bentuk maklumat yang berlebihan ini adalah dalam bentuk tidak berstruktur kerana ia dihasilkan untuk kegunaan manusia dan bukan untuk diproses oleh komputer.

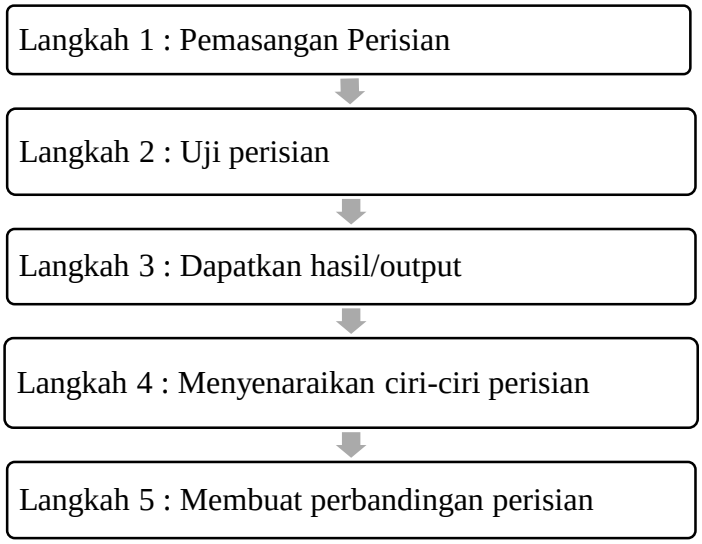
Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

Data-data sedia ada yang ada dimuat naik dalam media sosial terdiri daripada bentuk teks seperti hantaran dan komen, multimedia seperti video, imej dan audio dan jumlah bilangan butang suka dan butang kongsi (Kente 2017). Maklumat yang berlebihan ditakrifkan sebagai maklumat yang diberikan pada kadar yang terlalu cepat bagi seseorang untuk memproses (Hiltz & Plotnick 2013). Kadar maklumat yang dimuatnaik tidak mengikut topik atau mempunyai isi kandungan yang tidak penting dalam keadaan semasa (Hiltz & Plotnick 2013). Bukan itu sahaja, kajian Hiltz & Plotnick (2013). juga mendapati maklumat yang berlebihan ini juga kerap berlaku dan sentiasa memenuhi *"news feed"* pengguna apabila terdapat situasi berisiko tinggi seperti peristiwa kecemasan kebakaran, gempa bumi dan banjir. Dengan kebanjiran maklumat yang berlebihan, membolehkan kajian analisis teks diketengahkan memandangkan maklumat ini boleh dicapai dengan mudah tanpa perlu menggunakan kaedah penyelidikan secara tradisional seperti borang soal selidik. Oleh itu, kajian ini hanya tertumpu kepada alat pengekstrakkan data berbentuk teks sahaja.

KAEDAH KAJIAN

Kaedah kajian dalam kajian ini adalah berdasarkan langkah-langkah proses kajian yang dilakukan sepanjang kajian dijalankan yang telah diringkaskan. Terdapat lima langkah proses utama yang telah dikenalpasti untuk kajian ini (rujuk Rajah 1). Setiap langkah proses ini perlu dilakukan secara satu persatu mengikut aturan yang disediakan.

Rajah 1 : Proses eksplorasi alat pengekstrakan data teks janaan pengguna



Langkah utama dalam kajian ini ialah dengan memasang perisian yang dikehendaki mengikut keperluan spesikasi setiap perisian. Seterusnya, setiap perisian diuji dengan mengesktrak data sama ada daripada FB atau Twitter. Semasa proses ini dijalankan, ujian telah dilakukan sehingga output yang sebenar dalam bentuk fail Microsoft Excel diperoleh. Jika tiada output yang diperolehi, proses ujian ini akan diulang semula sehingga mencapai output yang dikehendaki. Pada langkah ketiga, output yang diperoleh pula akan dianalisa dengan mengetahui

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

jenis fail, nama bagi setiap kolum yang diberi dan menyemak tarikh data teks tersebut dijana. Setelah tiga langkah utama ini tamat, setiap perisian ini akan di rujuk semula untuk menyenaraikan ciri-ciri utama perisian yang perlu diketahui bagi menyempurnakan langkah seterusnya. Langkah terakhir dilengkapkan dengan membuat perbandingan bagi setiap perisian berdasarkan penemuan yang diperolehi sepanjang proses kajian.

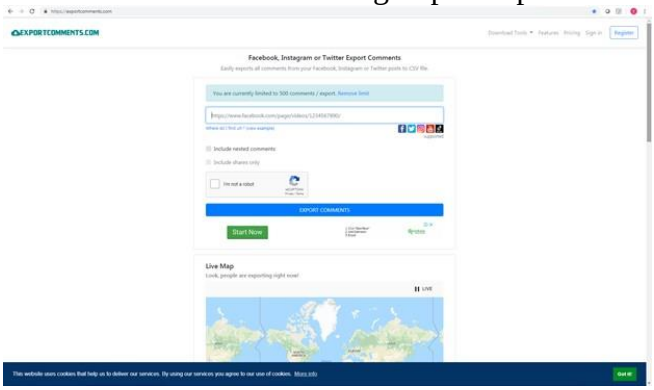
HASIL KAJIAN

Hasil daripada proses penerokaan bagi setiap alat pengekstrakan data teks, kajian ini menemui lima alat pengekstrakan data teks janaan pengguna yang boleh digunakan mengikut tahap iaitu tahap permulaan, tahap pertengahan dan tahap lanjutan. Bab ini membincangkan berkenaan kelebihan dan kelemahan setiap alat berdasarkan pengalaman yang diperolehi oleh pengkaji. Kajian ini mendapati bahawa, setiap alat mempunyai ciri istimewa yang tersendiri dan terpulang kepada individu untuk memilih alat yang bersesuaian dan boleh digunapakai.

(i) Aplikasi Atas Talian https://exportcomments.com

Laman sesawang https://exportcomments.com menyediakan aplikasi pengekstrakan data teks atas talian secara terus kepada pengguna (Rujuk Rajah 2). Menurut "Web Analysis for Exportcomments - exportcomments.com," (n.d.), aplikasi ini baru diperkenalkan dan ianya selamat untuk diguna. Aplikasi ini boleh mengeksport komen primari umum daripada media sosial Facebook, Instagram, Twitter, TikTok atau YouTube ke dalam bentuk fail hamparan format Microsoft Excel csv (Comma Delimited). Aplikasi ini amat sesuai untuk pengguna bagi tahap permulaan. Pengguna hanya perlu memasukkan URL yang ingin diekstrak dan laman sesawang tersebut akan menjalankan proses seterusnya.

Rajah 2 : Antaramuka laman sesawang https://exportcomments.com



Kelebihan

Aplikasi ini mesra pengguna. Menjimatkan ruang simpanan komputer kerana tiada pemasangan perisian perlu dibuat. Pengguna tidak semestinya mempunyai akaun berdaftar dengan media sosial. Fail hamparan format Microsoft Excel csv (Comma Delimited) yang dimuat turun daripada aplikasi ini membolehkan pengguna melihat profil pengguna media sosial secara terus.

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

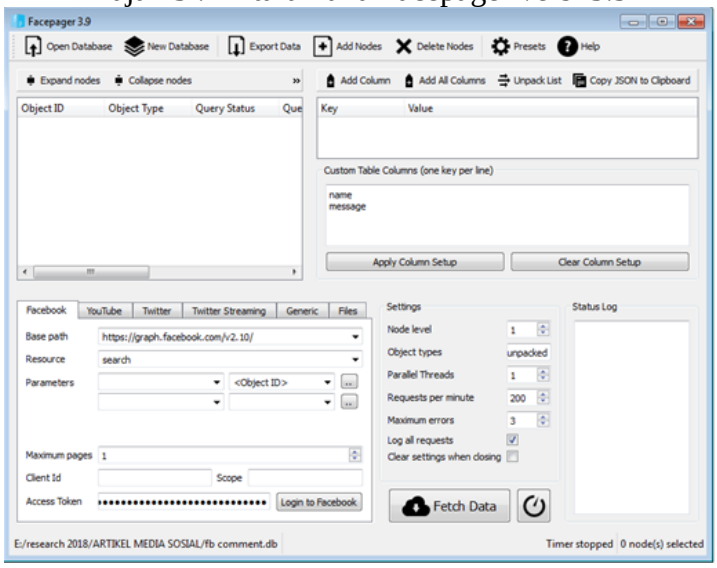
Kelemahan

Aplikasi ini memberi kemudahan percuma dan terhad untuk 500 komen sahaja yang akan dimuat turun. Bayaran akan dikenakan mengikut pakej sedia ada seperti pakej personal, pakej premium dan pakej perniagaan jika mahu mendapatkan data dalam kuantiti yang lebih banyak.

(ii) Perisian Facepager

Facepager adalah aplikasi berbentuk perisian terbuka (Keyling & Junger 2019). Facepager membolehkan pengumpulan data dapat didokumenkan dan amat relevan untuk kajian saintifik. Facepager tidak memerlukan kemahiran pengaturcaraan. Facepager boleh mengesktrak data berbentuk teks daripada FB, YouTube dan Twitter (Rujuk Rajah 3). Keperluan utama perisian ini ialah pengguna perlu mempunyai akaun media sosial yang berkenaan untuk mencapai data teks tersebut. Teks disimpan adalah dalam bentuk fail hamparan format Microsoft Excel csv (Comma Delimited).

Rajah 3 : Antaramuka Facepager versi 3.9



Kelebihan

Facepager ini amat sesuai untuk mengkstrak data teks daripada *fb page*. Facepager boleh mengekstrak data teks lebih daripada 100 hantaran dalam satu permintaan (Kente 2017) Penggunaan perisian ini mudah dengan enam langkah berikut:

1. Masukkan nama pangkalan data untuk meyimpan himpunan data teks dalam simpanan cakera keras
2. Masukkan nama nod iaitu nama FB page yang dikehendaki
3. Pilih jenis data teks yang diigini
4. Login ke akaun FB pengguna
5. Tekan *"Fetch Data"* untuk mengekstrak data teks tersebut
6. Ekspot data teks tersebut ke dalam format fail .csv
7. Selesai

e-Bangi Faculty of Social Sciences and Humanities Universiti Kebangsaan Malaysia
Journal of Social Sciences and Humanities

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

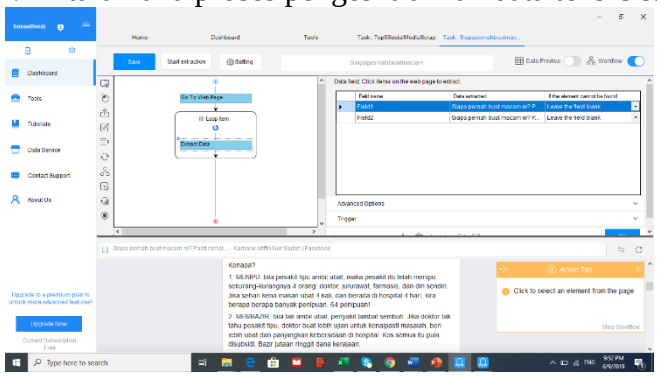
Kelemahan

Facepager memberikan hasil output dalam bentuk yang sukar dibaca iaitu segala data teks diletakkan dalam satu kolum. Oleh itu, pengguna perlu menukar format output tersebut secara manual dalam Microsoft Excel.

(iii) Perisian Octoparse

Octoparse adalah perisian pengekstrakan data teks untuk semua jenis laman sesawang yang ada seperti media sosial, blog, wiki dan juga laman sesawang berita (Rujuk Rajah 4). Pengguna hanya perlu mendaftar dan memasang perisian Octoparse untuk menggunakannya. Kemudian, pengguna hanya perlu memasukkan URL laman sesawang yang ingin diekstrak. Octoparse membolehkan pengguna memilih sendiri bahagian teks yang diperlukan dan mengabaikan pada bahagian yang tidak diperlukan. Tutorial disediakan di dalam perisian ini.

Rajah 4: Antaramuka proses pengesktrakkan data teks Octoparse



Kelebihan

Octoparse sesuai kepada pengguna yang mempunyai pengalaman atau tidak dalam pengalaman mengekstrak data teks. Octoparse menyediakan kemudahan panel operasi berbentuk visual yang sangat mesra pengguna dan mudah ("About Octoparse," n.d.). Tiga langkah mudah bagi menggunakan perisian ini:

1. Masukkan URL yang dikehendaki
2. Klik pada targert data yang hendak diekstrak
3. Jalankan proses pengekstrakan dan dapatkan data. Output boleh diperolehi dalam bentuk format fail Excel 2007 (xlsx), Excel 2003 (xlsx), JSON, Csv, HTML atau ekspot ke bentuk pangkalan data SQL.

Kelemahan

Octoparse tidak menggunakan pengaturcaraan tetapi pengalaman dan pengetahuan pengaturcaraan amat diperlukan untuk menggunakan Octoparse kerana pengguna sendiri yang perlu memilih data teks yang mana yang hendak diambil. Setiap pengguna Octoparse terhad kepada 10,000 URL sahaja bagi keseluruhan carian data teks. Jika bilangan URL telah dipenuhi, pengguna perlu membayar untuk meneruskan perkhidmatan perisian ini.

e-Bangi Faculty of Social Sciences and Humanities Universiti Kebangsaan Malaysia
Journal of Social Sciences and Humanities

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

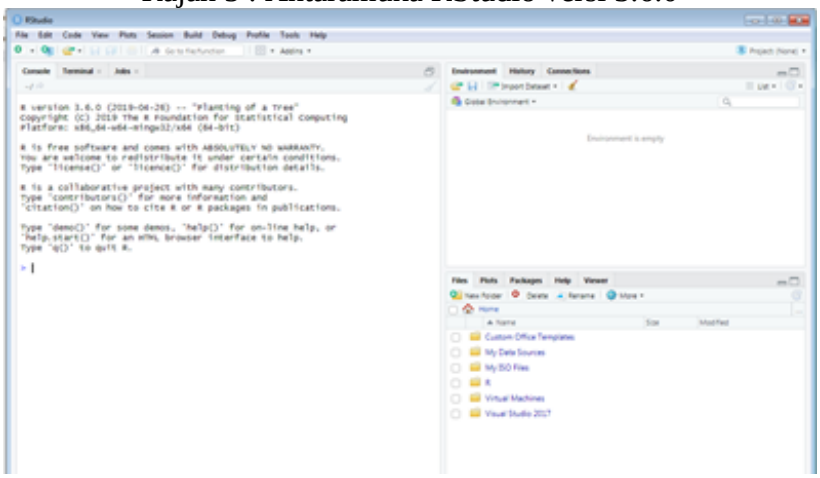
(iv) Perisian RStudio

RStudio merupakan perisian sumber terbuka yang membolehkan analisis data dilaksanakan secara analisis visual dan sokongan grafik serta analisis ramalan yang boleh digunakan untuk membuat keputusan yang lebih baik dalam mencari penyelesaian (Nasridinov & Park 2013). RStudio ialah sebuah perisian terbuka yang menggunakan persekitaran pembangunan bersepadu (IDE) (Rujuk Rajah 5). RStudio boleh didapati dalam dua bentuk versi iaitu RStudio Desktop untuk aplikasi desktop biasa dan RStudio Server yang membolehkan perisian ini diakses menggunakan laman pelayar semasa dari jarak yang jauh. Dalam kajian ini, versi RStudio Desktop digunakan. Bahasa pengaturcaraan R pula ialah sejenis bahasa pengaturcaraan yang mengkhususkan untuk persekitaran pengkomputeran berbentuk statistik dan grafik. Data yang diguna dan disimpan berbentuk fail format .csv (Comma Delimited) menggunakan perisian Microsoft Excel. Alat ini amat sesuai untuk mengekstrak data teks daripada Twitter dan Facebook. Aplikasi ini amat sesuai untuk pengguna bagi tahap lanjutan.

Kelebihan

Kelebihan RStudio adalah berdasarkan persekitaran pengkomputeran yang berupaya memberikan pengiraan berbentuk statistik yang boleh dipaparkan dalam bentuk grafik. Uniknya, Pengaturcaraan R ini menyediakan pakej *library* yang dibangunkan oleh orang perseorangan mahupun organisasi yang boleh dikongsi dan digunakan oleh para pengkaji lain bagi memudahkan pemprosesan data yang diperlukan. Dengan adanya pakej *library* yang dibangunkan ini, pengaturcaraan R ini berkeupayaan untuk mengekstrak data dengan menggunakan pakej twitteR untuk mencapai data daripada pelayan Twitter mengikut kriteria yang diinginkan. Pakej twitteR ini menyediakan kemudahan antara muka antara R dengan laman sesawang Twitter yang dikenali sebagai Application Programming Interface (API). Kemudahan ini membolehkan pengkaji untuk menyenarai tweet dengan menggunakan katakunci spesifik dan seterusnya mengumpulkan data (Rujuk Rajah 5).

Rajah 5 : Antaramuka RStudio versi 3.6.0

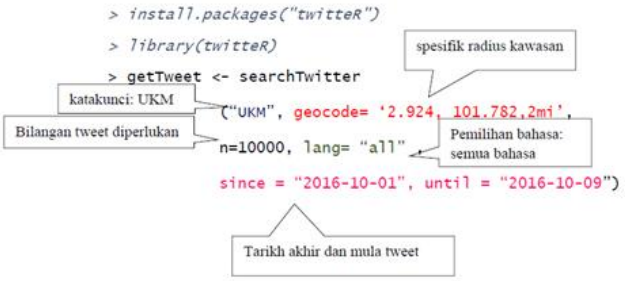


Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

Alat ini sesuai kepada pengguna yang mempunyai latarbelakang apa-apa bahasa pengaturcaran. Bahasa pengaturcaraan R digunakan dalam RStudio. Tutorial secara atas talian sama dalam bentuk video atau penulisan RStudio boleh diperolehi dengan mudah dengan menggunakan enjin carian jika ingin mencuba menggunakan alat ini. Jika dibandingkan dengan alat aplikasi laman sesawang https://exportcomments.com, alat ini membolehkan pengguna membuat carian data secara lebih spesifik dan tertumpu data yang dikehendaki sahaja. Contoh yang ditunjukkan dalam kajian ini ialah aturcara memuat turun tweet daripada Twitter (Rujuk Rajah 6). Terdapat lima faktor untuk fokus kepada data tertentu iaitu:

1. Katakunci
2. Spesifik radius kawasan (jika mahu)
3. Bilangan tweet diperlukan
4. Pemilihan bahasa
5. Tarikh mula dan akhir tweet

Rajah 4 : Contoh aturcara memuat turun data tweet dalam RStudio



Kelemahan

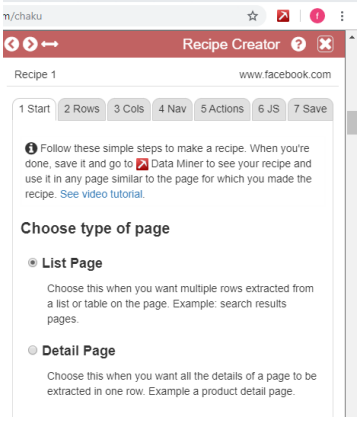
Kelemahan RStudio ini pula, jika pengguna menggunakan RStudio versi terkini, ada pakej *library* yang tidak boleh digunakan kerana keserasian sesuatu pakej masih belum diuji sepenuhnya oleh pasukan pengarturcara. Jadi, pengguna perlu menggunakan kaedah cuba dan buat untuk mengetahui versi RStudio dan pakej *library* sedia ada bagi memudahkan pengumpulan data. Inilah kelemahan perisian sumber terbuka secara umum.

(v) Perisian Dataminer

Data miner adalah perisian lanjutan yang dipasang pada pelayar Google Chrome yang membolehkan pengguna mengekstrak data teks dan menyimpannya dalam bentuk format fail csv. Jika tiada pelayar web Google Chrome, perisian ini tidak boleh dipasang. Cara pemasangan perisian ini mudah dengan mengikut tatacara pemasangan yang diberikan. Terdapat tujuh langkah untuk mengekstrak data teks dengan berpandukan kepada arahan yang diberikan dalam setiap langkah (Rujuk Rajah 7).

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

Rajah 5: Antaramuka perisian Dataminer



Kelebihan

Data miner boleh mengkstrak data teks semua jenis laman sesawang sama ada media sosial, laman sesawang berita, mahupun blog. Perisian ini menyimpan data sandaran selagi himpunan *data collection* dalam perisian ini tidak buang.

Kelemahan

Untuk menggunakan perisian ini, individu perlu mengetahui latarbelakang pengaturcaran pengetahuan asas pangkalan data yang terdiri daripada kolum dan baris. Perisian ini memerlukan individu memberikan input arahan berkenaan apakah jenis data (kolum) dan data apa yang hendak diekstrak (baris).

PERBINCANGAN

Berdasarkan lima alat pengekstrakan data teks janaan pengguna yang telah dibincangkan dalam bab Hasil Kajian, terdapat empat item utama yang dapat disimpulkan iaitu setiap alat ini merujuk kepada jenis perisian, tahap penggunaan, pengetahuan dan jenis laman sesawang yang sesuai digunakan (Rujuk Jadual 1). Hasil daripada pengujianan terhadap kelima-lima alat pengekstrakan data teks yang dijana pengguna ini, jenis perisian alat ini terbahagi kepada empat kategori iaitu aplikasi atas talian, perisian terbuka, pemasangan dan perisian lanjutan di Google Chrome. Tahap penggunaan dalam kajian ini dikelaskan kepada tiga bahagian iaitu permulaan, pertengahan dan lanjutan. Pengkelasan ini dibuat berdasarkan pengetahuan wajib yang diperlukan oleh seseorang sebelum menggunakan perisian ini melalui pengetahuan dan kemahiran asas penggunaan komputer, menulis aturcara program, pemahaman berkenaan pangkalan data seperti maksud kolum dan lajur serta perisian asas seperti Microsoft Excel amat diperlukan sebelum menggunakan perisian ini. Alat-alat ini amat sesuai digunakan untuk ke semua jenis laman sesawang media sosial dan bukan media sosial kerana yang penting data teks ini adalah terhasil dari janaan pengguna sendiri.

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

Jadual 1: Perbandingan ringkas alat pengekstrakan data teks janaan pengguna

Alat	**Jenis Perisian**	**Tahap Penggunaan**	**Pengetahuan**	**Jenis laman sesawang**
Exportcomments.com	Aplikasi atas talian	Permulaan	Asas literasi komputer Microsoft Excel	FB, Instagram, Twitter, TikTok, YouTube
Facepager	Perisian Terbuka	Pertengahan	Pengatucaraan Microsoft Excel	FB, YouTube, Twitter
Octoparse	Pemasangan (installation)	Pertengahan	Pengatucaraan Microsoft Excel	Semua media sosial, blog, wiki
RStudio	Perisian Terbuka	Lanjutan	Pengatucaraan Microsoft Excel	FB, Twitter
Dataminer	Perisian lanjutan Google Chrome	Lanjutan	Pengatucaraan Pangkalan Data Microsoft Excel	Semua media sosial, blog, wiki

Beberapa cadangan kajian masa hadapan telah dikenalpasti agar kajian ini dapat dikembangkan dan ditingkatkan lagi. Oleh itu, cadangan-cadangan tersebut adalah seperti berikut:

i. Kajian tingkah laku pengguna yang menggunakan ringkasan teks seperti perkataan "tidak" yang ditukar menjadi "x" dalam media sosial.
ii. Fokus menggunakan satu alat yang bersesuaian untuk mengekstrak data teks terhadap satu jenis media sosial untuk dijadikan data kepada kajian analisis sentimen.
iii. Membuat kajian pengkelasan data teks berkenaan tajuk yang menarik minat pengguna untuk terus membaca artikel tersebut.

KESIMPULAN

Berdasarkan objektif kajian yang dinyatakan di awal kajian ini, kajian ini telah menyenarai, mengkaji dan membuat perbandingan terhadap lima alat pengekstrakan data teks yang dijana pengguna (UGC). Kajian ini telah melalui lima proses utama bagi memenuhi objektif kajian. Setiap proses yang dibuat telah diulas secara terperinci terhadap setiap alat pengekstrakan data teks sebagaimana yang telah diterangkan dalam bab Kaedah Kajian. Secara keseluruhannya, kajian ini telah memberi pendedahan awal berkenaan bagaimana data teks UGC boleh diekstrak secara automatik oleh alat pengekstrakan data teks sedia ada dan boleh mengeluarkan output data teks dalam bentuk fail hamparan Microsoft Excel. Selepas melalui proses pengujianan bagi setiap alat, kajian ini mendapati bahawa alat pengekstrakan data teks ini sangat memudahkan kepada pengguna untuk mendapatkan data secara automatik dan sistematik. Antara kelebihan alat ini ialah memudahkan proses analisa komen-komen pengguna media sosial untuk tujuan pertandingan atau keperluan mengumpul komen berkenaan hasil produk jualan dengan cara yang cepat. Hasil rumusan dan penemuan kajian telah menunjukkan bahawa objektif kajian ini telah berjaya dicapai dalam skop yang telah ditetapkan. Cadangan masa hadapan juga telah dibincangkan supaya kajian ini akan lebih memberi manfaat serta menjadi rujukan kepada penyelidik yang mengkhusus kepada data teks janaan pengguna.

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

RUJUKAN

Ahmad, F., Ismail, N. H., Aziz, A. A. (2015) The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. *Applied Mathematical Sciences* (Vol. 9 No 129): 6415-6426 http://dx.doi.org/10.12988/ams.2015.53289

Amado, A., Cortez, P., Rita, P. & Moro, S. (2017) Research trends on Big Data in Marketing: A text mining and topicmodeling based literature analysis. *European Research on Management and Business Economics* 24: 1-7.

About Octoparse. (n.d). Retrieved from https://www.octoparse.com/blog/what-is-octoparse.
Dubras, R., & Tono, M. (2020) Digital 2020: 3.8 Billion People Use Social Media. Retrieved from https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media

Handayani, D., Bakar, N.S.A.A & Abuzuraida, M.A (2018) Sentiment Analysis for Malay Language: Systematic Literature Review. *2018 International Conference on Information and Communication Technology for the Muslim World*, 305-310.

Hiltz, S.R. & Plotnick, L. (2013) Dealing with Information Overload When Using Social Media for Emergency Management: Emerging Solutions. *Proceedings of the 10th International ISCRAM Conference,* 823-827

Huang, S., Peng, W., Li, J. & Lee, D. (2013) Sentiment and Topic Analysis on Social Media: A Multi-Task Multi-Label Classification Approach. *Proceedings of the 5th Annual ACM Web Science Conference,* 172-181.

Jalonen, H. (2014) Social Media – An Arena for Venting Negative Emotions. *International Conference on Communication, Media, Technology and Design:* 224-230.

Kente, M. (2017) Social Network Analysis. Department of Computer Science and Engineering University Of Gothenburg Retrieved from https://pdfs.semanticscholar.org/ee6e/78eb34e0f6287aaec7c1293d3a4cdd8ed270.pdf

Keyling, T & Junger, J. (2019) Facepager application. Retrieved from https://github.com/strohne/Facepager.

Kim, A. J., & Johnson, K. K. P. (2016). Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. *Computers in Human Behavior*, 58, 98–108. doi:10.1016/j.chb.2015.12.047

Korde, V. & Mahendar, C.M. (2012). Text Classification And Classifiers : A Survey. *International Journal of Artificial Intelligence & Applications (IJAIA Vol. 3, No. 2, March 2012)*: 85-99.

Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-Generated Content. IEEE Pervasive Computing, *7(4), 10–11.* doi:10.1109/mprv.2008.85

Nasridinov, A & Park, Y. H. (2013) Visual Analytics for Big Data Using R. *2013 International Conference on Cloud and Green Computing, Karlsruhe*, 564-565. doi: 10.1109/CGC.2013.96

Singh, T & Maini, R. (2013) A Comprehensive Review On Search Engine Optimization. *Journal of Global Research in Computer Science Volume 4 No 1*: 49-54

Sharma, S. & Srivastava, S.K. (2016) Review on Text Mining Algorithms. *International Journal of Computer Applications Volume 134:* 39-43.

Web Analysis for Exportcomments - exportcomments.com. (n.d.). Retrieved from

Vol. 17. No.3 (2020), 57-70. ISSN: 1823-884x

https://exportcomments.com.cutestat.com/.

MAKLUMAT PENULIS

FATIMAH RAHMAT
Jabatan Teknologi Maklumat dan Komunikasi
Politeknik Mersing
fatimah@pmj.edu.my

ZURAIHAH NGADENGON
Jabatan Teknologi Maklumat dan Komunikasi
Politeknik Mersing
zuraihah@pmj.edu.my

NURUL SHAKIRAH MOHD ZAWAWI
Jabatan Teknologi Maklumat dan Komunikasi
Politeknik Mersing
shakirah@pmj.edu.my